

Cost-effectiveness of natural forest regeneration and plantations for climate mitigation

In the format provided by the
authors and unedited

Supplementary Information for “Cost-effectiveness of natural forest regeneration and plantations for climate mitigation”

Full detail of construction of spatial data layers

The Materials and Methods section in the main text only provides an overview of the construction of the spatial data layers included in formulas (1), (2), and (3). Full detail on the construction of these layers is provided here in the Supplementary Information.

Implementation costs of natural regeneration and plantation

We generated implementation cost (\$/ha) surfaces for passive natural regeneration and forest plantations based on 355 observations of reported reforestation costs from 99 source documents spanning 26 LMICs in Africa, Asia and the Pacific, and Latin America and the Caribbean. We sourced the data from World Bank project documents and the peer-reviewed and grey literatures. All source documents were publicly available. We used machine learning to determine the linear regression model that provided the best cost predictions, and we used this model and spatial data layers to extrapolate implementation costs in 2020 US dollars per hectare across LMICs.

Project selection

In early 2020, we compiled a list of all World Bank projects initiated during 1992-2019 whose documents included at least one of the following search phrases: reforest*, afforest*, watershed rehabilitation, forest restoration, and tree planting. We defined “initiated” as the year of the project’s project appraisal document (PAD), which is the document reviewed by the Bank’s Board of Executive Directors during project approval. Our search yielded 383 projects, to which we added four earlier projects (two from 1987, one each from 1990 and 1991) that we had previously identified when evaluating the feasibility of extracting reforestation cost data from Bank documents. We used “reforestation” to encompass re-establishment of forest in areas where forests naturally occur. Projects where the previous land use was agriculture would be considered reforestation according to cit.⁸⁶ but would be considered afforestation according to cit.⁸⁷.

A review of the project abstracts indicated that 262 projects did not include reforestation components. The positive hits on the phrases used in the initial keyword search were due to incidental references in the project documents, e.g. in descriptions of the World Bank’s lending history in a country. For each of the remaining 125 projects, we visually searched the entire text of the PAD and, when it was available, the implementation completion report (ICR), for implementation cost data. We identified 59 projects containing usable data on implementation costs in the LMIC region. Given our objective to construct a spatial cost surface, we then dropped projects with insufficient information for identifying the Level 1 administrative subdivision (i.e., state or province) where the reforestation activity occurred. This lack of spatial detail often, but not always, affected older projects, as the World Bank began geolocating project locations after 2010. We also dropped projects that focused on special types of tree systems (e.g.,

agroforestry, shelterbelts, bamboo groves, roadside plantings, mangroves) that were not well-represented across the projects. We thus arrived at a final set of 48 World Bank projects.

Later in 2020 and continuing into early 2021, we conducted a snowball search for other sources of implementation cost data for the LMIC region during approximately the same period (early 1990s to present). We confined the search to the LMIC region for comparability to the World Bank data, all of which was collected from LMICs. We searched studies cited in cit.⁴³ and cit.⁸⁶, studies that they cited, and subsequent studies that cited these sources. We identified 74 studies containing usable data on implementation costs in the LMIC region, which reduced to a final set of 38 studies with the needed information, described above. These studies referred to reforestation by a range of implementers, from households and private companies to government agencies and nonprofit organizations, although, the identity of the implementer was often not evident.

The total number of observations in the final dataset was 355, with 231 from the 48 World Bank projects and 124 from the 38 other sources. Each observation referred to a data point for which a source document reported either implementation cost per unit area or both total implementation cost and total reforested area (so we could calculate cost per unit area). In some cases, a source document yielded just a single observation, such as the total cost and total area across an entire World Bank project. In most cases, however, each source document yielded multiple observations, such as total costs and total areas for distinct subprojects.

Data extraction

In our implementation cost database, plantations refer to forests that are artificially reestablished by human sowing of seeds or planting of seedlings. In contrast, natural regeneration refers to forest regeneration methods that rely on seeds that naturally occur on a site, i.e., seeds that are either present in the seed bank in the soil or transported to the site by the wind or seed-dispersing animals. Within the category of natural regeneration, our analysis focuses on passive natural regeneration, which entails minimal human interventions beyond evaluating a site to determine if natural seed sources are adequate to regenerate it and protecting regenerating trees on the site from predation by herbivores (livestock or wildlife) and human encroachment. Implementation costs for passive natural regeneration thus entail primarily site evaluation and protection measures such as fencing and guards. Also within the category of natural regeneration, assisted natural regeneration, which we did not include in our analysis, entails additional activities that vary according to site conditions. These activities can include site preparation, such as soil scarification, prescribed burning, and removal of invasive species, and actions to promote tree growth, such as fertilization and control of weeds, pests, and pathogens. The same activities commonly occur with plantations, making the essential difference between assisted natural regeneration and plantations being the source of regenerated trees: naturally occurring seeds in the case of assisted natural regeneration, and seeds sown or seedlings planted by humans in the case of plantations. Source documents indicated that project implementers typically chose planting instead of natural regeneration when natural seed sources were more limited (typically, land that had been in agricultural use for longer), project objectives emphasized wood production over biodiversity conservation, or both.

When extracting data on implementation costs from source documents, we attempted to record all direct monetary costs associated with the initial establishment of a new stand of trees (including site preparation costs) and subsequent management. Reported implementation cost data vary across sources within our database, with some estimates referring only to costs occurred in the initial year (the “establishment year”; e.g., fencing in the case of passive natural regeneration; site preparation and planting in the case of plantations) and other estimates also including management costs incurred in subsequent years (e.g., protection in the case of passive natural regeneration; protection, weed control, fertilization, etc. in the case of plantations). Sources also varied in terms of the number of years of additional management costs they reported. We followed the forestry principle of including costs through the “free to grow” stage: the age when trees are typically able to outcompete other vegetation. This age varies by location and site characteristics, but we assumed it to be 3-5 years. As described below, we included a dummy variable identifying these multiple-year cost estimates in the candidate variable set for our implementation cost model. We included administrative costs if they related directly to stand establishment and management but not broader program features (e.g., program design costs and coordination across multiple participants). The exclusion of indirect and nonmonetary costs and some administrative costs means we may underestimate the total economic costs of reforestation.

The source documents we reviewed did not report implementation costs in a standardized format. Most reported only an aggregate cost, without breaking it down by activity (e.g., site preparation, planting, fertilizing) or inputs (e.g. labor, fuel, seeds or seedlings, fertilizer). Given this heterogeneity, we developed an Excel template for data extraction that included some columns with predefined codes and others that were left undefined to allow the flexibility needed to record the available data. For observations that did not disaggregate cost by activities or inputs, we recorded the single aggregate cost; for those that disaggregated costs by activity or input, we recorded the disaggregated amounts and brief descriptions of them. For all costs, we recorded the currency (USD or local) and an indicator of whether they were reported as present values, which was rare. We processed the cost data to yield a single cost estimate for each observation, defined as the total implementation cost per hectare expressed in 2015 USD (later inflated to 2020 USD).

The template included additional columns for recording information that we used to generate 70 candidate variables that we hypothesized might be useful for predicting implementation cost per unit area (Table S2). These candidate variables formed nine groups:

1. Biophysical (16 variables; e.g., biome)
2. Socioeconomic (6 variables; e.g., national GDP per capita)
3. Land cover/land-use change (13 variables; e.g., agriculture being the prior land use)
4. Reforestation method (10 variables)
5. Species (4 variables)
6. Intended benefits (6 variables)
7. Area reforested (2 variables)
8. Cost characteristics (8 variables; e.g., cost includes only establishment year)
9. Source characteristics (5 variables; e.g., World Bank project year)

They included several geospatial variables based on the Level 1 subdivisions where the reforestation interventions occurred: mean elevation⁸⁸, mean annual maximum temperature, mean annual minimum temperature, and mean annual precipitation⁸⁹, biome area shares⁹⁰; population density⁹¹, and mean travel time to nearest city⁹². We considered including soil variables, but data were missing for too many of the subdivisions.

Statistical modeling

We analyzed implementation cost per hectare as a function of the candidate variables using lasso (“least absolute shrinkage and selection operator”). Lasso is a machine-learning regression method designed to select a model suitable for making predictions outside the data sample used for estimation^{93,94}. Lasso avoids overfitting by minimizing the out-of-sample prediction error subject to a constraint that the model must be sparse. We ln-transformed the dependent variable, implementation cost per hectare, in view of Box-Cox test results indicating that the ln-transformed model fit the data better than a purely linear model or a model with the inverse of implementation cost as the dependent variable. To ensure that source documents yielding multiple observations did not unduly influence the results, we weighted the variables by the inverse count of the number of observations from each source document.

The prediction model selected by lasso included just 10 of the 70 candidate variables (see Table S3). Two pertained to reforestation method, with predicted costs being higher for: (i) artificial regeneration (planting or direct seeding) and lower for (ii) passive natural regeneration compared to the reference case of assisted natural regeneration. Predicted costs were also higher: (iii) in countries with higher GDP per capita (ln-transformed); (iv) when costs were defined as including ones beyond the initial establishment year; and (v) in later years than earlier years (note that the annual increase is over and above inflation). Conversely, predicted costs were lower when: (vi) reforestation used exclusively native species; and (vii) the source document reported a single aggregate implementation cost without any detail by activities or inputs, which might reflect incomplete cost accounting.

The final three included variables pertained to three biomes: (viii) tropical and subtropical dry broad-leaved forests (negative effect); (ix) Mediterranean forests and woodlands (positive effect); and (x) deserts and xeric shrublands (negative effect). These variables give the area shares for the indicated biomes in the Level 1 subdivisions where reforestation occurred and did not necessarily indicate the biomes where reforestation occurred, as a given subdivision could include more than one biome. Their effects on implementation costs were in comparison to the group of all other biomes found in the Level 1 subdivisions, which were predominantly tropical and subtropical moist broad-leaved forests followed by temperate broad-leaved and mixed forests. Variables that were not selected by the machine learning model as useful predictors of implementation costs included population density, prior land use, and size of reforestation site (Table S2).

The mean implementation cost of reforestation across 355 observations was \$1,071/ha in 2020 USD (st. dev: \$1,103/ha) (Table S1). The machine learning model that we used to develop a spatially explicit implementation cost layer indicated that implementation costs were 9.6 times higher for plantations of introduced species than for passive natural regeneration, 1.4 times higher for plantations of introduced

species than for plantation of native species, increased with national income, and increased by 1.2% per year in real terms during the sample period (Table S3).

The direction of the effects of the ten variables on implementation costs (i.e., positive or negative) generally accords with intuition. One must be careful not to interpret lasso coefficients as causal effects, however, because lasso selects variables purely on the basis of their out-of-sample predictive value. The lasso model had reasonably high predictive accuracy: its out-of-sample R^2 was 0.86, and its mean prediction error (0.86) was much lower than the natural log of the weighted implementation cost per hectare (6.38).

Spatial extrapolation

We applied the lasso model described above to generate spatial implementation cost surfaces for LMICs in Africa, Asia and the Pacific, and Latin America and the Caribbean. We did not extrapolate costs to high income countries because our dataset did not include data from that region. We used dummy variables to generate a unique cost surface for each of three methods of reforestation: passive natural regeneration; plantations of native species; and plantations of introduced species. We later combined the two cost surfaces on plantations of native species and plantations of introduced species into a single surface on plantations, using information from Table 24 of cit.⁸⁷ on the percentage of plantations in each country that were native vs. introduced. In all cases, we set: the dummy variable for multiple-year cost estimates to one; the dummy variable for lack of cost details to zero; the three biome variables equal to their area shares in a pixel; the natural log of per-capita GDP equal to the latest year (2015) available in the gridded dataset cit.⁹⁵; and year equal to 2015. We later inflated values from 2015 USD to 2020 USD. Differences in costs across space within each surface were driven by the GDP per capita and biome variables.

Our spatial cost estimates are subject to at least three potential sources of bias. The first is the non-random locations of the reforestation projects in our sample. World Bank projects occurred only if the World Bank's Board of Executive Directors and its country government counterparts approved them. This approval process may have screened out reforestation projects in locations that were more costly based on unobserved factors, thus resulting in a sample that includes only lower-cost projects. A similar preference for lower-cost locations could conceivably affect studies from the peer-reviewed and grey literatures.

Second, costs reported in these literatures might also be biased downward if studies on projects that fail to be fully implemented due to unexpectedly high costs are not published. This bias could also affect reported costs for World Bank projects drawn from project appraisal documents, which are prepared before project implementation, instead of from implementation completion reports. Biases resulting from non-random locations and unpublished high-cost failures would result in underestimates of actual reforestation costs.

The third potential source of bias is the dominance of World Bank data in the sample: the sample included 231 observations from 48 World Bank projects (nearly 5 observations per project), compared to 124 observations from 38 other sources (closer to 3 observations per study). This dominance would

be a concern if World Bank projects were systematically more costly (or less costly) than other projects. However, this concerns us less than the other two potential sources of bias. As noted above, we weighted the observations in the lasso model to reduce the influence of source documents that provided larger numbers of observations. Furthermore, the set of potential predictor variables included an indicator variable for World Bank projects. Lasso did not select this variable, which indicates that information on the data source category (World Bank vs. other) was not useful for predicting reforestation costs.

For all three of these potential sources of bias, we have no reason to think they differentially affect natural regeneration vs. plantation projects; thus no reason to think they affect relative cost-effectiveness.

Opportunity cost of reforestation

Opportunity cost (\$/ha) in our study refers to the value of land if it is not reforested, i.e., its value in its continued current use. We assumed that this alternative use would be some form of agriculture, either cropland or pastureland. The opportunity cost of reforestation is the same for both methods of reforestation. That is, the opportunity cost of natural regeneration does not account for potential plantations, and vice versa.

We defined the opportunity cost as land rent, i.e., the portion of annual agricultural quasi-rent (essentially, short-run profit) that was attributable to land as a production input for raising crops or livestock. We generated a land rent surface across LMICs using a combination of gridded and household data. We used gridded data on gross crop revenue per hectare from the International Food Policy Research Institute (IFPRI)'s MapSPAM dataset^{23,24} and gridded data on livestock numbers (buffaloes, cattle, goats, horses, sheep) from FAO's Gridded Livestock of the World (GLW)²⁵. We used survey data from the World Bank's Living Standards Measurement Studies, compiled by FAO's Rural Livelihoods Information System (RuLIS) program²². The resolution of the opportunity cost surface was 5 arc-minutes (~10 km at the Equator), following the resolution of the MapSPAM and GLW datasets. Opportunity costs were generated for reference year 2019 and later inflated to 2020 USD.

Our approach makes two improvements upon previous global opportunity cost maps, such as cit.⁶² and its successors, which multiplied potential crop yields from agro-ecological zone maps by prevailing crop prices. First, we calculated net revenue instead of gross revenue, accounting for the share of gross revenue that is offset by production costs, using data from household surveys. Cit.⁶³ also estimated land rents instead of just gross revenues, but their estimates assumed a uniform 20% rent-share across countries and were based on data from only high-income countries (U.S., Canada, some European Union countries), not LMICs. Second, we estimated actual revenue rather than simulating potential revenue, drawing upon the best available data: the rich spatial MapSPAM dataset of where crops are actually grown, their yields, and their production values.

Our measure of actual gross revenue has several advantages over potential gross revenue. It is more likely to reflect land users' true opportunity costs, particularly in places where institutional factors prevent farmers (e.g., subsistence farmers) from planting crops that would access the highest market

prices, or within protected areas, or areas where agricultural activities occur despite legal prohibitions. And relevant for our study, actual gross revenue is the more appropriate opportunity cost of reforestation, where there is often already an existing agricultural land use, while potential gross revenue is the more appropriate opportunity cost of avoided deforestation, where an agricultural land use does not yet exist.

Land rent from cropland

We began with MapSPAM data on gridded crop revenue per hectare in 2010, which was the most recent year in the MapSPAM dataset. We updated the data to 2019 using country-specific ratios of crop revenue per hectare in 2019 compared to 2010, based on national data on total crop revenue and total cropland area from FAOSTAT⁹⁶. This procedure assumes that revenue changed by the same proportion in all areas of a country. While strictly untrue, this is the best available procedure until IFPRI updates MapSPAM to a more recent year.

Next, we multiplied the updated MapSPAM data by an estimate of the share of crop revenue that was attributable to land, i.e., the land-rent share. The land-rent share estimates differed across countries and, where data permitted, by Level 1 administrative subdivisions within countries. We developed the rent-share estimates through a two-step procedure.

In the first step, we used household survey data from RuLIS to econometrically estimate a model that related farm quasi-rent (equal to gross crop revenue minus expenditure on variable inputs, excluding any payments for land) to fixed inputs used in crop production. The estimation sample contained 229,859 observations from 196,327 unique farm households in 32 LMICs during 2004–2017. Table S4 shows the distribution of observations by country, and Table S5 shows the estimation results. The dependent variable in the model was the natural logarithm of quasi-rent, and fixed inputs were represented by the natural logarithms of cultivated area and family labor and a dummy variable that indicated if the farm was mechanized. The model also included year dummies and fixed effects for 242 regions. The regions were Level 1 subdivisions for most observations (222,771). For a small share (7,088), we could not determine a household's subdivision from the survey data—RuLIS protects respondent anonymity by reporting the administrative subdivision or country where a household is located instead of its geocoordinates—and instead used countries as the regions. (See asterisked countries in Table S4.) The year dummies controlled for unobserved factors that varied across time but not regions, and the region fixed effects controlled for unobserved factors that varied across regions but not time. We weighted the observations by the inverse count of observations by country to reduce the influence of variation in number of observations across countries.

Post-estimation, we calculated land rent for each observation by multiplying quasi-rent by the estimated coefficient on the log cultivated area variable, 0.325 (see Table S5). This procedure assumes that the coefficients on the fixed inputs in the log-log quasi-rent model can be interpreted as quasi-rent shares. This assumption is valid if production has constant returns to scale⁹⁷, which implies that the coefficients should sum to 1. The sum of the coefficients was 0.913, which was not significantly different from 1 ($p = 0.128$).

In the second step, we used sampling weights from the surveys to calculate mean values of crop revenue and land rent for each region (i.e., Level 1 subdivision or country) in the sample. We then calculated the land-rent share for each region by taking the ratio of mean land rent to mean crop revenue. This process yielded a sample of 1,032 rent shares, again from 32 LMICs during 2004-2017. Nearly all observations (1,006) referred to Level 1 subdivisions, with the remainder referring to countries. We econometrically related the natural logarithm of the land-rent shares to three spatial socioeconomic variables—the natural logarithm of the region’s gross domestic product per capita⁹⁴, natural logarithm of its population density⁹⁸, and strength of property rights (downscaled from the Rule of Law indicator in cit.⁹⁹—along with dummy variables for country classification by World Bank region, area shares for global ecological zones in the region¹⁰⁰, and an annual time trend. We choose 2015 as the reference year for the socioeconomic variables because it was the most recent year with subnational data on gross domestic product per capita⁹⁵. Table S6 shows the estimation results for the rent-share model.

We used this model to predict land-rent shares for Level 1 subdivisions in the LMICs, which we then multiplied by the gridded gross crop revenue data for 2019 to generate the cropland opportunity cost surface.

Land rent from pastureland

We used a similar procedure to generate the pastureland cost surface. We began with data from GLW on numbers of the five types of livestock in 2010. Data were not available for a later year. We estimated total gross revenue from livestock production in each grid cell by multiplying the livestock numbers by estimates of production value per animal in 2019, with the latter estimates based on country-level data on animal stocks and production value of livestock products¹⁰¹, and then summing across the livestock types. We adjusted the resulting estimates of total gross revenue per grid cell to include production only from pastured livestock, not livestock raised in feedlots, by using FAO estimates of national shares of meat production from grazing lands provided by the World Bank.

Unlike the cropland case, the household survey dataset contained too few observations on livestock production on pastureland, just 5,296 observations from 18 Level 1 subdivisions in four countries (Ethiopia, India, Peru, Vietnam), to enable us to estimate land-rent shares that varied spatially. Instead, we took a simpler approach. We regressed the natural logarithm of farm quasi-rent from pastured livestock on the natural logarithms of two fixed inputs, pastureland area and total livestock units, with year dummies and fixed effects for subdivisions also included. We weighted the observations by the inverse count of observations by country. Post-estimation, we determined the land-rent share by taking the ratio of the coefficient on pastureland area to the sum of that coefficient and the coefficient on total livestock units. The resulting share was 0.061 (Table S7). This low value implies that livestock contributed much more to quasi-rent from pastured livestock than did the land itself, which is suggestive of the rent dissipation that occurs when property rights for land are attenuated¹⁰².

We generated gridded estimates of total pastureland rent by multiplying this constant land-rent share by the gridded total revenue estimates. We converted total rent to rent per hectare by dividing total rent by total pastureland area in each grid cell. To estimate total pastureland area, we updated to 2015

a gridded global pastureland map for 2000 prepared by FAO^{103,104}. We updated this map by first regressing the pastureland share of grid cell surface area in 2000 on land-cover variables for 2000¹⁰⁵ and then using the regression coefficients and 2015 values of the land-cover variables to predict 2015 pastureland area within each cell.

Total land rent

We set the final opportunity cost surface equal to the cropland surface, with missing values or values of zero replaced with values from the pastureland surface. We prioritized the cropland surface for two reasons: we had more confidence in it because the household dataset included so many more observations on cropland than on pastureland; and it provided conservative (i.e., higher) estimates of opportunity cost at nearly all sites. We determined the opportunity cost of forgoing agricultural use of the reforested land for 30 years by calculating the 30-year sum of the present values of the annual land rent estimates, applying a time-discount rate of 5%, as justified below. The values in the opportunity cost surface are expressed in 2020 US dollars per hectare.

Likely plantation genus

We produced a map of the most likely plantation type to be planted across all LMICs at 1-km spatial resolution. To do so, we used data on the genus and spatial locations of planted forests across the globe from the Spatial Database of Planted Trees (SDPT)²⁶. We then used machine learning to model the likelihood of each genus of tree crop to be planted in that location using a suite of environmental and socio-economic covariates hypothesized to influence genus selection. This was a positive analysis of which plantation genus is most likely, based on historical distribution; not a normative analysis of which plantation genus is most preferable, based on native occurrence or ecosystem services provided.

Data description & preparation

Our model of the most likely plantation type is trained on the SDPT²⁶. This dataset is a non-systematic and non-representative compilation of locations of planted forests from various providers (e.g., national governments, nongovernmental organizations, independent researchers, or combinations of sources). Although the data are an incomplete representation of plantation types and locations, they are the best publicly available compilation of planted forests across the globe. Moreover, these data have been used in similar mapping exercises, such as creating global maps of planting years in plantations¹⁰⁶.

Prior to running our analysis, we filtered the SDPT to include only timber and fiber-producing plantations (i.e., excluding tree-crops such as oil palm, coffee, and fruit trees). In total, the dataset held 2,484,400 locations for the 15 genera of tree we considered: *Abies*, *Acacia*, *Betula*, *Cryptomeria*, *Cunninghamia*, *Eucalyptus*, *Larix*, *Picea*, *Pinus*, *Populus*, *Pseudotsuga*, *Quercus*, *Shorea*, *Tectona*, and *Tsuga*. The data were highly unbalanced (e.g., *Pinus* accounted for 52% of the total observations). Training random forest models on unbalanced datasets can lead to overpredicting the best-represented class. To avoid this, we randomly subsampled the data to extract 3,000 observations for each plantation genus. All genera had more than 3,000 observations, except *Cunninghamia* (n = 374) and *Cryptomeria* (n = 140). For these two genera, we resampled with replacement to reach 3,000 observations. This procedure likely overfits the

predictions for *Cunninghamia* and *Cryptomeria*, but it improves the overall predictive accuracy of the model.

Our training data included a suite of environmental, climatic, edaphic, and socio-economic covariates (see Table S8). The 64 climatic, edaphic, and environmental covariates were the same as those used to produce global maps of carbon accumulation from natural forest regrowth²⁰. We added 23 socio-economic covariates (Table S8) to account for non-environmental factors that might influence species selection for planting. All covariates were available as global maps, which we resampled to 1-km spatial resolution. The raster stack was resampled to 1-km resolution, with down-sampling performed via mean aggregation and up-sampling without interpolation. All preprocessing was performed in Google Earth Engine¹⁰⁷. We matched the full suite of covariates to all plantations in the SDPT using the centroid of each plantation location.

Identification of parsimonious predictors

We identified the most important subset of covariates for predicting the most likely plantation type using Variable Selection Using Random Forests (VSURF), a variable selection procedure available in the software program R¹⁰⁸. This package employs a widely-used variable selection procedure based on random forest scores of importance and a stepwise variable introduction strategy. We used the VSURF procedure to identify the subset of covariates that are most important for predictive accuracy, allowing for the design of a parsimonious predictive model. The method is computationally intensive, and thus we subsampled the full data frame of plantation-covariate combinations ($n = 2,484,400$) to extract 750 observations for each of the 15 different genera contained in the SDPT (again, sampling with replacement for *Cunninghamia* and *Cryptomeria*). The 750 observations per genus were determined to be a reasonable balance between representing the full covariate space and computational processing time.

Of the full suite of candidate covariates, the VSURF procedure identified 28 covariates to be included in the predictive model (Figure S2). Many of the identified variables represent fundamental constraints on tree growth (e.g., radiation of coldest quarter). Only three socio-economic variables (per capita GDP from 1990 and 2005; and distance to nearest medium-sized port) were included in the parsimonious set of predictors. Although plantation types are likely also influenced by relative differences in prices for different timber species across regions, there is no way to reliably predict how these relative prices might change in the future, and thus we did not include relative prices in the model. There is no reason to think this omission biases our predictions of plantation type.

Predictive modeling using Random Forests

We then trained a random forest model to predict the most likely plantation genus as a function of the parsimonious subset of covariates. Using the random forest classifier and the global maps of 28 covariates, we produced sixteen maps: i) fifteen maps indicating the likelihood of each genus being planted for each pixel of the globe, and ii) one map indicating the maximum-likelihood genus to be planted in each pixel. We did not tune the random forest parameters, but rather used default parameters (no. tree = 500, mtry = 3, nodesize = 1) in the R package {randomForest}¹⁰⁹.

Limitations

Our analysis of most likely plantation type has several limitations. First, the SDPT is not a statistical sample of global plantations and therefore is not perfectly representative of the spatial locations of global plantation forests. For example, the SDPT does not identify the genus of plantations in China, which accounts for roughly a quarter of all plantations on the globe. We partially addressed this shortcoming by gap-filling genus data in China (e.g., *Cunninghamia*) using an empirical database of carbon measurements in single-species plantations²¹. Given that our gap-filling was based on a limited number of observations, our predictions for *Cunninghamia* and *Cryptomeria* are likely to be overfit.

Our maps should not be taken as definitive or fully comprehensive predictions of the most likely type of plantation to be planted. Rather, they are suggestive of the dominant commercial genus that might be planted given currently available information on key climatic, edaphic, environmental, and socio-economic variables that are known to influence the type of plantation to be planted. Although the maps are sufficient for the purposes of our study, in practice the selection of species for planting involves site-level factors not fully captured by these variables.

Carbon accumulation in aboveground biomass

To produce maps of carbon accumulation over time in naturally regenerating forests and plantations, we used a modified version of the sigmoidal Chapman-Richards growth function. Specifically, we modeled spatial variation in two parameters of the Chapman-Richards function, the asymptotic limit of aboveground carbon (A) and the rate of carbon accumulation (k), as a function of spatially explicit environmental covariates. We then used these statistical relationships to produce 1-km global maps of A and k for each forest type, which in turn described aboveground carbon stocks as a function of stand age. Changes in residual aboveground carbon from prior land uses were not considered in this model. We obtained the training data for our growth functions from published datasets of aboveground carbon in naturally regenerating forests²⁰ and plantation forests²¹. A full description of the methods used to construct the datasets are contained in those studies, but we review the methods briefly here.

Datasets of aboveground carbon accumulation

The datasets were based on a Web of Science keyword search to identify studies with empirical estimates of carbon or biomass in stands undergoing some form of restoration of tree cover. The initial search included 11,360 studies, of which 5,464 were deemed relevant based on abstract review, and an additional 1,400 based on full review. The studies were then filtered to those with key information: (a) empirical measures of carbon (or biomass), (b) stand age, with stands between 0 and 100 years included, and (c) latitude and longitude, or a discernible geolocation (e.g., an identifiable place name). Of these final studies, 256 described natural forest regeneration and 424 studies described monoculture plantations.

Data cleaning and harmonization steps for the natural regeneration dataset included using the IPCC default biomass-to-carbon conversion factor of 0.47⁴⁷, excluding data points in mangroves and desert

biomes, excluding stands less than one year old, and for naturally regenerating forests, adding a biome-level average estimate of understory carbon to those studies that did not include understory in the aboveground carbon estimates. Understory carbon was not included in monoculture plantations given that understory biomass represents a minor component of total carbon stocks in plantation forests.

For this study, we further filtered the dataset to only include studies from LMICs. The final dataset included 2,309 observations of aboveground carbon stocks in naturally regenerating forests from 410 sites, and 3,289 observations of monocultures from 618 sites. There are several potential biases in these data. They only included systems that successfully regenerated and thus likely overestimate carbon accumulation in highly degraded sites or sites far from seed sources. Stand age may be increasingly uncertain in older stands. Lastly, carbon estimates may be overly high if sites were not fully cleared prior to regeneration and thus included residual biomass.

Processing of environmental covariates

To model spatial variation in carbon accumulation, we used the same raster stack of 64 environmental covariates that we used to model the most likely plantation genus. We only included the variables related to climatic, soil nutrient, soil chemical, soil physical, insolation, topographic, and nitrogen deposition variables²⁰. We did not include the socio-economic variables, as they should not drive variation in carbon accumulation.

It was infeasible to incorporate 64 covariates into a nonlinear regression, so we first performed a principal component analysis (PCA) to reduce the dimensionality of the data¹¹⁰. After masking biomes that do not support tree growth, we conducted the PCA on a sample of 5000 random locations from our covariate stack. Standard diagnostics for constructing principal components (PCs)—i.e., the Kaiser-Meyer-Olkin measure and the squared multiple correlations of each covariate with all the others—indicated that a separate PC should be constructed for aspect and hillshade. We therefore constructed a single PC from aspect and hillshade, and constructed an additional 10 PCs using the remaining 62 covariates. Of these 10 additional PCs, PC1 was strongly correlated with the temperature and insolation variables, PC2 was strongly correlated with precipitation and moisture availability variables, and PC3-PC10 correlated with the remaining predictors (climatic, edaphic, and topographical) (Figure S3). We then extracted the PC values for each geolocated observation of carbon accumulation in natural regenerating and plantation forests and used these data to estimate our growth model parameters.

Modified Chapman-Richards function

We used the carbon accumulation data and environmental covariates described above to produce 1-km resolution maps of the two parameters needed to fit a Chapman-Richards growth function. The Chapman-Richards function is a nonlinear growth model that is widely used within the forestry literature¹¹¹. The function is a generalization of the von Bertalanffy growth function, and its trademark is a flexible fit to empirical data^{112,113}. Moreover, the function is a theoretical model with biologically meaningful parameters:

$$y(t) = A * \left(1 - be^{-kt}\right)^{\frac{1}{1-m}} \quad (1)$$

Here $y(t)$ is the response variable, which in our study is aboveground carbon (tC/ha); t is the age of the forest in years; A is the asymptotic limit of the response variable; b is a scaling parameter; k is the rate of growth; and m is a shape parameter.

To reduce the statistical complexity of parameterization, it is common to fix both the scaling parameter and the shape parameter at assumed values instead of fitting them. We followed this procedure by setting $b=1$, fixing aboveground carbon to be 0 tC/ha at age 0. We set $m=0.5$, constraining the position of the curve's inflection point to a commonly used value within the literature. Moreover, a sensitivity analysis of fixing m at various values found $m=0.5$ to produce "best-fit" predictive models for the plantation data²¹.

To spatialize the Chapman-Richards growth function, we modeled asymptotic carbon stocks (A) and rates of growth (k) as functions of environmental covariates. The modified functional form for estimating aboveground carbon, incorporating both the fixed parameters and environmental covariates, was:

$$y_i(t) = \left(\beta_0^A + \sum_{m \in 1:M} \beta_m^A * x_{im} \right) \left(1 - e^{-(\beta_0^k + \sum_{n \in 1:N} \beta_n^k * x_{in})t} \right)^2 + \epsilon \quad (2)$$

Here i denotes a site; x_{im} is a set of M environmental covariates at site i hypothesized to influence the parameter A , the asymptotic limit of the response variable; and x_{in} is a set of N environmental covariates at site i hypothesized to influence k , the rate of growth. $\beta_0^A, \beta_m^A, \beta_0^k, \beta_n^k$ are coefficients to be estimated, and ϵ is a residual error term.

The resulting modified Chapman-Richards models thus produce spatially explicit estimates of carbon stocks as a function of stand age by incorporating spatial variation in the A and k parameters.

Parameter estimation & model validation

We estimated best-fit parameters for our modified Chapman-Richards function for eight different forest types: i) naturally regenerating forests, ii) *Pinus* plantations, iii) *Eucalyptus* plantations, iv) *Cunninghamia* plantations, v) deciduous broadleaf plantations, vi) evergreen broadleaf plantations, vii) deciduous needleleaf plantations, and viii) evergreen needleleaf plantations (Table S9). Our plantation database had >300 observations from more than 50 sites for each of the *Pinus*, *Eucalyptus*, and *Cunninghamia* genera, and we therefore produced individual models for these genera. For all other genera, we did not have sufficient data to produce genus-specific models (i.e., observations from fewer than 50 sites), and we produced one common model that incorporated a dummy variable for each of the four different plant functional types (i.e., deciduous broadleaf species, evergreen broadleaf species, deciduous needleleaf species, and evergreen needleleaf species).

For each of the eight models (i.e., forest types), we identified the most parsimonious set of principal components to be included. We sequentially added principal components to the model, updating the starting values to ensure model convergence after each addition. After all principal components were added, we then dropped the least significant principal component and re-estimated the model

parameters in a stepwise fashion, until all principal components were removed. We then identified the most parsimonious model specification using 10-fold cross-validation. We split the dataset into 10 equal folds by site to ensure true out-of-sample validation. For each fold, we then estimated model parameters using the remaining 9 folds and used these parameters to predict the aboveground carbon stocks of the withheld fold. We regressed the predicted and observed values for each withheld fold, calculated the root-mean-square prediction error, and averaged the prediction error across all ten folds. We then selected the model specification with the lowest prediction error as our most parsimonious model. In estimating the final model parameters, we accounted for spatial and temporal correlation in the data by clustering all standard errors by site ID (Ch. 8 in cit.[114](#)). All model estimation was performed in Stata/MP 17.

We ensured that our spatialized Chapman-Richards model had higher predictive accuracy than i) log-linear models that incorporated the environmental covariates, and ii) non-spatialized Chapman-Richards models. In both cases, our spatialized versions of the Chapman-Richards model produced lower predictive errors than the log-linear and aspatial models, suggesting that our modeling approach accounted for both nonlinear patterns in forest growth and spatial variation in constraints on growth patterns. As a final post-processing step, we truncated negative predicted values of A and k at zero for a small fraction of locations, as asymptotic carbon and carbon growth cannot be negative.

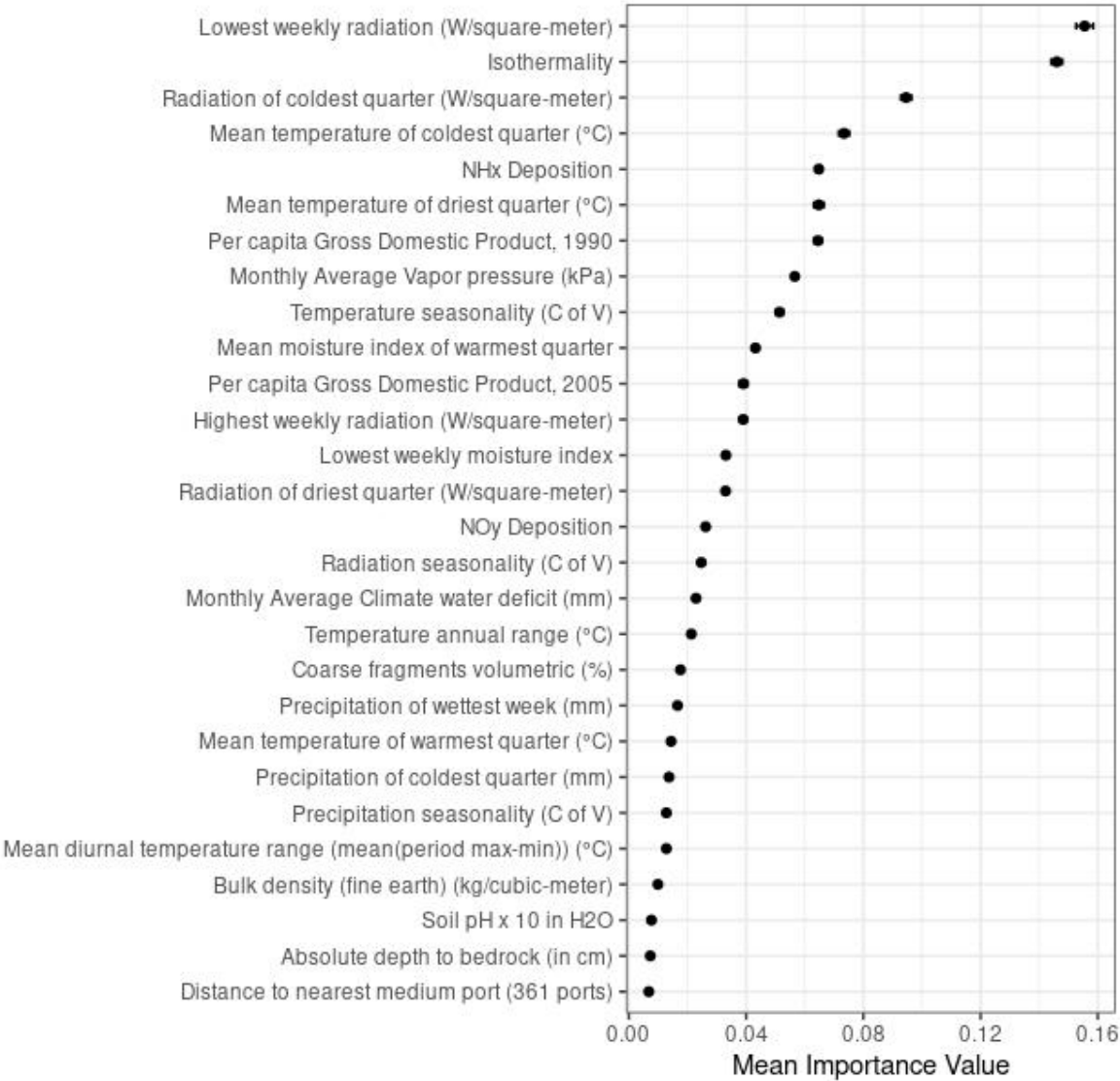
510 **Supplementary Information-only references**

- 511 86. IPCC (2018). Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of
512 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the
513 context of strengthening the global response to the threat of climate change, sustainable
514 development, and efforts to eradicate poverty. Intergovernmental Panel on Climate Change
515 (IPCC). Geneva, Switzerland.
- 516 87. FAO (2020). Global Forest Resources Assessment 2020: Main report. Food and Agricultural
517 Organization of the United Nations. Rome, Italy.
- 518 88. Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara. (2008). Hole-filled SRTM for the globe Version 4,
519 available from the CGIAR-CSI SRTM 90m Database. <https://srtm.csi.cgiar.org>.
- 520 89. Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. (2005). Very High Resolution
521 Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology*,
522 25:1965-1978.
- 523 90. TNC (2009). Global Ecoregions, Major Habitat Types, Biogeographical Realms and The Nature
524 Conservancy Terrestrial Assessment Units as of December 14, 2009. The Nature Conservancy,
525 Arlington, VA. <https://geospatial.tnc.org/datasets/TNC::terrestrial-ecoregions/about>
- 526 91. WorldPop (n.d.). Global Project Population Data: Estimated Residential Population per 100x100m
527 Grid Square. <https://hub.worldpop.org/project/categories?id=18>
- 528 92. Nelson, A., Weiss, D.J., van Etten, J. et al. A suite of global accessibility indicators. *Scientific Data*,
529 6:266 (2019). <https://doi.org/10.1038/s41597-019-0265-5>
- 530 93. Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*
531 *Statistical Society, Series B*, 58:267–288.
- 532 94. Hastie, T. J., Tibshirani, R.J., Wainwright., M. (2015). *Statistical Learning with Sparsity: The Lasso*
533 *and Generalizations*. CRC Press, Boca Raton, FL.
- 534 95. Kumm, M., Taka, M., Guillaume, J.H.A. (2018). Gridded global datasets for Gross Domestic
535 Product and Human Development Index over 1990–2015. *Scientific Data*, 5:180004.
- 536 96. FAO. (2020). FAOSTAT: Crops. Food and Agricultural Organization of the United Nations, Rome,
537 Italy. <http://www.fao.org/faostat/en/#data/QC>.
- 538 97. Chambers, R.G. (1988). *Applied Production Analysis: A Dual Approach*. Cambridge University Press.
539 Cambridge, UK.
- 540 98. CIESIN (2018). *Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision*
541 *11*. NASA Socioeconomic Data and Applications Center (SEDAC). Center for International Earth
542 Science Information Network, Columbia University, New York, NY, USA.
543 <https://doi.org/10.7927/H49C6VHW>.
- 544 99. World Bank (2020). *Worldwide Governance Indicators*.
545 <https://info.worldbank.org/governance/wgi/>
- 546 100. FAO. (2015). Global ecological zones for FAO forest reporting: 2010 update. Forest Resources
547 Assessment Working Paper 179. Food and Agricultural Organization of the United Nations, Rome,
548 Italy.
- 549 101. FAO. (2020). FAOSTAT: Livestock Primary. Food and Agricultural Organization of the United
550 Nations, Rome, Italy. <http://www.fao.org/faostat/en/#data/QL>.
- 551 102. Ostrom, E. (2008). Tragedy of the commons. The New Palgrave Dictionary of Economics. Palgrave
552 Macmillan, New York, NY.
- 553 103. FAO (2007). Occurrence of Pasture and Browse (FGGD). Food and Agricultural Organization of the
554 United Nations, Rome, Italy.
555 [https://data.apps.fao.org/map/catalog/srv/eng/catalog.search#/metadata/913e79a0-7591-11db-](https://data.apps.fao.org/map/catalog/srv/eng/catalog.search#/metadata/913e79a0-7591-11db-b9b2-000d939bc5d8)
556 [b9b2-000d939bc5d8](https://data.apps.fao.org/map/catalog/srv/eng/catalog.search#/metadata/913e79a0-7591-11db-b9b2-000d939bc5d8).

- 557 104. van Velthuisen, H., Huddleston, B., Fischer, G., Salvatore, M., Ataman, E., et al. 2007. Mapping
558 biophysical factors that influence agricultural production and rural vulnerability. *Environment and*
559 *Natural Resources Series No. 11*. Food and Agricultural Organization of the United Nations (FAO),
560 Rome, Italy.
- 561 105. ESA (2017). *Land Cover CCI Product User Guide, Version 2*. European Space Agency, Paris, France.
562 maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf.
- 563 106. Du, Z., Yu, L., Yang, J., Xu, Y., Chen, B., Peng, S., Zhang, T., Fu, H., Harris, N., Gong, P. (2022). A
564 global map of planting years of plantations. *Scientific Data*, 9:141.
- 565 107. Gorelick, N. et al. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone.
566 *Remote Sensing of Environment*, 202:18–27.
- 567 108. Genuer, R., Poggi, J.M., Tuleau-Malot, C. (2015). VSURF: An R package for variable selection using
568 random forests. *The R Journal*, 7:19–33.
- 569 109. Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.
- 570 110. Demšar, U. et al. (2013). Principal Component Analysis on Spatial Data: An Overview. – *Annals of*
571 *the Association of American Geographers*. 103:106–128.
- 572 111. Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*,
573 10:290–301.
- 574 112. Burkhardt, H.E., Tomé, M. (2012). *Modeling Forest Trees and Stands*. Springer Dordrecht, Berlin,
575 Germany. 458pp.
- 576 113. Salas-Eljatib, C. et al. (2021). Growth Equations in Forest Research: Mathematical Basis and Model
577 Similarities. *Current Forestry Reports*, 7(4): 230–244.
- 578 114. Angrist, J.D., Pischke, J.S. (2009). *Mostly harmless econometrics: an empiricist's companion*.
579 Princeton University Press. Princeton, NJ.
- 580 115. World Bank (2020). *World Development Indicators*.
581 <https://databank.worldbank.org/source/world-development-indicators>.
- 582 116. Trabucco, A., Zomer, R. (2019). Global Aridity Index and Potential Evapotranspiration (ET0) Climate
583 Database v2. figshare. Fileset. <https://doi.org/10.6084/m9.figshare.7504448.v3>
- 584 117. Fick, S.E., Hijmans, R.J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for
585 global land areas. *International Journal of Climatology*, 37(12):4302–4315.
- 586 118. Kriticos, D.J., Webber, B.L., Leriche, A., Ota, N., Macadam, I., Bathols, J. & Scott, J.K. (2012)
587 CliMond: global high resolution historical and future scenario climate surfaces for bioclimatic
588 modelling. *Methods in Ecology & Evolution*, 3:53–64.
- 589 119. Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C. (2018) Terraclimate, a high-
590 resolution global dataset of monthly climate and climatic water balance from 1958–2015, *Scientific*
591 *Data*, 5:170191.
- 592 120. Karlsson, K.G.; Anttila, K.; Trentmann, J.; Stengel, M.; Meirink, J.F.; Devasthale, A.; Hanschmann,
593 T.; Kothe, S.; Jääskeläinen, E.; Sedlar, J.; Benas, N.; van Zadelhoff, G.-J.; Schlundt, C.; Stein, D.;
594 Finkensieper, S.; Håkansson, N.; Hollmann, R.; Fuchs, P.; Werscheck, M. (2017): CLARA-A2: CM SAF
595 cCloud, Albedo and surface RAdiation dataset from AVHRR data - Edition 2, Satellite Application
596 Facility on Climate Monitoring, DOI:10.5676/EUM_SAF_CM/CLARA_AVHRR/V002.
- 597 121. Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., et al.
598 (2017) SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*
599 12(2): e0169748.
- 600 122. Danielson, J.J., Gesch, D.B. (2011) Global Multi-resolution Terrain Elevation Data 2010
601 (GMTED2010); Open-File Report 2011-1073. U.S. Geological Survey, U.S. Department of the
602 Interior, Washington, DC.
- 603 123. Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C.,
604 Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J.,

605 Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., van Vuuren, D. P.
606 (2010). Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive
607 gases and aerosols: methodology and application, *Atmospheric Chemistry and Physics*, 10:7017–
608 7039.
609

610 **Figure S1. Parsimonious set of predictors for classification of most likely plantation type.** Importance scores calculated by the Variable
 611 Selection Using Random Forests (VSURF) algorithm.



613 **Figure S2. Results of the principal component analysis to reduce the data dimensionality of our raster stack of 64 environmental covariates.**
614 The figure shows the principal component correlations (ranging from -1 to 1) with each original variable. The principal component for aspect
615 and hillshade is not shown.



617 **Table S1. Summary statistics for plantation projects in implementation cost database.** *Variable was included as natural logarithm in lasso
618 model. **Year of implementation was year of project appraisal document (PAD) in case of World Bank projects.

Variable	Units	All observations (n=355)		World Bank projects (n=231)		Other sources (n=124)	
		Mean	SD	Mean	SD	Mean	SD
Implementation cost per hectare* [dependent variable]	2011 USD/ha	\$1,071	\$1,103	\$929	\$862	\$1,336	\$1,416
Dummy: reforestation method = planting	0 or 1	0.90	0.30	0.95	0.22	0.82	0.38
Dummy: reforestation method = passive natural	0 or 1	0.07	0.25	0.08	0.27	0.04	0.20
Dummy: reforestation included native species	0 or 1	0.60	0.49	0.60	0.49	0.60	0.49
GDP per capita (PPP)*	2011 USD/person	\$7,431	\$5,738	\$6,552	\$5,209	\$9,068	\$6,315
Dummy: cost estimate spans multiple years	0 or 1	0.88	0.33	0.84	0.37	0.94	0.23
Dummy: cost estimate not disaggregated by inputs or activities	0 or 1	0.64	0.48	0.72	0.45	0.48	0.50
Area share: biome = tropical & subtropical dry broadleaved forest	Fraction (0 to 1)	0.06	0.13	0.03	0.08	0.12	0.18
Area share: biome = Mediterranean forests and woodlands	Fraction (0 to 1)	0.02	0.13	0.03	0.16	-	-
Area share: biome = deserts and xeric shrublands	Fraction (0 to 1)	0.08	0.18	0.05	0.14	0.14	0.24
Year of implementation**	Calendar year	2001.8	9.5	2000.5	8.7	2004.3	10.3

619

Table S2. Candidate covariates used to predict implementation costs in lasso model.

Covariate Type	Covariates	Data Source
Biophysical (16)	12 variables giving biome area shares for Level 1 subdivisions	TNC (2009) ⁹⁰
	3 climatic variables for Level 1 subdivisions: mean annual precipitation, mean annual maximum temperature, and mean annual minimum temperature (1970-2000)	Hijmans et al. (2005) ⁸⁹
	1 variable giving mean elevation of Level 1 subdivisions	Jarvis et al. (2008) ⁸⁸
Socioeconomic (6)	1 variable giving natural logarithm of national GDP per capita (purchasing power parity)	World Bank (2020a) ¹¹⁵
	1 dummy variable for projects in China	Source documents
	1 variable for population density of Level 1 subdivision	WorldPop (n.d.) ⁹¹
	3 variables for mean distance to nearest city within Level 1 subdivisions and interactions with artificial regeneration and natural regeneration dummies	Nelson et al. (2019) ⁹²
Land cover/land-use change (13)	3 dummy variables for broad prior land use: only forest, only nonforest, mix	Source documents
	3 dummy variables for more specific prior land use: forest, agriculture, other	
	2 dummy variables for prior agricultural use: cropland, pastureland	
	5 dummy variables for some (but not all) of project area involving specialized tree systems: silvopasture, other agroforestry, shelterbelt, bamboo, mangrove	
Reforestation method (10)	3 dummy variables for broad type of regeneration method: artificial, assisted natural, passive natural	Source documents
	2 dummy variables for type of artificial regeneration: planting, direct seeding	
	4 dummy variables for type of planting: monoculture, mixed species, enrichment, unspecified	
	1 dummy variable for coppicing	
Species (4)	4 dummy variables for type of regenerated tree species: natural, introduced (i.e., exotic), mix, unspecified	Source documents
Intended benefits (6)	6 dummy variables for intended benefits: wood, fodder, watershed protection, coastal protection, biodiversity conservation, carbon sequestration	Source documents
Area reforested (2)	1 dummy variable for total reforested area not reported	Source documents
	1 variable giving natural logarithm of total regenerated area	
Cost characteristics (8)	1 dummy variable for cost reported only per unit area (i.e., total cost and total area were not reported)	Source documents
	1 dummy variable for cost reported only as aggregate value (i.e., not disaggregated by activities or inputs)	
	3 dummy variables for cost reported for a multi-year period instead of just initial establishment year: yes, no, unspecified	
	3 dummy variables for time-discounting of costs that spanned multiple years: yes, no, unspecified	
Source characteristics (5)	4 dummy variables for type of source document: World Bank PAD or ICR, World Bank ICR, journal article, other	Source documents
	Establishment year (i.e., first year of implementation)	

Table S3. Implementation cost model: variables selected by lasso, with corresponding regression coefficients. See text for variable definitions. Dependent variable: natural logarithm of implementation cost per hectare, in constant 2011 USD. Lasso is not an inference-based technique and thus does not provide standard errors or other measures of statistical significance. Area shares refer to Level 1 subdivisions.

Variable	Coefficient
Dummy: regeneration method = planting	1.35
Dummy: regeneration method = passive natural	-0.916
Dummy: regeneration included native species	-0.356
ln(GDP per capita)	0.329
Dummy: cost estimate spans multiple years	0.206
Dummy: cost estimate not disaggregated by inputs or activities	-0.154
Area share: biome = tropical & subtropical dry broadleaved forest	-1.27
Area share: biome = Mediterranean forests and woodlands	0.725
Area share: biome = deserts and xeric shrublands	-1.04
Year of implementation	0.0123
Constant	-21.9

Table S4. Farm-level observations in estimation sample for cropland quasi-rent model, by country.
Total observations = 229,859. An observation refers to a farm household surveyed in a particular year during 2004–2017. Level 1 subdivisions could be determined for observations in all countries except those marked with an asterisk (*).

Country	Observations
Armenia*	5,726
Bangladesh	4,742
Bolivia*	214
Burkina Faso	6,711
Cambodia	7,240
Cameroon	4,940
Côte d'Ivoire	4,421
Ecuador	11,893
Ethiopia	5,381
Georgia	17,044
Ghana	8,720
Guatemala	8,511
India	25,350
Iraq	5,299
Kenya	6,140
Kyrgyzstan	2,438
Malawi	26,461
Mali	1,873
Mongolia*	292
Mozambique	7,014
Nepal	3,786
Nicaragua*	856
Niger	2,989
Nigeria	4,902
Pakistan	4,449
Panama	878
Peru	18,381
Rwanda	11,441
Tanzania	5,806
Timor-Leste	3,145
Uganda	7,734
Vietnam	5,082

Table S5. Fixed-effects estimation results: quasi-rent model for crops. Dependent variable: $\ln(\text{quasi-rent})$. Estimation sample: 229,859 observations from 196,327 unique farm households in 32 LMICs during 2004–2017. Model includes fixed effects for 242 regions (Level 1 administrative subdivisions or countries, depending on the survey), with observations weighted by the inverse count of observations by country to reduce the influence of variation in number of observations across countries. Coefficients give effect sizes; note that dependent variable is $\ln$ -transformed. Two-tailed t tests used to test significance, with standard errors robust and clustered by region. Estimates are from a single regression and thus have not been corrected for multiple comparisons. $R^2 = 0.854$; $F(15, 241) = 134$ ($p < 0.0001$); RMSE = 1.19.

Variable	Coefficient	Standard error	t statistic	$P > t $	95% confidence interval	
$\ln(\text{Cultivated area})$	0.325	0.030	11.00	0.000	0.267	0.383
$\ln(\text{Family labor})$	0.293	0.023	12.48	0.000	0.247	0.339
Dummy: mechanized	0.295	0.048	6.20	0.000	0.201	0.389
Dummy: 2004	-1.852	0.196	-9.46	0.000	-2.237	-1.466
Dummy: 2005	-1.229	0.182	-6.74	0.000	-1.588	-0.869
Dummy: 2006	-0.671	0.083	-8.05	0.000	-0.835	-0.507
Dummy: 2007	-0.110	0.210	-0.52	0.602	-0.523	0.304
Dummy: 2009	-0.381	0.155	-2.47	0.014	-0.686	-0.077
Dummy: 2010	0.118	0.164	0.72	0.473	-0.205	0.440
Dummy: 2011	-0.308	0.165	-1.87	0.062	-0.632	0.016
Dummy: 2012	0.008	0.144	0.06	0.954	-0.274	0.291
Dummy: 2013	0.063	0.127	0.50	0.619	-0.187	0.313
Dummy: 2015	0.263	0.117	2.25	0.025	0.033	0.494
Dummy: 2016	0.227	0.150	1.52	0.131	-0.068	0.522
Dummy: 2017	0.909	0.150	6.07	0.000	0.614	1.204
Constant	8.763	0.128	68.50	0.000	8.511	9.015

643 **Table S6. Estimation results: land-rent share model for crops.** Dependent variable: ln(land-rent share). Estimation sample: 1,011 observations in
644 32 LMICs during 2004–2017. Nearly all observations are Level 1 subdivisions, not countries, and were weighted by the subdivision's (or
645 country's) total cropland area. Reference category for World Bank region dummies: Sub-Saharan Africa. Reference category for area shares:
646 amalgamation of uncommon ecological zones, mainly subtropical and temperate steppes and all boreal zones. Coefficients give effect sizes; note
647 that dependent variable is ln-transformed. Two-tailed *t* tests used to test significance, with standard errors robust and clustered by country.
648 Estimates are from a single regression and thus have not been corrected for multiple comparisons. $R^2 = 0.623$; $F(20, 31) = 77.8$ ($p < 0.0001$); RMSE
649 = 0.238. ln(GDP per capita) and ln(Population density) are jointly significant at $p = 0.028$ ($F(2,31) = 4.03$).

Variable	Coefficient	Standard error	<i>t</i> statistic	P> <i>t</i>	95% confidence interval	
ln(GDP per capita)	0.047	0.074	0.63	0.531	-0.104	0.197
ln(Population density)	-0.054	0.031	-1.71	0.096	-0.117	0.010
Rule of law	-0.383	0.076	-5.02	0.000	-0.539	-0.228
Dummy: East Asia & Pacific	0.273	0.086	3.19	0.003	0.099	0.448
Dummy: Latin America & Caribbean	-0.087	0.152	-0.57	0.572	-0.397	0.223
Dummy: South Asia	0.269	0.103	2.61	0.014	0.059	0.479
Dummy: Middle East & N. Africa; Europe & C. Asia	0.338	0.125	2.71	0.011	0.084	0.592
Area share: Tropical rainforest	1.479	0.432	3.43	0.002	0.599	2.359
Area share: Tropical moist forest	1.471	0.422	3.48	0.002	0.609	2.332
Area share: Tropical dry forest	1.375	0.428	3.21	0.003	0.502	2.247
Area share: Tropical shrubland	1.329	0.414	3.21	0.003	0.485	2.173
Area share: Tropical desert	1.086	0.458	2.37	0.024	0.153	2.020
Area share: Tropical mountain system	1.340	0.446	3.01	0.005	0.431	2.249
Area share: Subtropical humid forest	1.365	0.425	3.21	0.003	0.498	2.233
Area share: Subtropical desert	1.408	0.608	2.32	0.027	0.169	2.647
Area share: Subtropical mountain system	1.141	0.451	2.53	0.017	0.221	2.061
Area share: Temperate continental forest	1.841	0.440	4.18	0.000	0.943	2.739
Area share: Temperate desert	0.249	0.431	0.58	0.568	-0.630	1.128
Area share: Temperate mountain system	1.267	0.424	2.99	0.005	0.403	2.131
Area share: Water	2.138	0.477	4.48	0.000	1.166	3.111
Year	0.061	0.016	3.85	0.001	0.029	0.093
Constant	-126.449	32.076	-3.94	0.000	-191.868	-61.029

652 **Table S7. Estimation results: quasi-rent model for livestock on pastureland.** Dependent variable: $\ln(\text{quasi-rent})$. Estimation sample: 5,296
653 observations from 5,082 unique farm households in 4 LMICs during 2010–2015. Model includes fixed effects for 18 regions (Level 1
654 administrative subdivisions), with observations weighted by the inverse count of observations by country to reduce the influence of variation in
655 number of observations across countries. Coefficients give effect sizes; note that dependent variable is $\ln$ -transformed. Two-tailed t tests used to
656 test significance, with standard errors robust and clustered by region. Estimates are from a single regression and thus have not been corrected
657 for multiple comparisons. $R^2 = 0.917$; $F(5, 17) = 91.1$ ($p < 0.0001$); RMSE = 1.02.

658

Variable	Coefficient	Standard error	t statistic	$P > t $	95% confidence interval	
$\ln(\text{Permanent meadow area})$	0.046	0.015	3.200	0.005	0.016	0.077
$\ln(\text{Total livestock units})$	0.522	0.057	9.120	0.000	0.401	0.643
Dummy: 2010	-0.287	0.044	-6.470	0.000	-0.381	-0.194
Dummy: 2013	-0.388	0.046	-8.360	0.000	-0.485	-0.290
Dummy: 2014	-0.019	0.086	-0.220	0.827	-0.201	0.163
Constant	9.940	0.047	209.970	0.000	9.840	10.039

659

Table S8. Candidate climatic, edaphic, environmental, and socio-economic covariates (n=82) used to predict the most likely plantation genus.

Covariate Type	Covariates	Data source
Climatic	Aridity Index	CGIAR (Trabucco and Zomer 2019) ^{116}
	Annual evapotranspiration	
	19 BioClimatic variables describing temperature and precipitation	BioClim (Fick and Hijmans 2017) ^{117}
	16 WorldClim variables describing insolation & moisture availability	WorldClim (Kriticos et al. 2012) ^{118}
	7 TerraClim variables describing drought indices, runoff, soil moisture, vapor pressure, and wind speed	TerraClim (Abatzoglou et al. 2018) ^{119}
	Average monthly shortwave radiation	EUMETSAT (Karlsson et al. 2017) ^{120}
Edaphic	Probability of occurrence of R horizon	SoilGrids, 250m (Hengl et al. 2017) ^{121}
	Absolute depth to bedrock	
	Bulk density (fine earth fraction)	
	Cation exchange capacity	
	Clay content	
	Silt content	
	Sand content	
	Coarse fragments volume	
	Soil organic C density	
	Soil pH in H ₂ O	
	Soil pH in KCl	
	Field capacity (soil water capacity until wilting point)	
Topographic	Aspect	GMTED (Danielson and Gesch 2011) ^{122}
	Elevation	
	Hillshade	
	Slope	
Environmental	NH _x Deposition	Lamarque et al. 2010. ^{123}
	NO _y Deposition	
Socio-economic	Per capita GDP (1990, 1995, 2000, 2005, 2010, & 2015)	Kummu et al. 2018 ^{95}

	Distance to cities of 12 different population sizes	Nelson et al. 2019 ^{92}
	Distance to ports of 5 different sizes	

662 **Table S9. Selection of carbon accumulation models for the 15 genera of candidate plantations.**

Plantation genus	Carbon accumulation model applied
<i>Abies</i>	Needleleaf Evergreen
<i>Acacia</i>	Broadleaf Evergreen
<i>Betula</i>	Broadleaf Deciduous
<i>Cryptomeria</i>	Needleleaf Evergreen
<i>Cunninghamia</i>	Cunninghamia
<i>Eucalyptus</i>	Eucalyptus
<i>Larix</i>	Needleleaf Deciduous
<i>Picea</i>	Needleleaf Evergreen
<i>Pinus</i>	Needleleaf Evergreen
<i>Populus</i>	Broadleaf Deciduous
<i>Pseudotsuga</i>	Needleleaf Evergreen
<i>Quercus</i>	Broadleaf Deciduous
<i>Shorea</i>	Broadleaf Deciduous
<i>Tectona</i>	Broadleaf Deciduous
<i>Tsuga</i>	Needleleaf Evergreen

663

664 **Table S10. Estimation results: Chapman-Richards models for carbon accumulation.** Dependent variable: above-ground carbon (tC/ha). PC =
665 principal component. Standard errors: robust and clustered by study site. Significance: *** = 1%, ** = 5%, * = 10%. See bottom of table for
666 number of observations, in-sample goodness-of-fit (R^2), and two measures of out-of-sample prediction errors (root mean squared prediction
667 error, RMSPE; mean absolute prediction error, MAPE).
668

Parameter	Natural regeneration	Plantation genus			
		Pinus	Eucalyptus	Cunninghamia	Other genera
A	108.5***	152.7***	201.3***	-145.6*	101.4***
	(6.697)	(17.52)	(18.26)	(78.66)	(12.21)
A: broadleaved evergreen					22.15**
					(10.58)
A: needle-leaved deciduous					82.45**
					(33.21)
A: needle-leaved evergreen					99.29***
					(21.98)
A × 1st PC		-17.86***		-181.3***	
		(4.364)		(32.61)	
A × 2nd PC		-12.43***	-29.62***		
		(2.587)	(4.825)		
A × 3rd PC	-4.481*			-74.93**	-8.041*
	(2.399)			(37.07)	(4.108)
A × 4th PC			-13.56***	-33.63**	-20.39***
			(3.706)	(13.70)	(3.289)
A × 5th PC		12.12***	-17.02***		-13.64***
		(4.653)	(5.909)		(3.956)
A × 6th PC	20.69***			338.2***	-9.140*
	(5.431)			(69.73)	(4.786)
A × 7th PC					-8.928***
					(3.036)

A × 8th PC			22.38***	-136.4***	14.41***
			(5.936)	(37.90)	(5.484)
A × 9th PC	-8.268*		13.70	-328.7***	24.24**
	(4.310)		(13.62)	(81.96)	(10.71)
A × 10th PC		9.662**	75.69***	-105.6***	5.991
		(4.093)	(7.837)	(37.56)	(5.074)
A × aspect/hillshade PC	-8.619***				
	(3.088)				
k	0.0346***	0.0534***	0.124***	0.119***	0.107***
	(0.00265)	(0.00627)	(0.00996)	(0.0155)	(0.0111)
k: broadleaved evergreen					-0.0199**
					(0.00937)
k: needle-leaved deciduous					-0.0626***
					(0.0106)
k: needle-leaved evergreen					-0.0735***
					(0.00961)
k × 1st PC		0.00232**		0.0248***	
		(0.00115)		(0.00363)	
k × 2nd PC	0.00650***	0.00898***	0.0225***	-0.00751***	0.00394***
	(0.000937)	(0.00174)	(0.00400)	(0.00280)	(0.000894)
k × 3rd PC				0.0130***	
				(0.00363)	
k × 4th PC		0.00323**			0.00499***
		(0.00130)			(0.00110)
k × 5th PC	-0.00535***	-0.00229			0.00506***
	(0.000917)	(0.00239)			(0.00125)
k × 6th PC			-0.0136	-0.0389***	0.00312*
			(0.00907)	(0.00485)	(0.00160)
k × 7th PC		-0.00257**		-0.00846***	
		(0.00129)		(0.00243)	
k × 8th PC		0.00726**	0.00958	0.0123***	

		(0.00289)	(0.00635)	(0.00310)	
k × 9th PC			0.0157	0.0495***	-0.0161***
			(0.0113)	(0.00738)	(0.00301)
k × 10th PC	0.00656***		-0.0553***	0.0283***	
	(0.00234)		(0.0153)	(0.00563)	
k × aspect/hillshade PC	0.00399***				
	(0.00141)				
Observations	2,309	1,192	323	461	1,313
R ²	0.787	0.847	0.886	0.801	0.811
% RMSPE	65%	54%	51%	66%	60%
% MAPE	44%	38%	31%	47%	43%