A novel framework to generate plant functional groups for ecological modelling

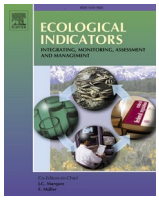
M. Calbi, G. Boenisch, Isabelle Boulangeat, D. Bunker, J.A. Catford, A. Changenet, V. Culshaw, A.S. Dias, T. Hauck, J. Joschinski, et al.

Ecological Indicators 166 (2024) 112370

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Original Articles

A novel framework to generate plant functional groups for ecological modelling

M. Calbi [a,*], G. Boenisch [b], I. Boulangeat [c], D. Bunker [d], J.A. Catford [e], A. Changenet [a,f], V. Culshaw [g], A.S. Dias [h], T. Hauck [i], J. Joschinski [i], J. Kattge [b,j], A. Mimet [k], M. Pianta [a], P. Poschlod [l], W.W. Weisser [g], E. Roccotiello [a]

[a] Department of Earth, Environment and Life Sciences (DISTAV), University of Genoa, Genoa, Italy
[b] Max Planck Institute for Biogeochemistry, Jena, Germany
[c] INRAE, Grenoble, France
[d] Department of Biological Sciences, New Jersey Institute of Technology, Newark, NJ 07102, USA
[e] Department of Geography, King's College London, 30 Aldwych, London WC2B 4BG, UK
[f] Swedish University of Agricultural Science, Remote Sensing of Forests Division, UMEÅ, Sweden
[g] Technical University of Munich, Terrestrial Ecology Research Group, Department of Life Science Systems, School of Life Sciences, Freising, Germany
[h] Institut für Physische Geographie, Biogeography and Biodiversity Lab, Goethe-Universität Frankfurt, Frankfurt am Main, Germany
[i] Studio Animal-Aided Design, Berlin, Germany
[j] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[k] Laboratoire BiodivAG, Faculté des Sciences, Université d'Angers, France
[l] Ecology and Conservation Biology, Institute of Plant Sciences, University of Regensburg, Regensburg, Germany

ARTICLE INFO

Keywords:
Functional ecology
Ecological niche
Hybrid dynamic models
Plant functional types

ABSTRACT

An effective way to reduce complexity in ecological modelling is by grouping species that share similar characteristics into functional groups or types. Often, the creation of plant functional groups (PFGs) is carried out for each case study in an *ad-hoc* way using a small set of traits. This limits the transferability of these PFGs to other geographical areas or study systems. We propose a novel generic framework to generate PFGs that considers the most important ecological dimensions, is applicable to case studies globally, and that emerges from patterns of functional redundancy across species. Based on most relevant and measured plant characteristics, we designed a multi-step process that includes: i) data harmonisation and missing values imputation; ii) species clustering based on multiple characteristics encompassing the main ecological dimensions featured in plant community ecological models (i.e., dispersal, competition, and demography) and iii) the combination of ecological dimension-specific groups into comprehensive PFGs. We demonstrate this framework by applying it to a global dataset of plant characteristics including a functional traits dataset and a plant-soil co-occurrence dataset for 19,102 species. Lastly, to test the ability of generated PFGs to summarise species' functional variation within plant communities, we correlate taxonomical and functional diversity indices calculated at the species and at the PFGs level across a global dataset of plant communities (sPlotOpen). Our framework generated 465 global, robust data-driven PFGs with non-overlapping combinations of traits for each ecological dimension divided by growth form. The validation returned positive correlation values between PFGs and species-level diversity metrics, supporting the ability of the obtained PFGs to capture functional and taxonomic diversity patterns across a variety of plant communities worldwide. The framework allows for the easy integration of newly available species characteristics data. The obtained global PFGs, covering all main known ecological processes and environmental conditions at small resolution, can increase the predictive power and accuracy of process-based models and help furthering varying-scale ecological studies.

* Corresponding author.
E-mail address: mariasolecalbi@hotmail.com (M. Calbi).

https://doi.org/10.1016/j.ecolind.2024.112370
Received 6 March 2024; Received in revised form 3 July 2024; Accepted 11 July 2024
Available online 18 July 2024

1. Introduction

Generating reliable predictions of ecological community dynamics under varying biotic and abiotic constraints and environmental changes is a central research topic in predictive ecology (Mouquet et al., 2015). In plant ecology, this topic has been explored through the development of predictive phenomenological, process-based or hybrid dynamic vegetation models (hybrid-DVMs; Boulangeat et al., 2012; Sitch et al., 2008). Performance and upscaling of vegetation models are hampered by the co-occurrence of ecological processes and the sheer number of species in natural systems (Schulze et al., 2019). The inherent ecological model complexity brings about three main trade-offs that the different modelling approaches must deal with simultaneously, which are the spatial scale, the specificity of the modelled entity, and the modelled processes. Typically, different modelling approaches rely on different spatial scales measured by extent and resolution (Boulangeat et al., 2012). Vegetation modelling approaches are either very broad and coarse (i.e., large geographical scale/low resolution) or highly context-specific (i.e., limited to specific geographical areas or habitats). In the case of modelling approaches with the same degree of complexity, there is still a trade-off between the extent and resolution: the larger the extent, the coarser the resolution. The extent is directly linked to the specificity of the modelled entity: the larger the extent, the more universal, or less study-specific the modelled entity is required to be. The spatial scale is also connected to the choice of the modelled entity through the differences in ecological processes taking place at the different spatial scales.

An ecological modelling entity is defined as the basic unit at which behaviour is modelled (Grimm et al., 2010). The specificity or universality of the modelled entity is driven by the spatial extent of the study and by the complexity of the modelled processes. Modelled entities are usually groups of individuals with similar characteristics. Ordered from the most specific to universal, and from higher to lower ecological resolution, these groups consist of individuals belonging to species, plant functional groups (PFGs) or plant functional types (PFTs). PFGs and PFTs are both created by grouping species by functional attributes or traits, and thus underlying convergent ecological strategies (Lavorel et al., 2007). Grouping species into PFGs (instead of simply using higher taxonomic ranks) has the advantage of keeping the focus on how species function rather than from which evolutionary history they originate, and reduces complexity. Generally, PFGs are based on a larger number of traits than PFTs that are usually broad groups of species with similar growth form, climate preferences and photosynthetic pathways (Wullschleger et al., 2014). PFTs are defined by grouping species having similar responses to abiotic constraints and similar effects on the main ecosystem processes (Walker 1992; Noble and Gitay, 1996) and are normally employed in dynamic global vegetation models (DGVMs) due to their universality. The choice of the characteristics of the PFGs or PFTs depends on the processes modelled at the scale of interest. The more numerous and complex the modelled processes, the more traits or associated parameters the modelling entity needs to encompass, and hence the higher the modelling entity specificity. Species resolution is the highest in specificity, being highly detailed but difficult to parametrize when a large number of species is considered simultaneously.

Existing PFTs/PFGs classifications have been driven by the assumption that plants are constrained in their performance by different resource availabilities (Lavorel et al., 2007), such as light, water and nutrients, competition and disturbance regimes (Grime, 1974; Smith et al., 1993, Tilman, 1988). Several bottom-up PFTs/PFGs classifications already exist but are in their majority tailored to a specific study region, or conversely, they are unable to capture fine scale processes because of the extent-resolution trade-off. These classifications are mostly based on sets of functional traits (Hodgson et al., 1999; Weiher et al., 1999) that relate to environmental factors and species coexistence mechanisms that can be found implemented in models. The main classifications employed in ecological studies and ecological modelling are listed in Table 1. While each of these approaches has brought substantial innovation and advancement to ecological modelling, it has been difficult to apply them across varying spatial scales and ecosystems. One major reason is overall trait data availability and species coverage, which has so far acted as a key limiting factor on the level of detail and number of groups to be developed.

The main ecological dimensions of a data set of various functional traits can be considered as the minimum number of latent variables required to describe it (Lee and Verleyson, 2007; Laughlin, 2014), or as essential groups of traits or characteristics (e.g. habitat, climatic or soil preferences) that are reduced in number (Winemiller et al., 2015). The PFT classification developed by Sitch et al. (2003) aims at representing the main axes of variation at a global scale and considers only the first two dimensions of the 'global spectrum of plant form and function' (Díaz et al., 2016) (plant size, leaf economics). The resulting groups are very broad and overlook e.g. dispersal or pollination mechanisms. In general, the most commonly measured traits in global databases are related to each main ecological dimension but are often used for separate PFTs/PFGs classifications. To be employed in a wide array of vegetation models and spatiotemporal contexts, while adequately summarising functional redundancy along the main axis of ecological variation including responses to biotic and abiotic factors, a successful PFT/PFG classification should include more ecological dimensions (Laughlin, 2014), yet this has been hindered so far by data availability. Due to the recent increase in data availability of plant traits (e.g. Kattge et al., 2020) and progress in gap-filling techniques (Debastiani et al., 2021), it has now become possible to include more traits and characteristics into the analyses and thus address more ecological dimensions. Boulangeat et al., (2012) for example, selected key functional traits related to six main ecological dimensions: resistance to disturbance, dispersal, tolerance to abiotic conditions, response to competitions, competitive effect, and demographic characteristics. While the approach of Boulangeat et al (2012) is robust, capturing essential ecological processes and retrieving a fair number of PFGs, it was designed for a regional scale and considers only locally dominant species. Hence, it cannot be easily transferred to another region. In our view, to find a solution around the extent-resolution trade-off, and to support the development of predictive modelling frameworks that can work seamlessly across habitat types and ecosystems, existing classifications, and in particular the approaches of Sitch et al. (2003) and Boulangeat et al (2012), require modifications (i.e. the possibility of including more ecological dimensions and species) to become widely applicable and to encompass all relevant trait variation globally.

A global framework to derive PFGs classifications based on emerging groups of species with similar characteristics (Lavorel et al., 1997) should:

i) include a reduced number of entities considering the trade-offs between the number of modelled entities and the level of details of the modelling outcome, while maximising species neutrality within groups and niche differentiation among groups (Hérault, 2007);
ii) be representative of all different biomes and continents, encompassing as much possible variation present across all plant species for a selection of attributes that relate to universal key ecological dimensions;
iii) maintain a conceptual link to usually employed modelling entities (e.g., PFTs by Sitch et al., 2003) to allow for its implementation in existing global modelling frameworks and the comparisons of modelling results.

The two most important challenges for universal PFGs classifications are thus the choice and selection of plant characteristics, and reducing the number of potential existing combinations of continuous values of the selected characteristics. Selected characteristics must capture and not just imply fundamental ecological processes (also known as the main

Table 1
Main existing PFGs/PFTs classifications frameworks employed in ecological studies and ecological modelling, their features, and main caveats for their implementation at varying spatial scales and across ecosystems. DVMs = dynamic vegetation models; DGVMs = dynamic global vegetation models; LPJ=Lund-Postdam-Jena model.

Name of the approach	Rationale	Modelling use	Scale	Variant (study-specific) or invariant	Advantages	Caveats
Life forms (Raunkiaer, 1934)	Classification based on the location of dormant meristems from where the plants start regrowing after the cold winter.	DGVMs	biome, region	invariant	Some correlation to height, leaf characteristics and phenology (Garnier et al., 2001; Lavorel et al., 2007); linked to overall responses of plants to disturbance (McIntyre et al., 1999), since the different locations of buds might reflect different success rate after perturbations.	No clear relationships with specific traits. Trait values ranges might not be constant within life forms (Foley et al., 1996; Steffen et al., 1996). Life forms can be used as trait variation proxies only in ecosystems including all life forms (Lavorel et al., 2007).
Plant Functional Types (PFTs) (Sitch et al., 2003)	PFTs based on combination of growth forms, leaf type (broadleaved/needle leaved) and leaf phenology type (evergreen/deciduous) that specify about 10 primary PFTs of which some are then further classified by climatic regions (tropical/temperate/boreal).	LPJ model; other DGVMs	biome, region	invariant	Synthetic and effective at global scale for modelling biomes dynamics. Explicit habitat preference for biomes.	Characterised by very few traits covering only two out of three dimensions of the 'global spectrum of plant form and function'.
Competitive-stress-tolerant-ruderal (CSR) scheme (Grime, 1974, 1977)	Multi-trait axes based on trade-offs between attributes for high resource acquisition in productive habitats and those for retention of resources in unproductive conditions; recurring plant specialisation strategies are identified. The C-S axis summarises the variation in responses to chances of rapid growth while the R axis reflects coping with disturbances.		landscape, region	variant	Innovative and comprehensive enough to be relevant at the global scale, allowing for comparisons among species, communities, and floras, especially in its late generalisation by Pierce et al., (2013, 2017).	Based on a small set of leaf traits; many ecological processes such as dispersal, pollination and below-ground interactions are implicitly inferred through correlations that are not well-supported or valid worldwide (Tilman, 1988).
Leaf-height-seed (LHS) scheme (Westoby, 1998).	Based on three traits: specific leaf area (SLA) reflects the same type of variation as the C-S axis in CSR; height and seed mass reflect separate aspects of coping with disturbance (axis R in CSR).		landscape, region	variant	Permits any species worldwide to be readily positioned within the scheme; captures a high portion of variability that is highly correlated to other relevant traits in specific geographical contexts (i.e., pine forests in Laughlin et al., 2010).	LHS axes independence was not confirmed for some geographical contexts (Mediterranean grazing systems in Golodets et al., 2009); does not provide reliable information on dispersal distances that cannot be directly related to seed mass or other plant attributes (Hughes et al., 1994).
Emergent Groups Approach (Lavorel et al., 1997; Pausas and Lavorel, 2003; Pillar and Sosinski, 2003; Hérault and Honnay, 2005; Hérault 2007: Boulangeat et al., 2012).	Groups emerging from sets of representative soft traits, allowing to identify sets of species with similar functional niches, thus convergent ecological strategies (Hérault and Honnay 2005).	hybrid DVMs and vital attribute-based models	landscape, region	variant	Outputs a relatively small number of Plant Functional Groups (PFGs), fulfilling the needs for functional equivalence within and of functional divergence between groups.	Highly context-specific, often ad-hoc and based on small sets of traits (Harrison et al., 2010; Boulangeat et al., 2012); usually lack simultaneous representation of herbaceous, parasitic, epiphytic, or aquatic species (Boulangeat et al., 2012), comparable sampling coverage of different biomes or continents (but see Pierce et al., 2013), and evenly distributed sampling in terms of evolutionary lineages; explicit habitat or soil preferences are almost never extended to PFGs.

(*continued on next page*)

Table 1 (*continued*)

Name of the approach	Rationale	Modelling use	Scale	Variant (study-specific) or invariant	Advantages	Caveats
Periodic Table of Niches (Winemiller et al., 2015).	Niche classification scheme based on functional redundancy, built on a suite of fundamental ecological niche dimensions and from related functional traits and performance data.		biome, region, landscape	invariant	Supported by convergent evolution that justifies periodicity in the niche space.	Species and traits evolve and change in relation to environmental contingencies.
Trait-flexible modelling (Van Bodegom et al., 2012; Scheiter et al., 2013; Pavlick et al., 2013; Berzaghi et al., 2020).	Describes individual plants by a set of traits, emerging from general correlations and dependencies, such a s the leaf and stem economics spectra (Wright et al., 2004; Baraloto et al., 2010).	Next generation DGVMs (e.g. DGVM2, Scheiter et al., 2013).	landscape, region	variant	Overcomed the simplifying classifications approaches of PFGs/PFTs that are assumed to have constant attribute values across the globe; allows for intra-specific variability and adaptation to the environment.	Data availability; may lead to overly complex models difficult to parametrize and use high computational power

ecological dimensions *sensu* Hérault, 2007; Boulangeat et al., 2012), important habitat filters and interspecific interactions (Lavorel et al., 1997).

Here, we generalise the methodological approach by Boulangeat et al. (2012) by implementing a broader set of traits and species and merge this approach with the periodic table of niches by Winemiller et al. (2015). This provides a data driven generic framework that can classify worldwide trait data to derive PFGs that are suitable for modelling plant biodiversity, vegetation dynamics and ecosystem functioning potentially anywhere on Earth. One foundational step of the proposed framework is the data driven clustering approach based on the generation of a small number of discrete categories from both continuous and categorical traits, thus facilitating the inclusion of new traits if a new or different ecological dimension is required by the model and moreover, new species trait data if new data becomes available. The specific aims of this paper are: i) to describe the conceptual approach of a novel PFG classification framework, its rationale and workflow; ii) to apply the workflow to classify a global set of species based on currently available trait data; and iii) to assess the generated classification in terms of its capacity to adequately summarise functional variation in plant communities globally.

2. Methods

2.1. The conceptual and methodological approach and its application to a global example use-case

The following section provides the details on the conceptual workflow to retrieve PFGs that are not region-specific, based on Boulangeat et al. (2012) and Winemiller et al. (2015) and on its application to a global example use-case. The workflow can in principle be applied to any study region (from a local to a global extent), ecological dimension or species pool of interest. The workflow focuses on the selection of relevant data; the requirements of the data to be subjected to the clustering procedure; the clustering and the validation procedures. The applied workflow is composed of seven main steps (Fig. 1).

2.2. Identification of relevant ecological dimensions and plant characteristics

This generic framework is designed to develop objective PFGs classifications based on the ecological dimensions and the geographical context of interest. The first conceptual step for building the classification scheme is to identify the relevant ecological dimensions and related characteristics (Garnier et al. 2017) for the case study. Ecological dimensions are axes of ecological variation that may comprise one or more traits that are assumed to be ecological equivalents with respect to community dynamics (Westoby et al., 2002; Hérault, 2007, Winemiller et al., 2015). Most implemented ecological dimensions are derived from a few important mechanisms that can be summarised into four categories: temporal dynamics (i.e., demography/life history/metabolism); species interactions (i.e., trophic mechanisms, defense); relation to abiotic conditions or habitats; and spatial dynamics (i.e., dispersal) (Boulangeat et al., 2012, Winemiller et al., 2015).

Previous work highlighted an asymptote of increasing predicting power at four to eight dimensions and suggested that the number of dimensions should be maximised using traits from multiple organs (Laughlin, 2014). The selection criteria for ecological dimensions has to do with the processes of interest to the modelling effort. Depending on the use-case this framework enables a specific selection and number of relevant ecological dimensions and respective traits and/or characteristics combinations, allowing in principle the usage of an unlimited number of traits, characteristics and dimensions.

For the applied framework, we based the selection of main ecological dimensions on Boulangeat et al. (2012), improving it with the inclusion of soil preferences, which is rarely accounted for in PFGs classification while being a key driver of plant communities assembly. The inclusion of soil type preference data was motivated by the fact that often closely related, and thus functionally similar species differ notably in soil preferences (Wherry, 1927). In Boulangeat et al. (2012), the selection of ecological dimensions aims at covering the main mechanisms of community assembly and biogeography, that could then be implemented in dynamic vegetation models. Here, the eight selected ecological dimensions represent temporal dynamics through form, demography and disturbance, spatial dynamics through dispersal, species interaction with competition response and competition effects, and relation to abiotic conditions with habitat and soil preferences. While Sitch et al. (2003) indicates that incorporating climatic zonation to PFTs adds substantial information, we did not include climatic preferences directly, as we assume that plant functional traits alone (considered in the habitat dimension) can adequately predict climatic niches of species (Medeiros et al., 2023).

We then identified the most measured and widely available traits or characteristics associated with these ecological dimensions by searching the TRY Plant Trait Database and selecting the traits with the most species and observations. TRY v.5 was initially queried for 62 widely measured candidate traits that were related to the selected ecological dimensions (TRY variable “TraitID”) (Supplementary Table 1). Overall,

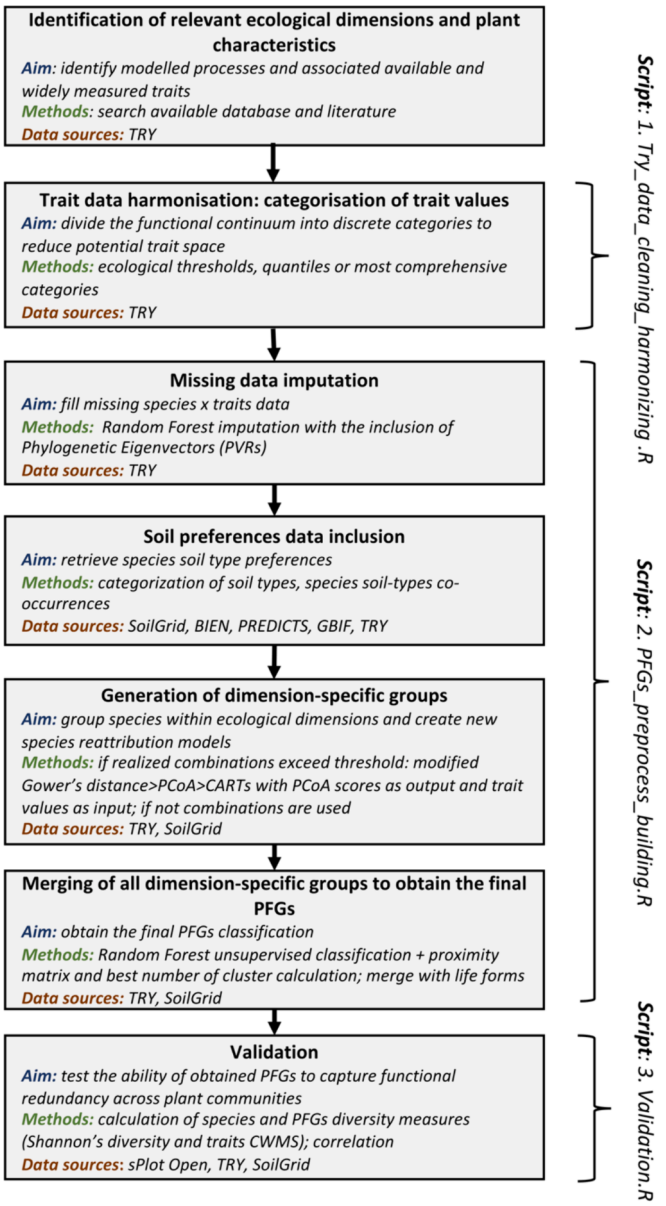


Fig. 1. PFGs building applied workflow with main aims, methods and data sources for each step. Corresponding supplementary R script are indicated. TRY=Plant Trait Database (Kattge et al., 2020); SoilGrid = global gridded soil information (Poggio et al., 2021); BIEN=Botanical Information and Ecology Network (Enquist et al., 2016); PREDICTS: Projecting Responses of Ecological Diversity In Changing Terrestrial Systems (Hudson et al., 2014); GBIF=and Global Biodiversity Information Facility (Telenius, 2011); sPlotOpen = an environmentally balanced, open-access, global dataset of vegetation plots (Sabatini et al., 2021); PCoA=principal coordinate analysis; CARTs = classification and regression trees; CWMs = community weighted means.

167,001 taxa had at least one observation for one of these traits, which summed to 4,399,663 trait observations. Taxa by traits data was requested through the TRY portal (https://www.try-db.org). First, we performed a species nomenclature standardisation procedure using The Plant List (TPL) taxonomic backbone (https://www.theplantlist.org), now a part of the WFO Plant List (https://wfoplantlist.org/plant-list). TPL is still a robust and reliable nomenclature system for plant species with an associated phylogenetic mega tree (Jin and Qian, 2019; 2022). We excluded Bryophytes, due to the lack of coverage in the TRY database. We used the *taxonlookup* R package (Pennell et al. 2016), to match the species names against the TPL names repository (downloaded from: https://github.com/nameMatch/Database) with the *nameMatch* function of the *U.Taxonstand* package (Zhang and Qian, 2023). Any species names with ambiguous or fuzzy matches and hybrid taxa were excluded.

After merging the information from related traits when possible and discarding traits that were not suitable for our purposes (*i.e.*, TraitID: 3096. Species habitat characterization: vegetation type was removed because it only contained data for Brazilian ecosystems) we obtained a dataset with 135,736 species by 49 traits (Supplementary Table 6). To further filter only the most measured traits we retained traits that had more than 10,000 sampled taxa and completed missing traits that are relevant for ecological modelling or for ecosystem services modelling (i.e., leaf phenology, mycorrhizal status, leaf thickness). Our final set of traits comprised 134,608 species and 19 traits: plant growth form, maturity, longevity, species tolerance to shade, leaf area per leaf dry mass (SLA), plant height vegetative, seed dry mass, pollination syndrome, dispersal syndrome, plant woodiness, plant resprouting capacity, seedbank longevity, leaf nitrogen (N) content per leaf dry mass, species tolerance to frost, rooting depth, N fixing capacity, leaf thickness, leaf phenology and mycorrhizas presence.

2.3. Trait data harmonisation: categorisation of trait values

Available trait data to be used for the classification of species in generic functional groups needs to be harmonised and categorised. In order to reduce complexity of trait combinations, continuous traits are divided into discrete ordinal categories (low, medium, high), based on trait data distribution or ecological thresholds/hypotheses when available. All data analyses were performed in R software (R Core Team, 2022).

We carried out a different procedure depending on whether the trait data was continuous, cateogorical, or mixed (Supplementary Table 1). For continuous traits (i.e. leaf area per leaf dry mass (SLA), plant height vegetative, seed dry mass, rooting depth, leaf nitrogen (N) content per leaf dry mass, and leaf thickness) we kept only records with standardised values (TRY variable "StdValue"), checked for outliers with Rosner's test using the function rosnerTest of the R package EnvStats (Millard et al., 2018) and removed extreme species values only when these were single observations and not repeated observations. Then we divided continuous traits into categories based either on ecological hypotheses/thresholds (e.g. for seed mass and height we based our categories on their relationships to dispersal agents and growth forms Westoby et al., 1990; Hughes et al., 1994) or based on trait values distribution into homogenous quantiles (*i.e.* <0.2, 0.2–0.8, >0.8) corresponding to low, medium, and high categories when no ecological threshold hypothesis was available. As most of the traits are normally distributed, dividing their values between the 20 % lowest values (i.e., extremely low values group), 20 % largest values (i.e., extremely high values group), and the 60 % remaining values are the "intermediate" category allows a good representation of extreme groups. For mixed traits that displayed both numerical and categorical values (i.e., maturity, longevity, species tolerance to shade, seedbank longevity, species tolerance to frost, leaf phenology, and mycorrhizas presence) we had to group numerical values into the most comprehensive categorization available across the various database composing TRY. Lastly, categorical traits (i.e., plant growth form, plant woodiness, pollination syndrome, dispersal syndrome, plant resprouting capacity, N fixing capacity) including many different categorizations depending on the database of origin within TRY were harmonised according to the most applicable categorization present in the data. Then, we calculated each trait mean value for each species for ordinal traits (rounding to the closest integer value) and the mode for strictly categorical traits. Finally, to obtain coverage between 10–15 % for the main evolutionary lineages considered (spermatophytes and ferns and allies) we removed all species with more than 15 missing trait values for ferns and allies and more than 14 missing trait values for spermatophytes, ensuring an equal sampling size of both phylogenetically distant lineages.

For data inspection, cleaning and data manipulation, we used rtry R package (Lam et al., 2022), data.table (Dowle et al., 2019) and tidyverse (Wickham and Wickham, 2017). The complete process of trait cleaning and homogenization can be found in the Supplementary R script "1. Try_data_cleaning_harmonizing.R".

2.4. Missing data imputation

Since trait data often contains missing data, we include a missing data imputation protocol that is robust enough for large species-traits datasets containing mixed variables (plant characteristics) that is performed on already categorised trait data, instead of raw continuous values, in order to minimise imputation errors. Our method follows Debastiani et al., (2021) and is based on the Random Forest algorithm (Breiman, 2001) including phylogenetic relatedness information in the form of Phylogenetic Eigenvectors (PVR) (Diniz-Filho et al., 1998).

In the applied framework, in order to calculate phylogenetic relatedness, we pruned a phylogenetic mega tree to our species pool. We downloaded the extended TPL plant mega tree (GBOTB.extended.TPL.tre) (Jin and Qian, 2022) and pruned the phylogeny to our species pool with the phylo.maker function of the V.Phylomaker2 R package (Jin and Qian, 2022), specifying 'scenarios = S3'. To maximise niche conservatisms (Harvey and Pagel, 1991) and ensure less variability between species in terms of overall physiology and phenotype when imputing data, we divided our species pool into eight major phylogenetic lineages (ferns and allies, gymnosperms, basal angiosperms, magnoliids, monocots, asterids, rosids and other eudicots) based on monophyletic clades of the TPL taxonomic backbone phylogenetic tree and generated phylogenetic trees by pruning the mega tree for each lineage. Secondly, we performed separately for each major lineage a PVR decomposition using the PVRdecomp function (PVR R package, Santos et al., 2018) and kept all PVRs explaining more than the 0.05 % of variance. PVRs are the eigenvectors of a Principal Coordinates Analysis (PCoA) fitted on species phylogenetic distances across a phylogenetic tree (Diniz-Filho et al., 1998). Then we added the obtained PVRs to our major lineages' species by traits matrix as if these were additional traits and imputed missing data in traits using the function missForest of the missForest R package (Stekhoven and Stekhoven, 2013) specifying default settings and merged back lineages by traits data into one dataset that comprised 45,009 species and 19 traits.

2.5. Soil preferences data inclusion

We computed and added to the main species by traits dataset a species soil preferences dataset, which assesses the relative abundance of occurrences for each species on 42 soil types. These soil preference profiles describe plant species generalism vs specialisation on certain soils and are based on the intersection of geolocated soil data with plant occurrence data. To obtain soil data, we simplified the soil database soilgrids250m (Poggio et al., 2021), which predicts several key soil properties worldwide at multiple depths with a resolution of 250 m. We used only the 5–15 cm depth layer and extracted the median estimates (Q0.5) of soil bulk density (BD), texture (grain size), pH and organic carbon (OC) content as well as their associated uncertainty estimates. To be able to discard soil estimates that we deemed too unreliable in any of the five soil variables, we sampled 100,000 random soils from the database and assessed the uncertainty distributions. Then, we used an independent representative worldwide sample (n = 100,000) of the soil data to classify soils. We discarded all data points that were above the 0.95 quantile in one of the five uncertainty distributions (23 % of all data), as well as soils with a coarse material fraction >15 %, and then determined the soil texture class according to the U.S. Department of Agriculture definition (12 classes), using the clay, sand, and silt content of the fine earth fraction. The relative proportion of texture classes differed widely, ranging from <0.1 % to 31 %. We joined adjacent classes: Silt was integrated in Silt Loam; Silty Clay in Silty Clay Loam; Sand in Loamy Sand; Sandy Clay in Clay. After merging, eight texture classes remained. Then, to retrieve reasonable texture-pH combinations since there was considerable variation in pH values within each texture class, with most textures showing clear bimodal patterns, we divided most texture classes into two groups according to pH. The resulting 14 texture/pH classes signify differences in abiotic conditions. These classes were further subdivided into biological activity classes. To do so, we split the classes along the 20th and 80th percentile of the organic carbon content distribution of the soils. Splitting along an organic carbon gradient is biologically meaningful, since soils with different organic carbon stocks may harbour different microbial and microfaunal functional groups. In total, we retrieved 42 soil classes. The classification that we derived from the representative sample was then applied to the full (simplified) database, thus obtaining a global dataset of all soil classes. The categorised soil data was then intersected with plant species occurrences obtained from the Botanical Information and Ecology Network (BIEN, Enquist et al., 2016), the Projecting Responses of Ecological Diversity In Changing Terrestrial Systems (PREDICTS, Hudson et al., 2014) and the Global Biodiversity Information Facility (GBIF, Telenius, 2011) databases. We considered only species records collected after the year 1980 to exclude potentially inaccurate historical records and with less than 100 m of coordinate uncertainty. Species occurrences were first screened for duplicates, nomenclatural or spatial errors and then were uniformed to the TPL nomenclature before merging the soil preference traits with the species by traits dataset. Since many species did not have enough occurrences in the databases (more than 50 occurrences) to compute reliable co-occurrences, or soil data were not available at the sampling locations, we obtained a species-soil co-occurrence matrix with 33,010 species. Accordingly, by merging the species by traits dataset we had to reduce our species pool, obtaining a dataset of 19,102 species for 61 traits. Soil preference data was not subjected to imputation. For accessing the BIEN data, we used the R package BIEN (archived version 1.2.4, from CRAN), and for data manipulation we used data.table, vroom, tidyverse, the developer version of scrubr (retrieved on 21 dec 2021, version 0.4.0, ropensci/scrubr), stringr and lubridate. We further used the packages terra, sf, gdalUtils and rworldmap for spatial data manipulation, soiltexture for retrieving soil textures, and MASS and pdfCluster for statistical analysis.

2.6. PFGs classification

Once a complete species by traits dataset is obtained, one can proceed with the ascendent clustering approach that leads to emergent PFGs. The clustering is decomposed into two main tasks: 1) Generation of ecological dimension-specific groups and attribution models; 2) merging all the ecological dimension-specific groups to obtain the final PFGs.

2.6.1. Generation of dimension-specific groups and attribution models

During the first task, for each ecological dimension if the number of unique realised combinations of trait categories exceeds a user-defined threshold, a reduction of each ecological dimension's complexity is carried out by performing a Principal Coordinate Analysis (PCoA) on the distance matrices. This analysis retrieves species scores as relative positions in an ordination space of several traits and characteristics. Then, the grouping of species by their relative position in the ordination space is performed. This allows to obtain dimension-specific groups while constructing a dimension-specific reattribution model. To this aim, a Classification and Regression Trees (CARTs) analysis following Winemiller et al. (2015) is applied. CARTs are built specifying as outputs the species scores on a reduced number of PCoA components (selected based on eigenvalues explanatory power) and using original categorised trait values as explanatory variables to build dendrograms. This, beside identifying meaningful groups, allows the generation of a reattribution model that can include more species when data becomes available.

In the applied framework, we first sorted obtained traits and

characteristics into the eight main ecological dimensions according to their relationships with these main axes of ecological variation (Boulangeat et al., 2012) (Table 2).

We had seven unique growth forms ("shrubs_trees", "herbs", "climber_lianas", "small_shrubs", "epiphytes", "aquatic" and "parasitic"). The form dimension is kept out of the clustering procedure and used as a top level grouping factor of actual PFGs as recommended by several authors (Lavorel et al., 1997; Landsberg, 1999, McIntyre et al., 1999), so that PFGs maintain a conceptual link with top-down PFTs classifications, to make the comparison with other vegetation modelling efforts possible and also be informative on three-dimensional plant growth patterns (ground rooted non-climber vs epiphyte vs ground rooted climber). Moreover, growth form alone is often highly correlated to other functional traits (Box, 1981) and thus including it in the clustering procedure could potentially lead to biassed results.

For each dimension except form, we calculated all the realised combinations of categorised trait values and soil types relative abundances for soil. When realised combinations were less than 50, we assigned each of them to a group while when realised combinations exceeded 50, we performed additional clustering and classifications steps as follows. First, we calculated a modified version of Gower's distance among species, that can handle a mix of categorical and ordinal data (Crossa and Franco, 2004), and exacerbates differences in adjacent ordinal categories by adding a multiplying factor (custom R function, kindly provided by Jonathan von Oppen). We set the multiplying factor to 10 and assigned weights to each trait to maximise the impact in the distance matrix of more important traits in current ecological modelling frameworks. We then performed a Principal Coordinates Analysis (PCoA) with the pcoa function of the ape R package (Paradis et al., 2019) on the resulting distance matrices and inspected the eigenvalues bar plot applying the broken stick criterion (Peres-Neto et al., 2005) to decide how many components scores to retain for each dimension. Finally, we fitted classification and regression trees (CARTs) with the mvpart function of the mvpart R package (De'Ath, 2007) using PCoA scores as outputs and traits values as inputs to cluster species based on their proximity in the ordination space but also to generate a model that we can use to reattribute new species to existing dimension-specific groups.

2.6.2. *Merging all the ecological dimension-specific groups to obtain the final PFGs*

In the second task, dimension-specific groups, either unique combinations of traits values or groups outputted by the CARTs, are then merged to build the comprehensive PFGs classification with an unsupervised Random Forest classification that is used to generate a proximity matrix. The unsupervised Random Forest algorithm does not impose any predefined hierarchical order while merging all dimension-specific groups but instead assesses *a posteriori* the most important dimensions in creating the PFGs. To minimise the subjectivity of clustering, the proximity matrix is used to calculate the optimal number of clusters with a statistical test. Finally, obtained clusters are combined with the growth forms dimension that was not included into the Random Forest algorithm to retrieve the actual PFGs.

The species by dimension-specific groups dataset excluding growth forms was used to perform the Random Forest unsupervised classification. We used the randomForest function of the randomForest R package (Liaw and Wiener, 2002), specifying proximity = TRUE and 10,000 trees. The proximity matrix was then used to calculate the optimal number of clusters with the function Nbclust of the Nbclust R package (Charrad et al., 2014) specifying a maximum of 200 clusters, "ward.D2" as clustering method and "cindex" as preferred index. The concordance index, or C-index (Hubert and Levin, 1976) is a commonly used metric that evaluates the performance of a predictive statistical model in discriminating between cases with different outcomes or clusters (Blanche et al., 2019). The lower value of the C-index corresponds to the best number of clusters. We then combined the obtained best partition with the seven growth forms to obtain the final PFGs. Moreover, we calculated variable importance in terms of mean decrease in Gini Index (Gini, 1955) to assess the different contributions of the different dimensions. The mean decrease in Gini Index is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini Index, the higher the importance of the variable in the model (Martinez-Taboada and Redondo, 2020).

Lastly, we derived PFGs traits calculating the average values and the standard deviation of numerical traits and the modes of categorical traits. Since often TRY species observations for the trait longevity do not indicate maximum lifespan but rather recorded individuals age, to minimise possible errors in long-lived species groups, PFGs longevity was calculated as the maximum observed longevity value for shrubs and trees PFGs and as the average value for all other growth forms. The complete processes of missing data imputation, the creation of each ecological dimension-specific groups and attribution models and the merging of all dimension-specific groups to obtain the final PFGs can be found in the Supplementary R script "2.PFGs_preprocess_building.R" (Supplementary Material).

Table 2
Selected ecological dimensions and associated traits. Extracted from the TRY database.

Dimension	Trait
FORM	plant growth form
DEMOGRAPHY	plant lifespan, plant maturity
COMPETITION RESPONSE	plant nitrogen (N) fixation capacity, species tolerance to shade
COMPETITION EFFECT	leaf area per leaf dry mass; plant height vegetative; seed dry mass
DISPERSAL	pollination syndrome; dispersal syndrome
DISTURBANCE	woodiness, plant resprouting capacity, seedbank longevity, mycorrhiza status
HABITAT	leaf phenology type, leaf thickness, leaf nitrogen (N) content per leaf dry mass, species tolerance to frost, rooting depth
SOIL	soil types relative abundance of occurrences data

2.7. *Validation*

The validation step of the generated PFGs follows the PFGs validation method proposed by Boulangeat et al. (2012). Boulangeat et al. (2012) argue that since PFGs trait selection and grouping parameters are a simplification of reality, it is important to assess how much information is actually lost during the process. This validation method tests the ability of PFGs to summarise plant diversity across plant communities by comparing and correlating species-based and PFG-based taxonomic and functional diversity metrics calculated from plant community data. If PFG-level metrics are positively correlated with species-level metrics, then the PFG classification is robust and can be effectively used for plant communities modelling (Boulangeat et al., 2012).

To validate the PFGs outputted by our applied framework we downloaded the open access sPlotOpen, the biggest public global vegetation database, containing plant species co-occurrence data for 95,104 vegetation plots (Sabatini et al., 2021). First, we aligned the sPlotOpen species nomenclature to the TPL nomenclature to ensure compatibility, as performed for the TRY data. Second, we excluded all plots where less than 70 % of occurring species were present in our PFGs classification, retaining 64,224 plots. Next, we averaged the species cover values within each PFG for each plot to obtain a separate plot by PFGs abundance matrix. We calculated Shannon's diversity index for both matrices and Community Weighted Means (CWMs) metrics for all PFGs functional traits, extracting species values from our cleaned and categorised TRY PFGs building dataset. Prior to indices computation, all species that were not present in the PFGs were excluded from plot data.

CWMs are calculated as the mean value of trait values weighted by species/PFGs relative abundances within a sample (Roscher et al., 2012)

and relate to the functional identity of plant communities by focusing on the contribution of dominant species (McLaren and Turkington, 2010). This measure has been shown to be tightly linked to ecosystem functioning (Boulangeat et al., 2012 and references therein). Shannon's diversity was calculated using the diversity function of the R package vegan (Oksanen et al., 2007), while CWMs for ordinal traits were computed calculating the weighted mean of traits by either percent cover or average cover in the case of species and PFGs respectively. CWMs for categorical traits were computed by first converting each trait category to a dummy variable and then computing the CWM for each new dummy variable. Correlations between species and PFGs indices were calculated as Pearson's coefficients with the *corr.test* function of the psych R package (Revelle and Revelle, 2015). The complete process of validation can be found in the R script "3.Validation.R" (Supplementary Material).

3. Results

3.1. Species coverage and inputed data coverage

The here employed pool of 19,102 species represented about 5.45 % of the 350,699 accepted species names in the TPL database (https://www.theplantlist.org), and included 4,493 genera, 380 families and 83 orders out of the 17,020 accepted genera, 642 accepted families and 85 listed orders in the TPL database. Species coverage varied greatly among families, spanning from 100 % in some monospecific families (e.g. Psilotaceae, Ginkgoaceae, and Welwitschiaceae) to values below 1 % for some, mostly tropical herbaceous or epiphytic, species families (e.g. Gesneriaceae, Cyclanthaceae, Begoniaceae, Orchidaceae). Species coverage in species rich families such as Asteraceae, Fabaceae, Rubiaceae and Poaceae was overall lower than 10 %, with 4,98 %, 7.38 %, 3.41 % and 10 % respectively. By comparing the native distributions of employed taxa found in the World Checklist of Vascular Plants (WCPV, https://powo.science.kew.org/about-wcvp) with all recorded native distributions in the WCPV, coverage by continent was the highest for North America and Europe and the lowest for tropical Asia, Antarctica and the Pacific. Additional information on species coverage can be found in Supplementary Fig. 1 and Supplementary Tables 2 and 3.

Out of the obtained 1,165,222 cells in the species by features matrix, 362,938 contained species TRY traits data, of which 214,145 contained imputed data (59 %). Only 135 species did not contain any missing entry. The highest proportion of missing data (>80 %) across all traits in species rich families was recorded for several Fern and allies families (e.g. Cyatheaceae, Pteridaceae and Selaginellaceae). Conversely, species rich families with the lowest missing data proportion were Juglandaceae, Pinaceae and Betulaceae. Additional details on relative proportions of imputed data by lineages can be found in Supplementary Table 4.

3.2. PFGs features

We obtained 12 unique realised combinations for the demography dimension, 11 for the competition response dimension, 23 for the dispersal dimension and 46 for the disturbance dimension. Other dimensions than the soil dimension had a number of realised unique combinations higher than 50 such as competition effect: 85, and habitat: 132. After further grouping the large number of unique combinations we obtained 5, 46 and 11 groups for the competition effect, habitat and soil dimensions, respectively.

When merging dimension-specific groups to obtain the final PFGs classification, the Nbclust function identified 182 groups as the best partition for our data. We combined the obtained best partition scheme with the seven growth forms retrieving an optimal number of 465 PFGs, of which 27 were composed by aquatic species, 72 climbers_lianas, 26 epiphytes, 150 herbs, 8 parasitic, 125 shrubs_trees, and 57 were composed by small_shrubs species. Overall, 221 PFGs contained more than 10 species with the three most numerous groups containing 2111, 1 938 and 1235 species respectively (Table 3, Supplementary Tables 5,6). The ten most species rich PFGs (Table 3) comprised mostly herbs, shrubs trees and only one climber and lianas group and belonged to mostly 2 clusters of characteristics values (cluster number 3 and number 5). All 10 most species rich PFGs displayed generalism in soil preferences, occurring on all soil types. Ninety-four PFGs contained only one species (Supplementary Tables 5, 6). The most important dimensions that emerged in the Random Forest classification were dispersal, habitat, disturbance and competition response. The least important were demography, competition effect, and soil (Table 4).

The obtained PFGs in general represented non-overlapping combinations of characteristics within each of the ecological dimensions divided by growth forms (Fig. 2).

The most influential traits highlighted in Fig. 2, underline the two main uncorrelated gradients in PFGs distribution in the traits space. These are the one including SLA, leaf_N leaf_thickness and frost_tolerance, and the one including height, woodiness, and seed mass. Regarding the most species rich PFGs, for shrubs_trees, small_shrubs, climbers and herbs the most numerous PFGs belonged to the two same clusters (number 5 and number 3). This means that most species rich PFGs of these growth forms shared very similar combinations of traits. Aquatic species-richest group belonged to one of these clusters. Epiphytes and parasites' species-richest groups belonged to completely different clusters (Fig. 3).

3.3. PFGs validation

Overall, the validation process highlighted strong correlations between species-level and PFGs-level diversity indices, which suggests that our classification was robust and summarised adequately the main diversity trends (Table 5, Supplementary Fig. 2). Functional identity (CWMs) was successfully captured after the reduction of plant species into the 465 PFGs and there were significant positive correlations between species-level and PFG-level CWMs values for most of the traits used for building our PFGs classification (Table 5), suggesting that the number of obtained PFGs is sufficient to capture the main features of plant communities worldwide. Lastly, our classification of 465 PFGs also captured much of the variation of taxonomic diversity across plots (Pearson's correlation = 0.95, $p < 0.001$).

4. Discussion

Our research presents a theoretical and applied framework to create a usable number of discrete and globally valid PFGs that are detailed enough to capture the functional divergence between species and the key ecological dimensions underpinning vegetation diversity, and hence are suitable for use in large scale dynamic vegetation models. We described the conceptual and applied approach of the framework, its rationale and workflow. The approach relies on first, dividing the functional continuum (Westoby et al., 2002) into discrete categories to reduce the potential trait space and to accommodate many more species (without increasing the number of PFGs) if their traits values fall into the identified trait values ranges. We further grouped these trait combinations by their functional proximity to retrieve the final PFGs. Lastly, we validated the outcomes of the applied framework using an extensive dataset of plant communities distributed globally, highlighting its potential in capturing functional redundancy patterns worldwide.

Using the most comprehensive and currently available global trait database of plant functional traits (TRY, Kattge et al., 2020) and soil preference data, we assigned 19,102 plant species to 465 distinct PFGs based on a set of 19 functional traits with two to seven modalities each, and on preference data for 42 soil types. This number is higher than that obtained in other existing PFG/PFT classifications (i.e., in the order of tens) that are coarser or constrained to a specific region (e.g., Sitch et al., 2003, Boulangeat et al. 2012, Pierce et al., 2017). Also, the number of

Table 3
10 most species rich plant functional groups (PFGs) and associated trait values ranges. Categories thresholds values can be found in the Supplementary Table 1. spp = number of species; dispers. = dispersal syndrome; pollin. = pollination syndrome; leaf phen. = leaf phenology; leaf N = leaf nitrogen (N) content; SLA = specific leaf area; myco. = mycorrhizas presence; N fix. = N fixation. Soil preferences is not displayed as all 10 most species rich groups occur on all soil types.

PFG	spp	dispers.	pollin.	leaf phen.	leaf N (mg/g)	maturity (yrs)	seed mass (mg)	frost tolerance	height (m)	SLA (mm^2/mg^{-1})	seedbank longevity (yrs)	myco.	woodiness	leaf thickness (mm)	longevity (yrs)	root depth (m)	shade tolerance	N fix.	resprouting
shrubs trees_5	2111	zoochory	animal	evergreen	12.9–27.07	0–5	2–100	no	5–15	8.08–25.4	1–5	yes	woody	0.17–0.31	100+	<1	50–75	no	yes
herbs_5	1938	zoochory	animal	deciduous	12.9–27.07	0–5	0.1–2	yes	0.1–0.5	8.08–25.4	1–5	yes	non-woody	0.17–0.31	1–2	<1	25–50	no	no
herbs_3	1235	unassisted	animal	deciduous	12.9–27.07	0–5	0.1–2	yes	0.1–0.5	8.08–25.4	1–5	yes	non-woody	0.17–0.31	1–2	<1	0	no	no
shrubs trees_3	753	anemochory	animal	evergreen	12.9–27.07	0–5	2–100	no	0.5–5	8.08–25.4	1–5	yes	woody	0.17–0.31	100+	<1	0	no	no
herbs_25	589	hydrochory	animal	deciduous	12.9–27.07	0–5	0.1–2	yes	0–0.1	8.08–25.4	1–5	yes	non-woody	0.17–0.31	1–2	<1	0	no	no
herbs_54	421	anemochory	abiotic	deciduous	12.9–27.07	0–5	0.1–2	yes	0.1–0.5	8.08–25.4	1–5	yes	non-woody	0.17–0.31	2–100	<1	25–50	no	no
shrubs trees_49	356	zoochory	animal	evergreen	12.9–27.07	0–5	2–100	no	5–15	8.08–25.4	1–5	yes	woody	0.17–0.31	2–100	<1	75–100	no	no
herbs_28	316	hydrochory	animal	deciduous	12.9–27.07	0–5	0.1–2	yes	0.1–0.5	8.08–25.4	<1	yes	non-woody	0.17–0.31	2–100	<1	0	no	no
climber lianas_5	254	zoochory	animal	evergreen	12.9–27.07	0–5	2–100	no	0.5–5	8.08–25.4	<1	yes	semi-woody	0.17–0.31	1–2	<1	50–75	no	no
shrubs trees_22	230	zoochory	animal	evergreen	12.9–27.07	5–10	2–100	no	5–15	8.08–25.4	1	yes	woody	0.17–0.31	100+	1–5	75–100	no	yes

Table 4
Unsupervised random forest variable importance for each ecological dimension. Importance is expressed as mean decrease in Gini Index.

Dimension	Importance
dispersal	2465.52
habitat	2448.23
disturbance	2375.26
competition response	1791.27
demography	1573.78
competition effect	933.86
soil	95.22

species is higher that what usually included in PFTs/PFGs classifications (e.g. Boulangeat et al., 2012; Pierce et al., 2017), but lower than the 46,047 species examined in Diaz et al. (2016)'s global form and function spectrum. Given that our PFGs classification framework is global, which means that it can potentially represent all plant species and work seamlessly across different ecosystems, we believe that 465 is a reasonable maximum number of entities for varying scale plant ecological modelling. This number represents a good, computationally sound compromise between the species level (in the hundreds of thousands) and the PFTs' level (in the order of tens).

The here proposed framework allows in principle the inclusion of an unlimited number of plant features and ecological dimensions. However, in the presented global example-use case, trait selection was carried out to ensure adequate coverage of important ecological dimensions (species interactions, habitat affinities, spatial and temporal dynamics), that are involved in key processes determining species distribution in space and species dynamics in time. Such processes are those found in ecological models, ecological theories and therefore important for ecological analyses. To limit our selection to an informative set of traits and features, we gathered and retained only those with the most observations by taxa available in public global repositories. For example, among selected traits, we did not include direct climate preferences, as in Sitch et al. (2003), or drought tolerance as we assumed that these would be correlated with the other functional traits used in the classification. A different selection of features is entirely possible within the proposed framework and would probably lead to a partially different PFGs classification.

The use of growth form as a highest level of our PFGs classification was necessary to ensure that subsequent nested species groupings were based on functional traits that ranged over a similar scale (e.g. plant height is not comparable between tree and herbs). Having PFGs divided by growth forms is also useful for the comparison with functional groups or types derived via other means, in order to relate suites of traits to specific plant behaviours (Lavorel et al., 1999, Symstad, 2002). Furthermore, it would allow to include 3d growth patterns (e.g., in landscapes planning, or projects aiming at combining urban ecology and architecture such as ECOLOPES (Weisser et al., 2023), and to assimilate our classification to the PFTs that are currently used in global scale DGVMs (Sitch et al., 2003).

Our PFGs classification framework was validated by testing its ability to capture essential functional and taxonomic variations among plant communities without losing an excessive amount of information. The high positive correlation between most of the CWMs of species traits vs the CWMs of PFGs traits confirmed that the framework largely achieved this aim. In particular, six quantitative traits and two dispersal modalities showed correlation > 0.6 (Table 5). Among these, the dispersal modalities correlation is reflected by the high importance in Random Forest classification of the dispersal dimension. Also, many of the traits with the highest correlation coefficients belonged to the disturbance dimension, which also had a high importance value in the Random Forest classification. Only the longevity CWM showed a low positive correlation value. This may reflect the difficulty in retrieving realistic long-lived PFGs longevity values and may be partially due to errors in

Fig. 2. Principal Coordinates Analysis (PCoA) of plant functional groups (PFGs) in the plant characteristics space (including functional traits and soil preferences), symbols are PFGs centroids. Arrows show the gradient of the seven fitted variables (traits) most correlated ($r^2 > |0.2|$) with the first two ordination axes: woodiness, frost tolerance, height, seed_mass, SLA, leaf_N, and leaf_thickness. Bigger symbols represent the two most species rich PFGs for each growth form. SLA=specific leaf area; leaf_N=leaf Nitrogen content.

the used trait data that were further propagated throught the missing data imputation procedure.

The distribution of obtained PFGs in the plant characteristics ordination space overall resembles Diaz et al.'s (2016) spectrum of plant form and function. However, one unexpected result is the positive relationships of higher frost tolerance values with herbaceous species, while reasonably expecting woody species to exhibit a higher frost tolerance. This may be related to the inclusion of many tropical tree and shrubs species in our species pool not balanced by as many tropical grasses that are underrepresented in functional traits data. In our classification, across growth forms, the most numerous PFGs belonged to the same dimension cluster for the climber_lianas, herbs, small_shrubs and shrubs_trees growth forms and partially for the aquatic growth form. This suggests that these few particular combinations of features are the most common, and perhaps successful, in terrestrial plants adapted to similar environments, while different combinations of features are more successful in plants that are adapted to different living conditions, such as epiphytes, parasitic plants and in part aquatic plants. Moreover, the most numerous PFGs were often composed of several species belonging to few or related genera. This is consistent with evolutionary histories that make species of the same phylum share overall trait values (Peterson, 2011). This result is likely inflated by trait imputation, that in our case involved overall more than half of our data entries (59 %) and for several ferns and allies families up to ca. 90 % of entries, and may

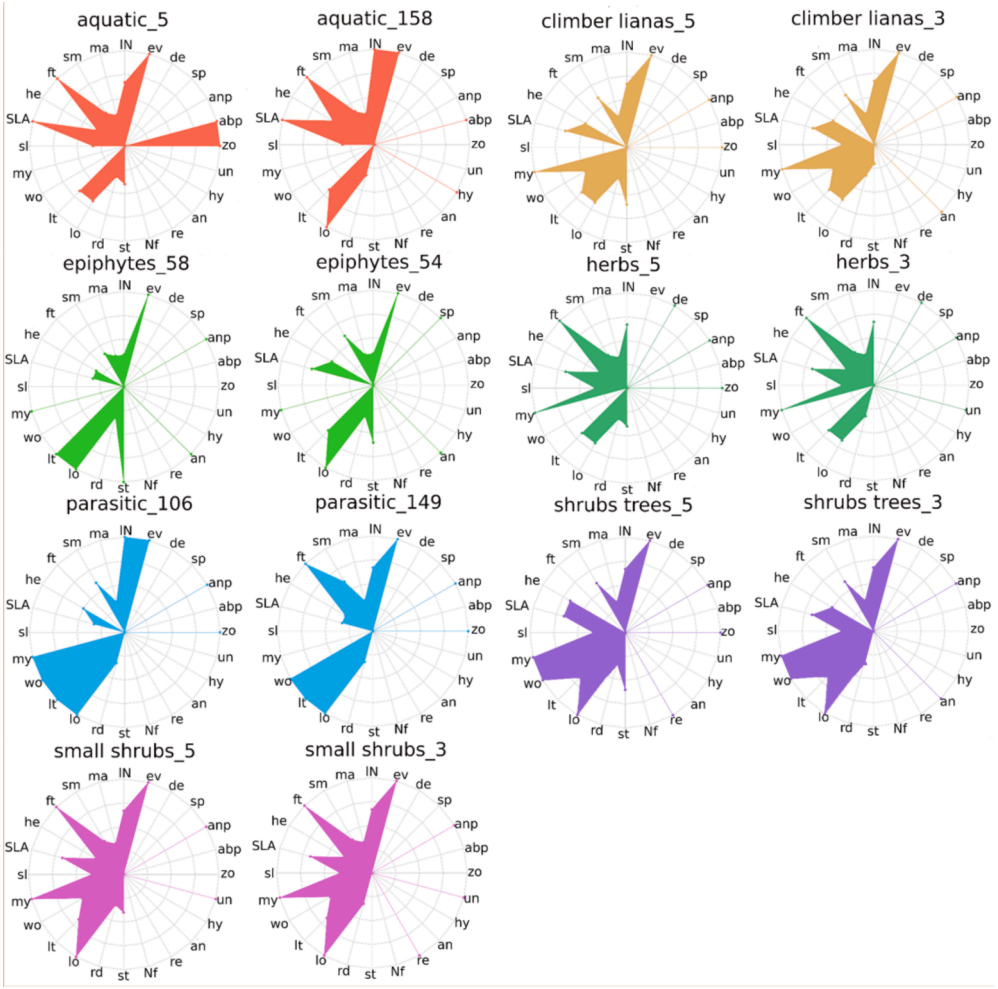


Fig. 3. Radar charts of the scaled values of traits (soil preferences not included) of the two most numerous PFGs for each considered growth form. Representative species are species that were selected as examples. Aquatic_5 (14 spp; representative sp: *Azolla filiculoides* Lam.); aquatic_158 (11 sp: representative sp: *Vallisneria spiralis* L.); climber lianas_5 (254 spp; representative sp: *Lonicera caprifolium* L.); climber lianas_3 (119 spp; representative sp: *Luffa acutangula* (L.) Roxb.); epiphytes_58 (26 spp; representative sp: *Tillandsia polystachia* (L.) L.); epiphytes_54(9 spp; representative sp: *Microgramma lycopodioides* (L.) Copel.); herbs_5 (1398 spp; representative sp: *Bidens pilosa* L.); herbs_3 (1235 spp; representative sp: *Viola tricolor* L.); parasitic_106 (12 spp; representative sp: *Oryctanthus cordifolius* (C.Presl) Urb.); parasitic_149 (9; representative sp: *Phoradendron piperoides* (Kunth) Trel.); small shrubs_5 (104 spp; representative sp: *Genista sagittalis* L.); small shrubs_3 (59 spp; representative sp: *Juniperus horizontalis* Moench); shrubs trees_5 (2111 spp; representative sp: *Sterculia apetala* (Jacq.) H.Karst.); shrubs trees_3 (753 spp; representative sp: *Banksia integrifolia* L.f.). lN=leaf nitrogen; ev = evergreen; de = deciduous; sp = self-pollinating; anp = animal pollination; abp = abiotic pollination; zo = zoochory; un = unassisted dispersal; hy = hydrochory; an = anemochory; re = resprouting; Nf = nitrogen fixing; st = shade tolerance; rd = root depth; lo = longevity; lt = leaf thickness; wo = woodiness; my = mychorrizas; sl = seed longevity; SLA=specific leaf area; he = height; ft = frost tolerance; sm = seed mass; ma = maturity.

partly reflect the greater sampling efforts in traits databases on easily accessible, well-studied, species-rich genera or species complexes. Conversely, many PFGs contained only a single species. In this case, separately for each growth form, species that belong to these PFGs are different enough to be single groups, and the fact that there is no other species with the same trait values ranges is a sign of functional originality. This may also be exacerbated by a scarce sampling effort for species with similar trait data.

With regards to other PFGs/PFTs classifications (Table 1), our PFGs retain a strong conceptual and methodological connection with what proposed by Boulangeat et al. (2012) and Winemiller et al. (2015). However, contrarily to Boulangeat and the majority of classifications listed in Table 1, our PFGs are suitable to be modelled at a variety of scales and extents simultaneously. Also, the here produced PFGs are not only based on dominant species (as in Boulangeat et al., 2012), endorsing the importance of rare species and possibly functional rarity in PFGs/PFTs classifications. Moreover, in contrast to Winemiller et al. (2015), we did not combine ecological dimensions specifying a hierarchical order, allowing the emergence of the more important ecological dimensions from the classification procedure.

Finally, our PFGs encompass a larger set of traits that other globally applicable PFGs/PFTs classifications such as Sitch et al. (2003)'s, CSR, LHS, and even Diaz et al., (2016)'s. This means that pivotal ecological processes such as dispersal, pollination and below-ground interactions are not implicitly inferred through correlations but rather explicitly addressed with the inclusion of relevant traits (e.g. the presence of mycorrhizal associations and dispersal or pollination syndromes).

4.1. Perspectives and further developments

Within PFGs, species share a suite of traits that can be adaptive to a set of environmental conditions (Symstad, 2002) and can point towards assembly mechanisms for specific plant communities (Kindscher and Wells, 1995; Weiher et al., 1998). The PFGs provided here can aid in understanding the mechanisms behind ecological relationships, or the lack thereof, and between functional diversity and ecosystem

Table 5
Correlation between PFGs-level and species-level CWMs for all used quantitative traits and qualitative traits modalities across sPlotOpen data (Sabatini et al., 2021). All correlation coefficient were statistically significant ($p < 0.001$). leaf_N=leaf nitrogen content; sla = specific leaf area. (N) number of observations: 64,224.

trait	trait modality	Pearson's r
leaf N	–	0.49
maturity	–	0.7
seed mass	–	0.63
frost tolerance	–	0.46
height	–	0.76
sla	–	0.44
seed bank longevity	–	0.65
mycorrhizas	–	0.66
woodiness	–	0.93
leaf thickness	–	0.32
longevity	–	0.17
root depth	–	0.36
N fixation	–	0.49
resprouting	–	0.52
dispersal syndrome	anemochory	0.48
dispersal syndrome	hemerochory	0.31
dispersal syndrome	hydrochory	0.73
dispersal syndrome	myrmecochory	0.6
dispersal syndrome	unassisted	0.26
dispersal syndrome	zoochory	0.48
pollination syndrome	abiotic pollination	0.52
pollination syndrome	animal	0.49
leaf phenology type	deciduous	0.34
leaf phenology type	evergreen	0.4
leaf phenology type	semi deciduous	0.36

functioning by highlighting the most important traits. Moreover, these PFGs can be used to understand community assemblage mechanisms and to functionally characterise plant communities, becoming indicators of peculiar or rare trait combinations and thus ecological rarity within plant communities in the light of conservation strategies. Our PFGs may have application in: i) general ecological/botanical studies e.g., studying trade-offs between dispersal strategies and local adaptation, defining if a species is a generalist or a specialist; ii) ecosystems classification and mapping; e.g., to build ecological maps of vegetation for modelling animal distribution; iii) ecological restoration and conservation strategies; e.g., to efficiently plan the restoration of a degraded habitat it is wise to use PFGs diversity instead of just species diversity to ensure a variety of ecological functions (Cadotte et al., 2011); iv) Hybrid-DVMs, that are models combining Habitat Suitability Models (HSMs) and small-scale process-based models (Gallien et al., 2010), modelling single species or functional groups of intermediate complexity to simulate regional/landscape-wide processes as FATEHD (Boulangeat et al., 2014) and the ECOLOPES ecological model (Joschinski and Culshaw, unpublished, Weisser et al., 2023).

The approach presented here did not consider trait variability in space due to intragroup or even intraspecific trait variability, species plasticity and local adaptations (see Berzaghi et al., 2020). Those are, by hypothesis, considered relatively low compared to inter-group and inter-specific trait variability. Our framework also assumes that trait values do not change over time and will still be valid under future environmental conditions (Clark and Gelfand, 2006). Further development to adapt PFGs to trait-flexible models might offer even more opportunities for predictive approaches. However, considering more intraspecific variation is still strongly limited by data availability. While the analysis here is based on currently available data and thus limited to it, the framework is adapted to periodical updates as soon as more trait data becomes available, or trait database errors are corrected. The employed species pool also is highly biassed according to the different sampling efforts by taxonomic groups, geographical areas, growth form and evolutionary lineages inherent to the employed trait database. The representativity of our PFGs is thus directly related to the ability of this species pool and plant characteristics data, given its quality, to represent the real functional variability worldwide, which is not possible to evaluate. However, the same framework can be applied in any sub-region which may be better documented. Moreover, our choice of setting the limits of clustering procedures to 50 groups or 200 for within ecological-dimensions traits combination and PFGs random forest classification respectively, is of course, arbitrary. Nevertheless, our general approach allows for less stringent maximum number of combinations thus for more functional groups as soon as computational power increases.

On a global scale and for design or ecological restoration purposes many additional species features may become relevant, such as species native range, their invasive potential, or their conservation status. Here, we did not include any information on species distribution in the classification to focus on the intrinsic characteristics of species, however it would be interesting to carry out future studies investigating how such information may affect PFGs classification. The native/exotic status would be relative to the geographical area of interest and not inherent to the species themselves so it should be considered *a posteriori* in the screening of the species within each PFGs. Similarly, additional plant features that are also relevant in the context of ecological restoration, urban environment modelling or planning, such as the allergenicity, poisonous or medicinal/gastronomical properties, or the aesthetic features of species should be considered *a posteriori* as filters for species selection within PFGs. For now, our classification does not account for such traits and to our knowledge the available data for these traits is sparse and sometimes difficult to obtain. Nevertheless, our framework allows, in principle, for their inclusion, either as foundational PFGs traits if needed, or as *a posteriori* filter. Finally, by further processing the outputs of ecological models implementing our PFGs classification or the classification itself, one could potentially derive some important ecosystem services related to the used set of traits (including supporting, provisioning, regulating and cultural ecosystem services) across different plant communities and assemblages that can be instrumental in ecological restoration and urban planning strategies.

5. Conclusions

We developed and tested a generic framework that builds case-specific PFGs that consider a limited set of informative, widely measured traits and arise from global functional redundancy patterns. Our robust and globally applicable PFGs can be implemented to model plant responses and community dynamics in different ecosystems and even at a global scale. In principle, our classification framework allows to retrieve functional groups that maintain a high level of detail and characterization of response to abiotic conditions, species interactions, and spatial and temporal dynamics, something that no other approach allows in the present time. Furthermore, the ability of the proposed PFGs classification framework to provide effective functional groups for different ecological modelling efforts can be tested with different selections of species, ecological dimensions and species traits and features.

CRediT authorship contribution statement

M. Calbi: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **G. Boenisch:** Writing – review & editing, Methodology, Data curation. **I. Boulangeat:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **D. Bunker:** Writing – review & editing, Data curation. **J.A. Catford:** Writing – review & editing, Data curation. **A. Changenet:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **V. Culshaw:** Writing – review & editing, Methodology, Conceptualization. **A.S. Dias:** Writing – review & editing, Data curation. **T. Hauck:** Writing – review & editing, Investigation, Conceptualization. **J. Joschinski:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **J. Kattge:**

Writing – review & editing, Supervision, Methodology, Data curation. **A. Mimet:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **M. Pianta:** Writing – review & editing, Methodology, Data curation. **P. Poschlod:** Writing – review & editing, Data curation. **W.W. Weisser:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **E. Roccotiello:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The study has been supported by the TRY initiative on plant traits (http://www.try--db.org). The TRY initiative and database is hosted, developed, and maintained by J. Kattge and G. Bönisch (Max--Planck--Institute for Biogeochemistry, Jena, Germany). TRY is/has been supported by DIVERSITAS, IGBP, the Global Land Project, the UK Natural Environment Research Council (NERC) through its program QUEST (Quantifying and Understanding the Earth System), the French Foundation for Biodiversity Research (FRB), and GIS "Climat, Environnement et Société" France. The authors would like to gratefully acknowledge the help of Jonathan von Oppen and Angelino Carta and the support received from authors' institutions. We also gratefully acknowledge the funding of the EU H2020 FET-OPEN project ECOLOPES (GRANT AGREEMENT NUMBER 964414).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2024.112370.

References

Baraloto, C., Timothy Paine, C.E., Poorter, L., Beauchene, J., Bonal, D., Domenach, A.M., Chave, J., 2010. Decoupled leaf and stem economics in rain forest trees. Ecol. Lett. 13 (11), 1338–1347. https://doi.org/10.1111/j.1461-0248.2010.01517.x.

Berzaghi, F., Wright, I.J., Kramer, K., Oddou-Muratorio, S., Bohn, F.J., Reyer, C.P., Sabaté, S., Sanders, T.G.M., Hartig, F., 2020. Towards a new generation of trait-flexible vegetation models. Trends Ecol. Evol. 35 (3), 191–205. https://doi.org/10.1016/j.tree.2019.11.006.

Blanche, P., Kattan, M.W., Gerds, T.A., 2019. The c-index is not proper for the evaluation of-year predicted risks. Biostatistics 20 (2), 347–357.

Boulangeat, I., Philippe, P., Abdulhak, S., Douzet, R., Garraud, L., Lavergne, S., Lavorel, S., Van Es, J., Vittoz, P., Thuiller, W., 2012. Improving plant functional groups for dynamic models of biodiversity: at the crossroads between functional and community ecology. Glob. Chang. Biol. 18 (11), 3464–3475. https://doi.org/10.1111/j.1365-2486.2012.02783.x.

Boulangeat, I., Damien, G., Thuiller, W., 2014. FATE-HD: A spatially and temporally explicit integrated model for predicting vegetation structure and diversity at regional scale. Glob. Chang. Biol. 20, 2368–2378. https://doi.org/10.1111/gcb.12466.

Box, E.O., 1981. Macroclimate and plant forms: an introduction to predictive modeling in phytogeography. In: Tasks for Vegetation Science series. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-8680-0.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cadotte, M.W., Carscadden, K., Mirotchnick, N., 2011. Beyond species: functional diversity and the maintenance of ecological processes and services. J. Appl. Ecol. 48 (5), 1079–1087. https://doi.org/10.1111/j.1365-2664.2011.02048.x.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. NbClust: an R package for determining the relevant number of clusters in a data set. J. Stat. Softw. 61 (6), 1–36. https://doi.org/10.18637/jss.v061.i06.

Clark, J.S., Gelfand, A.E., 2006. A future for models and data in environmental science. Trends Ecol. Evol. 21 (7), 375–380. https://doi.org/10.1016/j.tree.2006.03.016.

Crossa, J., Franco, J., 2004. Statistical methods for classifying genotypes. Euphytica 137, 19–37.

De'Ath, G. 2007. The mvpart package. http://cran. r-project. org/doc/packages/mvpart.pdf.

Debastiani, V.J., Bastazini, V.A.G., Pillar, V.D., 2021. Using phylogenetic information to impute missing functional trait values in ecological databases. Eco. Inform. 63, 101315 https://doi.org/10.1016/j.ecoinf.2021.101315.

Díaz, S., Kattge, J., Cornelissen, J.H., Wright, I.J., Lavorel, S., Dray, S., Reu, B., Kleyer, M., Wirth, C., Prentice, I.C., Garnier, E., Bönisch, G., Westoby, M., Poorter, H., Reich, P.B., Moles, A.T., Dickie, J., Gillison, A.N., Zanne, A.E., Gorné, L. D., 2016. The global spectrum of plant form and function. Nature 529, 167–171. https://doi.org/10.1038/nature16489.

Diniz-Filho, J.A.F., Sant'Ana, C.E.R.D., Bini, L.M., 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52 (5), 1247–1262. https://doi.org/10.1111/j.1558-5646.1998.tb02006.x.

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X. 2019. Package 'data. table'. Extension of 'data. frame, 596.

Enquist, B. J., Condit, R., Peet, R. K., Schildhauer, M., Thiers, B. M. 2016. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity PeerJ Preprints e2615v2. https://doi.org/10.7287/peerj.preprints.2615v2.

Foley, J.A., Prentice, I.C., Ramankutty, N., Levis, S., Pollard, D., Sitch, S., Haxeltine, A., 1996. An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. Glob. Biogeochem. Cycles 10 (4), 603–628. https://doi.org/10.1029/96GB02692.

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I., Thuiller, W., 2010. Predicting potential distributions of invasive species: where to go from here? Divers. Distrib. 16 (3), 331–342. https://doi.org/10.1111/j.1472-4642.2010.00652.x.

Garnier, E., Laurent, G., Bellmann, A., Debain, S., Berthelier, P., Ducout, B., Roumet, C., Navas, M.L., 2001. Consistency of species ranking based on functional leaf traits. New Phytol. 152 (1), 69–83. https://doi.org/10.1046/j.0028-646x.2001.00239.x.

Garnier, E., Stahl, U., Laporte, M.A., Kattge, J., Mougenot, I., Kühn, I., Klotz, S., 2017. Towards a thesaurus of plant characteristics: an ecological contribution. J. Ecol. 105 (2), 298–309. https://doi.org/10.1111/1365-2745.12698.

Gini, C. 1955. Memorie di metodologia statistica (Vol. 1). Libr. goliardica.

Golodets, C., Sternberg, M., Kigel, J., 2009. A community-level test of the leaf-height-seed ecology strategy scheme in relation to grazing conditions. J. Veg. Sci. 20 (3), 392–402. https://doi.org/10.1111/j.1654-1103.2009.01071.x.

Grime, J.P., 1974. Vegetation classification by reference to strategies. Nature 250, 26–31. https://doi.org/10.1038/250026a0.

Grime, J.P., 1977. Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. Am. Nat. 111 (982), 1169–1194. https://www.jstor.org/stable/2460262.

Grimm, V., Berger, U., DeAngelis, D.L., Polhill, J.G., Giske, J., Railsback, S.F., 2010. The ODD protocol: a review and first update. Ecol. Model. 221 (23), 2760–2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019.

Harrison, S.P., Prentice, I.C., Barboni, D., Kohfeld, K.E., Ni, J., Sutra, J.P., 2010. Ecophysiological and bioclimatic foundations for a global plant functional classification. J. Veg. Sci. 21 (2), 300–317. https://doi.org/10.1111/j.1654-1103.2009.01144.x.

Harvey, P.H., Pagel, M.R., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press.

Hérault, B., 2007. Reconciling niche and neutrality through the Emergent Group approach. Perspect. Plant Ecol. Evol. Systemat. 9 (2), 71–78. https://doi.org/10.1016/j.ppees.2007.08.001.

Hérault, B., Honnay, O., 2005. The relative importance of local, regional and historical factors determining the distribution of plants in fragmented riverine forests: an emergent group approach. J. Biogeogr. 32 (12), 2069–2081. https://doi.org/10.1111/j.1365-2699.2005.01351.x.

Hodgson, J.G., Wilson, P.J., Hunt, R., Grime, J.P., Thompson, K., 1999. Allocating CSR plant functional types: a soft approach to a hard problem. Oikos 85 (2), 282–294. https://doi.org/10.2307/3546494.

Hubert, L.J., Levin, J.R., 1976. Evaluating object set partitions: Free-sort analysis and some generalizations. J. Verbal Learn. Verbal Behav. 15 (4), 459–470. https://doi.org/10.1016/S0022-5371(76)90041-4.

Hudson, L.N., Newbold, T., Contu, S., Hill, S.L.L., Lysenko, I., De Palma, A., Phillips, H.R. P., Senior, R.A., Bennett, D.J., Booth, H., Choimes, A., Correia, D.L.P., Day, J., Echeverría-Londoño, S., Garon, M., Harrison, M.L.K., Ingram, D.J., Jung, M., Kemp, V., Purvis, A., 2014. The PREDICTS database: a global database of how local terrestrial biodiversity responds to human impacts. Ecol. Evol. 4 (24), 4701–4735. https://doi.org/10.1002/ece3.1303.

Hughes, L., Dunlop, M., French, K., Leishman, M.R., Rice, B., Rodgerson, L., Westoby, M., 1994. Predicting dispersal spectra: a minimal set of hypotheses based on plant attributes. J. Ecol. 82 (4), 933–950. https://doi.org/10.2307/2261456.

Jin, Y., Qian, H., 2019. V. PhyloMaker: an R package that can generate very large phylogenies for vascular plants. Ecography 42 (8), 1353–1359. https://doi.org/10.1111/ecog.04434.

Jin, Y., Qian, H., 2022. V. PhyloMaker2: An updated and enlarged R package that can generate very large phylogenies for vascular plants. Plant Diversity 44 (4), 335–339. https://doi.org/10.1016/j.pld.2022.05.005.

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I.C., Leadley, P., Tautenhahn, S., Werner, G.D.A., Aakala, T., Abedi, M., Acosta, A.T.R., Adamidis, G.C., Adamson, K., Aiba, M., Albert, C.H., Alcántara, J.M., Alcázar, C., Aleixo, I., Ali, H., Wirth, C., 2020.

TRY plant trait database—Enhanced coverage and open access. Glob. Chang. Biol. 26 (1), 119–188. https://doi.org/10.1111/gcb.14904.
Kindscher, K., Wells, P.V., 1995. Prairie plant guilds: a multivariate analysis of prairie species based on ecological and morphological traits. Vegetatio 117, 29–50. https://doi.org/10.1007/BF00033257.
Lam, O.H.Y., Tautenhahn, S., Walther, G., Boenisch, G., Baddam, P., Kattge, J., 2022. The'rtry'R package for preprocessing plant trait data. In: EGU General Assembly Conference Abstracts, Vienna, Austria. https://doi.org/10.5194/egusphere-egu22-13251.
Landsberg, J. 1999. Response and effect-different reasons for classifying plant functional types under grazing. In D. Eldridge, D. Freudenberger (Eds.), People and rangelands: Building the future. Proceedings of the VI International Rangeland Congress, Townsville, Australia (pp. 911-915). Aitkenvale, International Rangeland Congress, Inc. .
Laughlin, D.C., 2014. The intrinsic dimensionality of plant traits and its relevance to community assembly. J. Ecol. 102 (1), 186–193. https://doi.org/10.1111/1365-2745.12187.
Laughlin, D.C., Leppert, J.J., Moore, M.M., Sieg, C.H., 2010. A multi-trait test of the leaf-height-seed plant strategy scheme with 133 species from a pine forest flora. Funct. Ecol. 24 (3), 493–501. https://doi.org/10.1111/j.1365-2435.2009.01672.x.
Lavorel, S., McIntyre, S., Landsberg, J., Forbes, T.D.A., 1997. Plant functional classifications: from general groups to specific groups based on response to disturbance. Trends Ecol. Evol. 12 (12), 474–478. https://doi.org/10.1016/S0169-5347(97)01219-6.
Lavorel, S., Rochette, C., Lebreton, J.D., 1999. Functional groups for response to disturbance in Mediterranean old fields. Oikos 84 (3), 480–498. https://doi.org/10.2307/3546427.
Lavorel, S., Díaz, S., Cornelissen, J.H.C., Garnier, E., Harrison, S.P., McIntyre, S., Pausas, J.G., Pérez-Harguindeguy, N., Roumet, C., Urcelay, C., 2007. Plant functional types: are we getting any closer to the Holy Grail? In: Canadell, J.G., Pataki, D.E., Pitelka, L.F. (Eds.), Terrestrial Ecosystems in a Changing World. Global Change — The IGBP Series. Springer, Berlin, Heidelberg, pp. 149–164. https://doi.org/10.1007/978-3-540-32730-1_13.
Lee, J.A., Verleyson, M., 2007. Nonlinear Dimensionality Reduction. Springer, New York, NY https://doi.org/10.1007/978-0-387-39351-3.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2 (3), 18–22. https://CRAN.R-project.org/doc/Rnews/.
Martinez-Taboada, F., Redondo, J.I., 2020. Variable importance plot (mean decrease accuracy and mean decrease Gini). Figure. PLoS ONE. https://doi.org/10.1371/journal.pone.0230799.g002.
McIntyre, S., Lavorel, S., Landsberg, J., Forbes, T.D.A., 1999. Disturbance response in vegetation–towards a global perspective on functional traits. J. Veg. Sci. 10 (5), 621–630. https://doi.org/10.2307/3237077.
McLaren, J.R., Turkington, R., 2010. Ecosystem properties determined by plant functional group identity. J. Ecol. 98 (2), 459–469.
Medeiros, C.D., Henry, C., Trueba, S., Anghel, I., Guerrero, S.D.D.D.L., Pivovaroff, A., Fletcher, L.R., John, G.P., Lutz, J.A., Alonzo, R.M., Sack, L., 2023. Predicting plant species climate niches on the basis of mechanistic traits. Funct. Ecol. 37 (11), 2786–2808. https://doi.org/10.1111/1365-2435.14422.
Millard, S. P., Kowarik, A., Kowarik, M. A. 2018. Package 'EnvStats'. Package for Environmental Statistics. Version, 2, 31-32.
Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kéfi, S., Kergoat, G.J., Lavorel, S., Le Gall, L., Meslin, L., Morand, S., Loreau, M., 2015. Predictive ecology in a changing world. J. Appl. Ecol. 52 (5), 1293–1310. https://doi.org/10.1111/1365-2664.12482.
Noble, I.R., Gitay, H., 1996. A functional classification for predicting the dynamics of landscapes. J. Veg. Sci. 7 (3), 329–336. https://doi.org/10.2307/3236276.
Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., Suggests, M.A.S.S., 2007. The Vegan Package. Community Ecology Package 10 (631–637), 719.
Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H. S., Desper, R. 2019. Package 'ape'. Analyses of phylogenetics and evolution, version, 2(4), 47.
Pausas, J.G., Lavorel, S., 2003. A hierarchical deductive approach for functional types in disturbed ecosystems. J. Veg. Sci. 14 (3), 409–416. https://doi.org/10.1111/j.1654-1103.2003.tb02166.x.
Pavlick, R., Drewry, D.T., Bohn, K., Reu, B., Kleidon, A., 2013. The Jena Diversity-Dynamic Global Vegetation Model (JeDi-DGVM): a diverse approach to representing terrestrial biogeography and biogeochemistry based on plant functional trade-offs. Biogeosciences 10 (6), 4137–4177. https://doi.org/10.5194/bg-10-4137-2013.
Pennell, M.W., FitzJohn, R.G., Cornwell, W.K., 2016. A simple approach for maximizing the overlap of phylogenetic and comparative data. Methods Ecol. Evol. 7 (6), 751–758. https://doi.org/10.1111/2041-210X.12517.
Peres-Neto, P.S., Jackson, D.A., Somers, K.M., 2005. How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. Comput. Stat. Data Anal. 49, 974–997. https://doi.org/10.1016/j.csda.2004.06.015.
Peterson, A.T., 2011. Ecological niche conservatism: A time-structured review of evidence. J. Biogeogr. 38 (5), 817–827. https://doi.org/10.1111/j.1365-2699.2010.02456.x.
Pierce, S., Brusa, G., Vagge, I., Cerabolini, B.E., 2013. Allocating CSR plant functional types: the use of leaf economics and size traits to classify woody and herbaceous vascular plants. Funct. Ecol. 27 (4), 1002–1010. https://doi.org/10.1111/1365-2435.12095.
Pierce, S., Negreiros, D., Cerabolini, B.E., Kattge, J., Díaz, S., Kleyer, M., Tampucci, D., 2017. A global method for calculating plant CSR ecological strategies applied across biomes world-wide. Funct. Ecol. 31 (2), 444–457. https://doi.org/10.1111/1365-2435.12722.
Pillar, V.D., Sosinski Jr, E.E., 2003. An improved method for searching plant functional types by numerical analysis. J. Veg. Sci. 14 (3), 323–332. https://doi.org/10.1111/j.1654-1103.2003.tb02158.x.
Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240. https://doi.org/10.5194/soil-7-217-2021.
R Core Team. 2022. R: A language and environment for statistical computing. R R package version 1.5.1, https://github.com/tidyverse/stringr, https://stringr.tidyverse.org.
Raunkiaer, C., 1934. The Life Forms of Plants and Statistical Plant Geography; Being the Collected Papers of C, Raunkiær. Clarendon Press, Oxford.
Revelle, W., Revelle, M.W., 2015. Package 'psych'. The Comprehensive R Archive Network 337 (338).
Roscher, C., Schumacher, J., Gubsch, M., Lipowsky, A., Weigelt, A., Buchmann, N., Schulze, E.D., 2012. Using plant functional traits to explain diversity–productivity relationships. PLoS One 7 (5), e36760.
Sabatini, F.M., Lenoir, J., Hattab, T., Arnst, E.A., Chytrý, M., Dengler, J., De Ruffray, P., Hennekens, S.M., Jandt, U., Jansen, F., Jiménez-Alfaro, B., Kattge, J., Levesley, A., Pillar, V.D., Purschke, O., Sandel, B., Sultana, F., Aavik, T., Aćić, S., Wagner, V., 2021. sPlotOpen–An environmentally balanced, open-access, global dataset of vegetation plots. Glob. Ecol. Biogeogr. 30 (9), 1740–1764. https://doi.org/10.1111/geb.13346.
Santos, T., Diniz-Filho, J. A., e Luis, T. R., Bini, M., Santos, M. T. 2018. Package 'PVR'. Phylogenetic Eigenvectors Regression and Phylogentic Signal-Representation, 3274.
Scheiter, S., Langan, L., Higgins, S.I., 2013. Next-generation dynamic global vegetation models: learning from community ecology. New Phytol. 198 (3), 957–969. https://doi.org/10.1111/nph.12210.
Schulze, E.D., Beck, E., Buchmann, N., Clemens, S., Müller-Hohenstein, K., Scherer-Lorenzen, M., 2019. Dynamic Global Vegetation Models: Contribution by S. Zaehle. In: Plant Ecology. Springer, Berlin, Heidelberg, pp. 843–863. https://doi.org/10.1007/978-3-662-56233-8_22.
Sitch, S., Smith, B., Prentice, I.C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J.O., Levis, S., Lucht, W., Sykes, M.T., Thonicke, K., Venevsky, S., 2003. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Glob. Chang. Biol. 9 (2), 161–185. https://doi.org/10.1046/j.1365-2486.2003.00569.x.
Sitch, S., Huntingford, C., Gedney, N., Levy, P.E., Lomas, M., Piao, S.L., Bettis, R., Ciais, P., Cox, P., Friedlingstein, P., Jones, C.D., Prentice, I.C., Woodward, F.I., 2008. Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs). Glob. Chang. Biol. 14 (9), 2015–2039. https://doi.org/10.1111/j.1365-2486.2008.01626.x.
Smith, T.M., Shugart, H.H., Woodward, F.I., Burton, P.J., 1993. Plant functional types. In: Solomon, A.M., Shugart, H.H. (Eds.), Vegetation Dynamics & Global Change. Springer, Boston, MA, pp. 272–292. https://doi.org/10.1007/978-1-4615-2816-6_14.
Steffen, W.L., Cramer, W., Plöchl, M., Bugmann, H., 1996. Global vegetation models: incorporating transient changes to structure and composition. J. Veg. Sci. 7 (3), 321–328. https://doi.org/10.2307/3236275.
Stekhoven, D. J., Stekhoven, M. D. J. 2013. Package 'missForest'. R package version, 1.
Symstad, A.J., 2002. An overview of ecological plant classification systems: linking functional response and functional effect groups. Mod. Trends Appl. Terrestr. Ecol. 13–50.
Telenius, A., 2011. Biodiversity information goes public: GBIF at your service. Nord. J. Bot. 29 (3), 378–381. https://doi.org/10.1111/j.1756-1051.2011.01167.x.
Tilman, D. 1988. Plant Strategies and the Dynamics and Structure of Plant Communities. Monographs in Population Biology series (Vol. 26). Princeton University Press. https://doi.org/10.2307/j.ctvx5w9ws.
Van Bodegom, P.M., Douma, J.C., Witte, J.P.M., Ordoñez, J.C., Bartholomeus, R.P., Aerts, R., 2012. Going beyond limitations of plant functional types when predicting global ecosystem–atmosphere fluxes: exploring the merits of traits-based approaches. Glob. Ecol. Biogeogr. 21 (6), 625–636. https://doi.org/10.1111/j.1466-8238.2011.00717.x.
Walker, B.H., 1992. Biodiversity and ecological redundancy. Conserv. Biol. 6 (1), 18–23. http://www.jstor.org/stable/2385847.
Weiher, E., Clarke, G.P., Keddy, P.A., 1998. Community assembly rules, morphological dispersion, and the coexistence of plant species. Oikos 81 (2), 309–322. https://doi.org/10.2307/3547051.
Weiher, E., van der Werf, A., Thompson, K., Roderick, M., Garnier, E., Eriksson, O., 1999. Challenging Theophrastus: A common core list of plant traits for functional ecology. J. Veg. Sci. 10 (5), 609–620. https://doi.org/10.2307/3237076.
Weisser, W.W., Hensel, M., Barath, S., Culshaw, V., Grobman, Y.J., Hauck, T.E., Joschinski, J., Ludwig, F., Mimet, A., Perini, K., Roccotiello, E., Schloter, M., Shwartz, A., Sunguroğlu Hensel, D., Vogler, V., 2023. Creating ecologically sound buildings by integrating ecology, architecture and computational design. People Nat. 5 (1), 4–20. https://doi.org/10.1002/pan3.10411.
Westoby, M., 1998. A leaf-height-seed (LHS) plant ecology strategy scheme. Plant Soil 199 (2), 213–227. http://www.jstor.org/stable/42948252.
Westoby, M., Rice, B., Howell, J., 1990. Seed size and plant growth form as factors in dispersal spectra: ecological archives E071-002. Ecology 71 (4), 1307–1315.
Westoby, M., Falster, D.S., Moles, A.T., Vesk, P.A., Wright, I.J., 2002. Plant ecological strategies: some leading dimensions of variation between species. Annu. Rev. Ecol. Syst. 33 (1), 125–159.

Wherry, E.T., 1927. Divergent soil reaction preferences of related plants. Ecology 8 (2), 197–206.
Wickham, H., Wickham, M. H. 2017. Package tidyverse. Easily install and load the 'Tidyverse.
Winemiller, K.O., Fitzgerald, D.B., Bower, L.M., Pianka, E.R., 2015. Functional traits, convergent evolution, and periodic tables of niches. Ecol. Lett. 18 (8), 737–751. https://doi.org/10.1111/ele.12462.
Wright, I.J., Reich, P.B., Westoby, M., Ackerly, D.D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J.H., Diemer, M., Flexas, J., Garnier, E., Groom, P. K., Gulias, J., Hikosaka, K., Lamont, B.B., Lee, T., Lee, W., Lusk, C., Villar, R., 2004. The worldwide leaf economics spectrum. Nature 428 (6985), 821–827. https://doi.org/10.1038/nature02403.
Wullschleger, S.D., Epstein, H.E., Box, E.O., Euskirchen, E.S., Goswami, S., Iversen, C.M., Kattge, J., Norby, R.J., van Bodegom, P.M., Xu, X., 2014. Plant functional types in Earth system models: past experiences and future directions for application of dynamic vegetation models in high-latitude ecosystems. Ann. Bot. 114 (1), 1–16. https://doi.org/10.1093/aob/mcu077.
Zhang, J., Qian, H., 2023. U. Taxonstand: An R package for standardizing scientific names of plants and animals. Plant Diversity 45 (1), 1–5. https://doi.org/10.1016/j.pld.2022.09.001.