Supplementary Materials for

Title: Multispecies grasslands produce more yield from lower nitrogen inputs across a climatic gradient

Authors: James O'Malley[1], John A. Finn[2]*, Carsten S. Malisch[3], Matthias Suter[4], Sebastian T. Meyer[5], Giovanni Peratoner[6], Marie-Noëlle Thivierge[7], Diego Abalos[3], Paul R. Adler[8], T. Martijn Bezemer[9,10], Alistair D. Black[11], Åshild Ergon[12], Barbara Golińska[13], Guylain Grange[1,2], Josef Hakl[14], Nyncke J. Hoekstra[15], Olivier Huguenin-Elie[4], Jingying Jing[16], Jacob M. Jungers[17], Julie Lajeunesse[18], Ralf Loges[19], Gaëtan Louarn[20], Andreas Lüscher[4], Thomas Moloney[21], Christopher K. Reynolds[22], Ievina Sturite[23], Ali Sultan Khan[2], Rishabh Vishwakarma[1], Yingjun Zhang[16], Feng Zhu[24], Caroline Brophy[1]

Affiliations:

[1]School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland.

[2]Environment, Soils and Land Use Dept., Teagasc, Johnstown Castle, Co. Wexford, Ireland.

[3]Department of Agroecology, Aarhus University, Tjele, Denmark.

[4]Agroscope, Reckenholzstrasse, Zurich, Switzerland.

[5]School of Life Sciences, Technical University of Munich, Munich, Germany.

[6]Laimburg Research Centre, Laimburg 6, Vadena/Pfatten, Italy.

[7]Agriculture and Agri-Food Canada, Quebec Research and Development Centre, Quebec City, Quebec, Canada.

[8]United States Department of Agriculture, Agricultural Research Service, PA, USA.

[9]Institute of Biology, Leiden University, Leiden, The Netherlands.

[10]Netherlands Institute of Ecology, NIOO-KNAW, Wageningen, The Netherlands.

[11]Department of Agricultural Sciences, Lincoln University, Christchurch, New Zealand.

[12]Dept. of Plant Sciences, Norwegian University of Life Sciences, Norway.

[13]Department of Grassland and Natural Landscape Sciences, Poznan University of Life Sciences, Poland.

[14]Department of Agroecology and Crop Production, Czech University of Life Sciences Prague, Czech Republic.

[15]Louis Bolk Institute, Bunnik, The Netherlands.

[16]College of Grassland Science and Technology, China Agricultural University, Beijing, China.

[17]Department of Agronomy and Plant Genetics, University of Minnesota, MN, USA.

[18]Agriculture and Agri-Food Canada, Research farm, Normandin, Quebec, Canada.

[19]Institute of Animal Nutrition and Physiology, Group of Organic Farming, Kiel University, Kiel, Germany.

[20]INRAE UR4 URP3F, Lusignan, France.

[21]DLF, Ballymountain, Ferrybank, Co. Waterford, Ireland.

[22]School of Agriculture, Policy and Development, University of Reading, Early Gate, UK.

[23]Department of Grassland and Livestock, Norwegian Institute of Bioeconomy Research (NIBIO), Ås station Steinkjer, Norway.

[24]Hebei Key Laboratory of Soil Ecology, Key Laboratory of Agricultural Water Resources, Center for Agricultural Resources Research, Institute of Genetic and Developmental Biology, Chinese Academy of Sciences, Shijiazhuang, China.

Corresponding author: John.Finn@teagasc.ie

The PDF file includes:

Materials and Methods
Figs. S1 to S9
Tables S1 to S4
References (*52-61*)

Other Supplementary Materials for this manuscript include the following:

Data S1 (available on Dryad (*51*))

Materials and Methods

Experimental design of the LegacyNet common experiment

LegacyNet is a voluntary network of experiments that spans 26 sites across 15 countries. The aim of LegacyNet is to test the benefits of multispecies grassland leys within crop rotations. All sites within the network implemented a common two-stage experimental design that consisted of a grassland ley phase followed immediately by a follow-on crop phase. A range of plant communities in the grassland phase, from monocultures of each of the six component species to six-species mixtures, were tested with the objective of identifying optimal species mixtures that maximize forage production and legacy effects. Conducting a common multi-site experiment across a variety of environment conditions (climate, soil type, local management etc.) increases the ability to draw more general inference and conclusions about effect sizes. The focus of this paper is on the biomass yield response measured during the grassland ley phase.

At each site, six species were selected from three functional groups (two grasses, two legumes, and two herbs; denoted G1, G2, L1, L2, H1, H2) for use in the grassland phase. Species were selected for their high forage yields and quality, and complementary traits regarding the manner of nitrogen acquisition and growth form. Furthermore, the species were chosen to be representative of the most commonly used forage species at their location and are therefore best suited to the climatic, soil, and management conditions of the region; the species used at each site are listed in Data S1. The first grass species, G1, was chosen as the grass species that would be most commonly sown in the site's locality, this was *L. perenne* at 21 of 26 sites. As the second grass species, G2, *P. pratense* was used at 17 of 26 sites. The first legume species, L1, was *T. pratense* in 23 of 26 sites, while L2 was *T. repens* at 20 of 26 sites; both species are sown in productive temperate grasslands that include legumes. The first herb species, H1 was *C. intybus* at 25 sites which is commonly sown in leys that include herbs (e.g., (*42*)), and the second herb species, H2, was *P. lanceolata* at 24 of 26 sites.

The experimental design of the grassland phase was a simplex design comprising 33 unique communities of systematically varying sown proportions of each of six species. There were monocultures of each of the six species, and mixture plots with two, three, four, or six species (where species were sown in equal proportion). The sown proportions of each species for the design communities are shown in Table S1 (communities 1-33). Monocultures and the six-species mixture were replicated three times each, while remaining communities were each established in one experimental unit (Table S1; communities 1-33). Note that although some mixtures are not replicated beyond one experimental unit, replication is achieved from the spread of individual points across the sown species proportions (as is possible for any continuous design space). The application rate of synthetic nitrogen (N) fertilizer was also manipulated experimentally in the design; most of the plots across the simplex design (Table S1, communities 1-33) were managed at a moderate N level (with average application rate of inorganic nitrogen fertiliser across sites of 108.7 kg ha^{-1} yr^{-1}), while five (additional) replicated plots of the G1 grass monoculture were managed at a high N level (Table S1, community 34; with average application rate of inorganic nitrogen fertiliser across sites of 260.5 kg ha^{-1} yr^{-1}). Specific levels of N application were dictated by local practice (values shown in Data S1, two sites used zero N fertilizer as their moderate level). The high N level was typically at least twice the moderate N level (the difference between the two rates was on average 152 kg ha^{-1}). At each site, the seeding rate for monocultures was the locally recommended seeding rate for the establishment of monocultures of the respective species. To calculate the intended sown species proportion for each species in mixture communities, the sown proportion for each species (Table S1) was multiplied by its corresponding monoculture seeding rate, and seeds for all species were combined to create the seed mixture for sowing in the field plot. At each site, the 52 plant communities were allocated to field plots in a fully randomized design. All plots measured a minimum of 3 m × 5 m. Across all sites, there were a total of 1,382 plots (while most sites had just the 52 plots listed in Table S1, there were some deviations to the design at a small number of individual sites).

During the grassland phase, plots were regularly harvested via mechanical harvester at time intervals dictated by local practice (the total number of harvests in the grassland phase for each site is shown in Data S1). At each harvest, the dry biomass yield (t ha^{-1}) of each plot was measured by cutting and weighing the fresh mass by a plot harvester and by determining its dry matter (DM) content on a herbage sub-sample that was oven-dried to constant weight. At most harvests, a sample of the harvest biomass was taken for forage quality analysis. No weeding took place throughout the duration of the experiment (except at sites CN2 and CN3), but there was a cleaning cut within the first six months of establishment at some sites.

Weather data recorded at each site included maximum, minimum, and mean daily temperatures in °C and daily precipitation in mm (including snow and irrigation applications); these variables were recorded monthly in the case of CN2 and CN3. A range of summary statistics were calculated across the grassland phase period (from the establishment date to the date of the final harvest) for each of these variables, these were: average daily temperature, average daily precipitation, averages of the ten highest and lowest daily temperatures, and antecedent precipitation index (API, an estimate of soil moisture values). This gave rise to one value per site for each of these variables.

Plots were maintained during the grassland phase for a minimum of 18 months, after which they were terminated and a follow-on crop of a pure grass ley, cereal, or maize was established. The duration of the grassland phase was 24 months on average across sites. The follow-on crop was grown and measured for one full growing season and retained the same plot structure as the grassland phase. Follow-on crop measurements are not part of the analyses presented here.

Calculation of the yield response variable

The total number of harvests over the grassland phase and the duration of the phase varied by site (Data S1). The typical length of a growing season also varied substantially across sites. To account for these site variations, a yield per growing season for each plot at each site was calculated by first calculating the average daily yield of each plot during the grassland phase and then multiplying it by an estimate of the average number of growing days in a single growing season for the site. We define the following notation used in the calculation of the yield response variable:

- H_k is the total number of harvests taken at site *k* over the grassland phase.
- $m_{k[q]h}$ is the DM yield of harvest h for plot q at site k. The DM yields were summed across all harvests in the grassland phase to give the total DM yield per plot.
- GDL_k (growing days in the grassland phase) is the total number of days during the grassland phase at site k where the plots were deemed to be growing. It was calculated at each site by summing up the number of days in each year where the plots were harvested, between the 1st of January to the final harvest where the ten-day rolling average temperature exceeded a threshold of 5°C. If the plots were harvested in the same year as establishment,

the number of growing days for that year were counted from the date of establishment (rather than the 1st of January) to the final harvest. The number of growing days in each year at site k were then summed to give GDL_k.

- Thus, $\frac{\sum_{h=1}^{H_k} m_{k[q]h}}{GDL_k}$ estimates the average daily yield of each plot during the grassland phase at a given site.
- $GDSS_k$ (growing days in a single season) is an estimate of the number of days in a single growing season of site k. This variable was calculated for each full (calendar) year where the experiment was running at each site; the year of termination was only included if the final harvest in that year was deemed to be at the end of the growing season. Thus, for a single year, we computed the number of days where the ten-day rolling average temperature exceeded a threshold of 5°C between the 1st of January (or from the 1st of July in the case of New Zealand), and the date of the site's final harvest in that year. Where there was more than one full year of grassland phase at a site, the average over the available years was taken as the $GDSS_k$ value.

Finally, $y_{k[q]}$, the yield per growing season for plot q at site k was computed as:

$$y_{k[q]} = \frac{(\sum_{h=1}^{H_k} m_{k[q]h})}{GDL_k} * (GDSS_k) \qquad \text{Eq. (1)}$$

We refer to y as 'yield per growing season' or 'yield' for short throughout the main text and its units are t ha^{-1}.

Statistical analysis using Diversity-Interactions modelling

Traditional approaches to modelling the biodiversity and ecosystem function relationship focus on the number of species (species richness) as the main determinant of ecosystem function. More recently, the Diversity-Interactions modelling approach extends this to assess how species composition, richness and initial proportions jointly affect ecosystem function (*28-30*).

We modelled the LegacyNet plot-level yield data from across all sites using the Diversity-Interactions (DI) modelling framework, fitted as a random coefficients linear mixed-effects model (*52*), where the response, $y_{k[q]}$, was yield per growing season (t ha^{-1}), calculated as described in Eq. (1). This allowed us to quantify the effects of manipulating species diversity on yield in the multispecies grasslands across our multi-site international-scale experiment. The DI model takes the general form:

$$y = [Identity\ terms] + [Interaction\ terms] + [High\ N\ term] + \varepsilon \qquad \text{Eq. (2)}$$

The model explicitly includes species' identity effects through sown species' proportion predictors (six species' identity parameters), species' pairwise interactions that can take varying forms (some versions more parsimonious than others), and a term for the high N grass monoculture (one parameter). The model implicitly includes species richness as an explanatory variable (*30*).

A possible model specification is:

$$y_{k[q]} = \sum_{i=1}^{6}(\beta_i + b_{ik})p_{k[q]i} + \sum_{1 \le i < j \le 6}(\delta_{ij} + d_{ijk})\left(p_{k[q]i}p_{k[q]j}\right)^{\theta} + (\alpha + a_k)X_{k[q]} + \varepsilon_{k[q]} \quad \text{Eq. (3)}$$

Where $p_{k[q]i}$ represents the sown proportion of species i in plot q at site k. $X_{k[q]}$ is coded as one for high N monoculture plots and zero otherwise. When $\mathrm{X}_{\mathrm{k[q]}} = 1$, each $\mathrm{p}_{\mathrm{k[q]i}}$ is set to zero, hence the overall expected response for the high N monocultures is α. The lowercase b, d, and a terms represent the random coefficients in the model. The random terms assumptions are:

$$(b_{1k}, \dots, b_{6k}, d_{12k}, \dots, d_{56k}, a_k)^T \sim MVN(\mathbf{0}, \boldsymbol{D})$$

$$\varepsilon_{k[q]} \sim N(0, \sigma_k^2)$$

Hence, there are overall (across site) effects β, δ, and α, that vary from site to site according to their corresponding random effect, with the variance of each random coefficient capturing the nature of the spread across sites. The variance-covariance matrix **D** can be structured in many forms, for example, a parsimonious and biologically motivated structure related to the meaning of the various predictors in the model is:

$$\mathbf{D} = \begin{pmatrix} \sigma^2_{\mathrm{b}_1} & \cdots & \sigma_{\mathrm{SS}} & \sigma_{\mathrm{SI}} & \cdots & \sigma_{\mathrm{SI}} & \sigma_{\mathrm{SN}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{\mathrm{SS}} & \cdots & \sigma^2_{\mathrm{b}_6} & \sigma_{\mathrm{SI}} & \cdots & \sigma_{\mathrm{SI}} & \sigma_{\mathrm{SN}} \\ \sigma_{\mathrm{SI}} & \cdots & \sigma_{\mathrm{SI}} & \sigma^2_{\mathrm{d}_{12}} & \cdots & \sigma_{\mathrm{II}} & \sigma_{\mathrm{IN}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{\mathrm{SI}} & \cdots & \sigma_{\mathrm{SI}} & \sigma_{\mathrm{II}} & \cdots & \sigma^2_{\mathrm{d}_{56}} & \sigma_{\mathrm{IN}} \\ \sigma_{\mathrm{SN}} & \cdots & \sigma_{\mathrm{SN}} & \sigma_{\mathrm{IN}} & \cdots & \sigma_{\mathrm{IN}} & \sigma^2_{\mathrm{a}} \end{pmatrix} \quad \text{Eq. (4)}$$

Where each random coefficient has a unique variance term, to allow it to vary from site to site: $\sigma^2_{\mathrm{b}_i}$ for the random effect associated with each species identity term; $\sigma^2_{\mathrm{d}_{ij}}$ for the random effect associated with each pairwise interaction between species i and j; and σ^2_{a} for the random effect associated with the high N term. Each pair of species identity random effects (b terms) have a common covariance (σ_{SS}); each pair of interaction random effects (d terms) have a common covariance (σ_{II}); each pair of species identity-interaction random effects have a common covariance (σ_{SI}), while the high N random effect has a covariance with the species identity random effects (σ_{SN}) and with the interaction random effects (σ_{IN}).

Model selection process

Our model selection procedure was adapted from West, Welch, and Galecki (*53*) for a random coefficients Diversity-Interactions modelling framework. There were two main components to our model selection process: select the best structure for the fixed effect interaction terms, including whether the θ parameter differed from 1, and select the best random coefficients structure to capture variation from site to site. The following process was used.

1. We fitted a Diversity-Interactions model (*28-30*) that assumed the full pairwise fixed effect interaction structure (i.e., all pairs of species interacted uniquely giving 15 individual interaction terms). Site was included as a fixed blocking factor to account for site-to-site variation (site as a fixed effect is used only for this first step in the selection process, mixed-effects models will be used in subsequent steps to capture site-to-site variation). This model was fitted firstly with θ set to 1, and secondly with θ estimated freely. The model with θ = 1 was fitted using ordinary least squares estimation, while the model with θ estimated was fitted via profile likelihood (*30*), and the two models were compared using the sample-size corrected Akaike's Information Criterion AICc.

2. We fitted a random coefficients DI model, still assuming the full pairwise interaction structure, with all identity terms, species interaction terms and the high N term as random coefficients varying from site to site. Following (*32*), we used the value of θ selected in step 1; i.e., the θ parameter was not re-estimated in this step of the selection process. We tested the variations in the structure of the variance-covariance matrix for the random effects (i.e., matrix D), including:
    - Variance components (the same variances for each random effect and zero covariances between each pair; 1 parameter),
    - Compound symmetry with homogeneous variances (equal variances and equal covariances; 1 + 1 parameters),
    - A biologically motivated covariance structure related to the meaning of the predictors in the model: unique variances for all random effects, covariances for all pairs of identity effects, all pairs of identity-interaction effects, all pairs of interaction-interaction random effects, all high N-ID effects, all high N-interaction effects; 22 + 5 parameters, as specified in the matrix D example shown in Eq. (4).

   These models were fitted using Restricted Maximum Likelihood (REML) and compared using AICc. Note that the variance-covariance structure that would assume unique variances and unique covariances between each pair of random effects (22 + 231 = 253 parameters) was not fitted due to convergence issues related to computational scale.

3. We assessed the assumption that the within-site error variance was constant across all sites and compared it to a heterogeneous structure that allowed the within-site error term variance to differ by site. These two models were estimated with REML and compared using AICc.

4. We compared the model selected in step 3 to alternative ways to model the pairwise interactions (*28, 30*). We tested the assumptions that the fixed interaction effects were dictated by functional group membership, that all interactions were equal, and that all interaction terms were equal to 0. We followed a two-step procedure to test this, whereby random effects and fixed effects were changed one at a time; models that differed only in random effects were fitted using REML for comparison, while models that differed only in fixed effects were fitted using Maximum Likelihood (ML) for comparison.

5. We re-estimated θ using the model selected in step 4 to confirm its final estimate.

6. Finally, we tested a range of weather variables one by one for inclusion as fixed effects in the model selected in step 5 to select our final model (Sattherthwaite denominator degrees of freedom were used during the tests). The variables investigated were average daily temperature, average daily precipitation, averages of the ten highest and lowest daily temperatures, antecedent precipitation index (API, an estimate of soil moisture values), each calculated per site over the grassland phase. Interactions between the weather variables with species identity and species interaction terms and with the high N term, and quadratic terms of the weather variables were also tested.

As part of the model selection process, model diagnostics were checked via a range of measures. These included plots of residuals and random effect estimates, and Cookes Distance and the PRESS statistic to assess the influence of individual sites on parameter estimates and model predictions respectively.

The final model selected

The final model was fitted using REML and had six species identity effect terms (one each for G1, G2, L1, L2, H1, H2), six functional group species pairwise interaction terms (grass-legume, grass-herb, legume-herb, grass-grass, legume-legume, herb-herb) that each included the non-linear θ parameter that differed from 1, and a high N term. All of these were fitted as random coefficients, i.e., an overall effect across sites was estimated and an associated variance term captured its variation from site to site. The structure of the variance covariance matrix for the random effects was similar to the matrix D example provided in Eq. (4). We found evidence of a quadratic effect of average daily temperature, and a positive interaction between average daily temperature and the GL and HL interactions. Although the p-value for the quadratic temperature effect in Table S2 is non-significant, model diagnostics and tests revealed that the quadratic effect was important, but more variable at the extreme lower end than the middle of the temperature gradient; this was investigated using 'leave-one-out' statistics where sites were omitted one by one. The predictive power of the model was best when a quadratic form of the temperature effect was included. We did not find evidence for the inclusion of other tested weather variables.

The estimates of the fixed effects from the final model are in Table S2, and random effect variance component estimates are in Table S3. Note that the average daily temperature variable was centered across sites to aid in the interpretation of the fixed effects estimates in Table S2, where the average across sites of the average daily temperature variable was 9.43ºC. Thus, all parameters in Table S2 (excluding the three weather-related variables) can be interpreted for a temperature of 9.43°C.

Deriving inference from the final Diversity-Interactions model

Community comparison tests such as overyielding (communities outperforming the weighted average of their component monoculture performances), transgressive overyielding (communities outperforming the best monoculture), and tests of comparison against the high N grass monoculture and the two-species 70:30 G1:L2 community, were conducted using predictions from the fitted final DI model. We describe the general principle of these comparisons.

Assume it is of interest to test if a community M is significantly different from another community B. Let x_M be a vector containing the predictor variables for community M, and x_B be the vector of predictor variables for community B. We calculate the contrast comparing the predicted yields of communities M and B as:

$$\hat{C} = \hat{y}_M - \hat{y}_B$$

Where $\hat{y}_M$ is the predicted yield of M, $\hat{y}_B$ is the predicted yield of B.
We calculate *t*, as

$$t = \frac{\hat{C} - 0}{SE(\hat{C})}$$

Assuming a null hypothesis that the contrast is equal to 0 and where the standard error of $\hat{C}$ can be calculated using the variance-covariance matrix of our fixed model parameters, denoted **V**.

$$SE(\hat{C}) = \sqrt{(x_M - x_B)\boldsymbol{V}(x_M - x_B)'}$$

We infer significance of the comparison between mixtures M and B, if *t* is greater than 2, or less than minus 2.

For tests performed across sites, we used the fixed effects estimates and fixed effects covariance matrices from our fitted mixed model (Tables S2 and S3). In the case of tests at the individual site level, we used the model fitted to each site individually (estimates not shown).

Some of our selected communities for comparison are in the experimental design (e.g., the high N grass monoculture), while some are not (e.g., the predicted two-species 70:30 G1:L2 community) but can be predicted since they are part of the continuous design space in our experiment that is modelled via the Diversity-Interactions modelling approach.

Testing for functional redundancy

Functional redundancy refers to the situation where two (or more) species fulfil identical roles in an ecosystem and thus substitution of one by another does not impact the ecosystem function of interest (*28, 54*). A pair of species is said to be functionally redundant, if they have identical monoculture performances (identity effects), do not interact with one another, and have the same interaction strengths with each of the other species in the ecosystem (*28*). Under the DI framework, the presence/absence of functional redundancy can be assessed by fitting a DI model where these conditions are respected and testing if it fits the data better than the DI model containing separate identity and interaction effects for each species. For each site separately, the functional redundancy between the two species within each functional group was tested (separately for each functional group). For example, to assess functional redundancy between the two grasses, G1 and G2 at a particular site, the model in Eq. (5) was compared to that in Eq. (6). Similar comparisons were also performed to assess functional redundancy between the two legume species and between the two herb species.

$$\begin{aligned}\hat{y} = \hat{\beta}_G(P_{G1} + P_{G2}) + \hat{\beta}_{L1}P_{L1} + \hat{\beta}_{L2}P_{L2} + \hat{\beta}_{H1}P_{H1} + \hat{\beta}_{H2}P_{H2} + \hat{\alpha}X +\\ \hat{\delta}_{GL1}(P_{G1} + P_{G2})^{\hat{\theta}}P_{L1}^{\hat{\theta}} + \hat{\delta}_{GL2}(P_{G1} + P_{G2})^{\hat{\theta}}P_{L2}^{\hat{\theta}} +\\ \hat{\delta}_{GH1}(P_{G1} + P_{G2})^{\hat{\theta}}P_{H1}^{\hat{\theta}} + \hat{\delta}_{GH2}(P_{G1} + P_{G2})^{\hat{\theta}}P_{H2}^{\hat{\theta}} +\\ \hat{\delta}_{L1L2}(P_{L1}P_{L2})^{\hat{\theta}} + \hat{\delta}_{L1H1}(P_{L1}P_{H1})^{\hat{\theta}} + \hat{\delta}_{L1H2}(P_{L1}P_{H2})^{\hat{\theta}} +\end{aligned} \quad \text{Eq. (5)}$$

$$\hat{\delta}_{L2H1}(P_{L2}P_{H1})^{\hat{\theta}} + \hat{\delta}_{L2H2}(P_{L2}P_{H2})^{\hat{\theta}} + \hat{\delta}_{H1H2}(P_{H1}P_{H2})^{\hat{\theta}}$$

$$\begin{aligned}\hat{y} = {} & \hat{\beta}_{G1}P_{G1} + \hat{\beta}_{G2}P_{G2} + \hat{\beta}_{L1}P_{L1} + \hat{\beta}_{L2}P_{L2} + \hat{\beta}_{H1}P_{H1} + \hat{\beta}_{H2}P_{H2} + \hat{\alpha}X + \\ & \hat{\delta}_{G1G2}(P_{G1}P_{G2})^{\hat{\theta}} + \hat{\delta}_{L1L2}(P_{L1}P_{L2})^{\hat{\theta}} + \hat{\delta}_{H1H2}(P_{H1}P_{H2})^{\hat{\theta}} \\ & \hat{\delta}_{G1L1}(P_{G1}P_{L1})^{\hat{\theta}} + \hat{\delta}_{G1L2}(P_{G1}P_{L2})^{\hat{\theta}} + \hat{\delta}_{G1H1}(P_{G1}P_{H1})^{\hat{\theta}} + \hat{\delta}_{G1H2}(P_{G1}P_{H2})^{\hat{\theta}} + \\ & \hat{\delta}_{G2L1}(P_{G2}P_{L1})^{\hat{\theta}} + \hat{\delta}_{G2L2}(P_{G2}P_{L2})^{\hat{\theta}} + \hat{\delta}_{G2H1}(P_{G2}P_{H1})^{\hat{\theta}} + \hat{\delta}_{G2H2}(P_{G2}P_{H2})^{\hat{\theta}} + \\ & \hat{\delta}_{L1H1}(P_{L1}P_{H1})^{\hat{\theta}} + \hat{\delta}_{L1H2}(P_{L1}P_{H2})^{\hat{\theta}} + \hat{\delta}_{L2H1}(P_{L2}P_{H1})^{\hat{\theta}} + \hat{\delta}_{L2H2}(P_{L2}P_{H2})^{\hat{\theta}}\end{aligned} \qquad \text{Eq. (6)}$$

If $\theta = 1$, the model in equation (5) is nested within the model in equation (6) and they can be compared using an F-test. However, if the estimate of θ differs between the two models, they are only partially nested and can be compared using Vuong's test (*55*). If the model in equation (5) fits the data as well as or better than the model in equation (6) for a given site, then there was evidence of functional redundancy for that pair of species at the site.

Software

Step 1 in the model selection process was carried out using the `DImodels` package (*30*) in R version 4.4.2 (*56*), while remaining statistical analyses were carried out using SAS software (*57*) (copyright © 2020 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA). The figures presented in the main text and supplementary materials were generated using the `ggplot2` (*58*), `ggmap` (*59*), `DImodelsVis` (*60*) and `PieGlyph` (*61*) R packages.

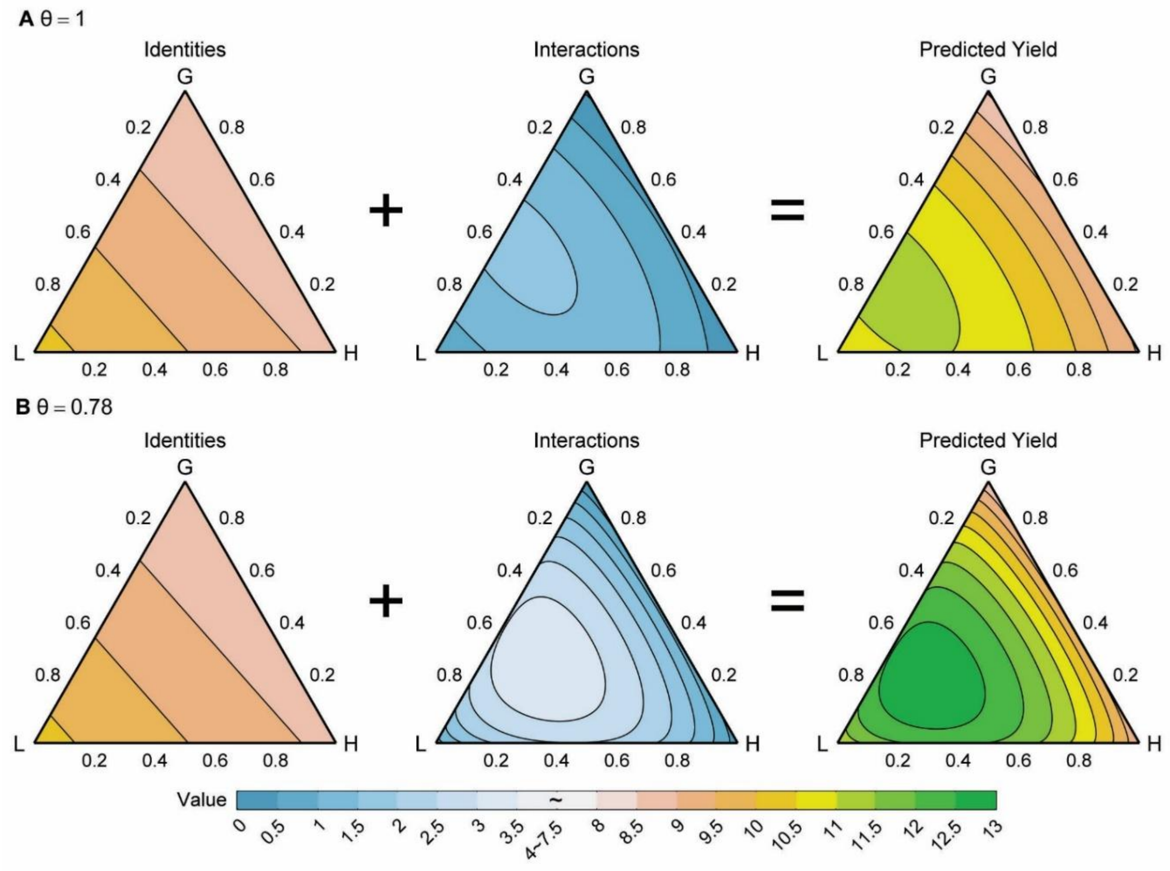


Fig. S1.

Effect of the non-linear theta (θ) parameter in the final fitted Diversity-Interactions model. In each row, predictions from the model are split into a component for the net identity effects (first ternary, 'Identities'), a component for the net interactions (second ternary, 'Interactions') and the combined predictions (third ternary, 'Predicted Yield') as the sown proportions of grass (G), legume (L) and herb (H) vary. Only the value of θ differs between the two rows: **(A)** θ is equal to 1, and **(B)** θ is equal to 0.7816041, the estimate of θ in the final model. The legend is truncated between 4 and 7.5 for ease of readability.

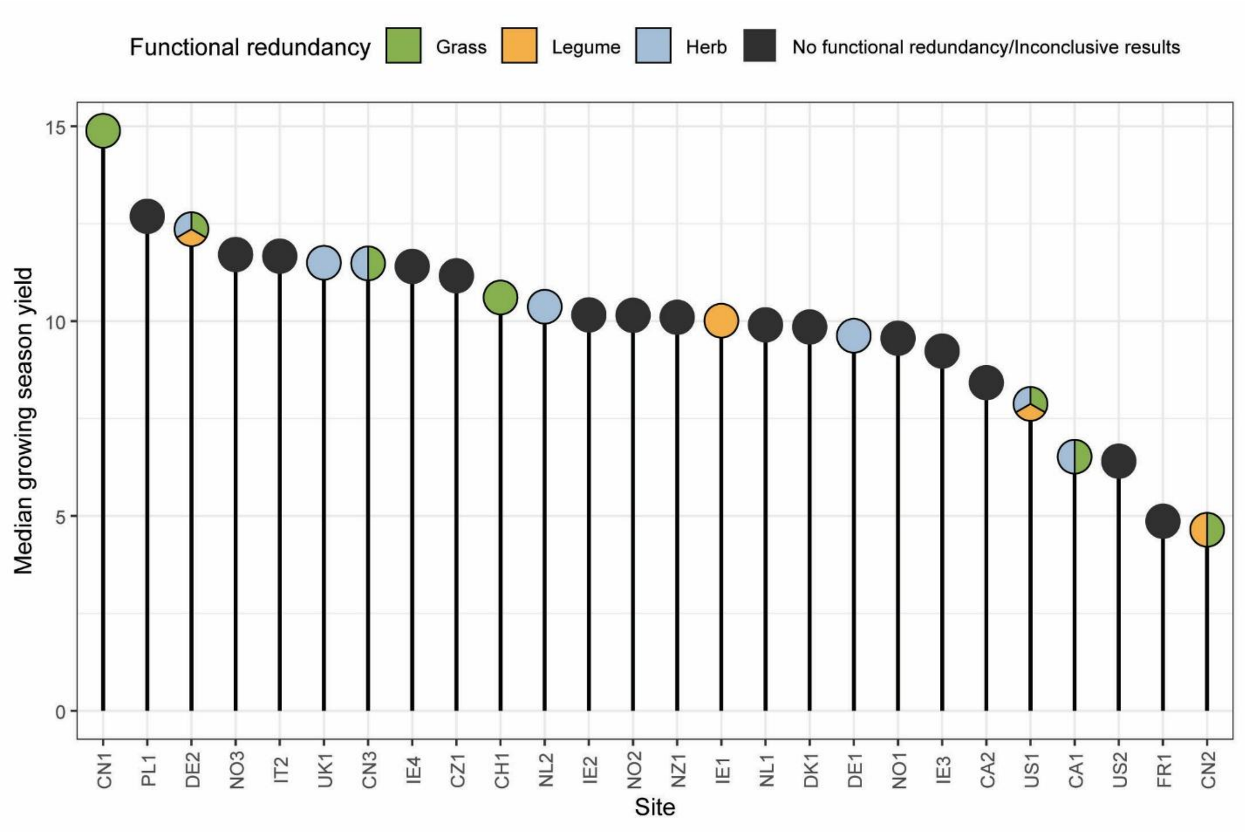
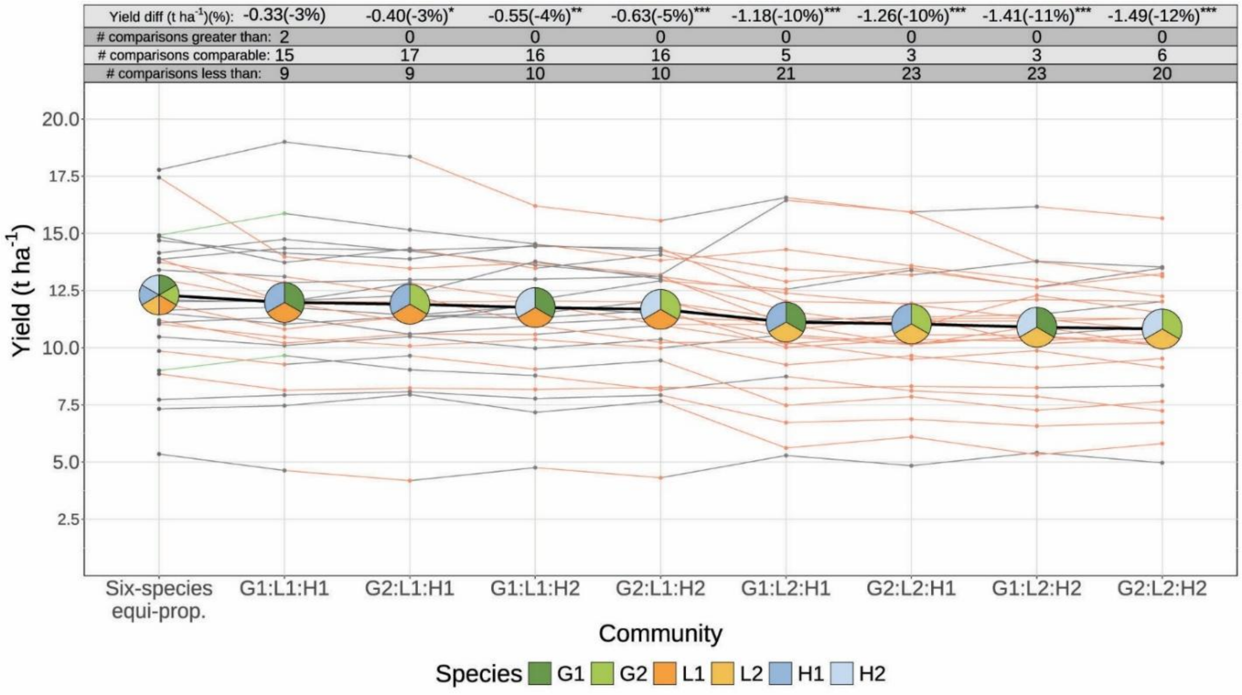


Fig. S2. Functional redundancy within functional groups. The presence/absence of functional redundancy between the two grass (green), the two legume (orange), and the two herb (blue) species at each site is shown using pie-glyphs. The pie-glyphs with green, orange and/or blue in them indicate the presence of functional redundancy between the species from the respective functional groups, while those colored in dark grey represent sites with no redundancy (or inconclusive results) between the pair of species in each functional group (details in (*27*)). The sites along the x-axis are arranged in decreasing order of their median yields.



Yield diff (t ha^{-1})(%):	-0.33(-3%)	-0.40(-3%)*	-0.55(-4%)**	-0.63(-5%)***	-1.18(-10%)***	-1.26(-10%)***	-1.41(-11%)***	-1.49(-12%)***
# comparisons greater than:	2	0	0	0	0	0	0	0
# comparisons comparable:	15	17	16	16	5	3	3	6
# comparisons less than:	9	9	10	10	21	23	23	20

Fig. S3. Predicted yields of equi-proportional three-species mixtures were all (except one) lower than the yield of the six-species equi-proportional mixture. Predicted yields of the six-species equi-proportional mixture and each possible equi-proportional three-species three-functional-group mixture, across sites (black line and pie-glyphs that illustrate the sown species proportions) and at each site (colored lines and smaller points). There are eight possible equi-proportional three-species three-functional-group combinations (G1:L1:H1 to G2:L2:H2 labels across the x-axis) and they are arranged in order of decreasing predicted yield across sites. Predictions at site level are calculated by fitting the final model (excluding climate variables) to the data from each individual site. The predictions from each individual site are joined by lines; both the color of a site prediction and the line to the left of it for three-species mixtures are green if the prediction was more than 2 SEs greater than the six-species mixture, grey if it was comparable (within ± 2SE), and orange if more than 2 SEs lower. Across sites, the difference in yield between each three-species mixture and the six-species mixture is shown at the top of the figure (with significance indicated by: *** p < 0.001, ** p < 0.01, * p < 0.05). The number of sites where the three-species mixture combination was greater than, comparable to, and less than the six-species mixture at the site is also listed at the top of the graph. A total of 208 site-level comparisons were made (26 sites * 8 three-species communities).

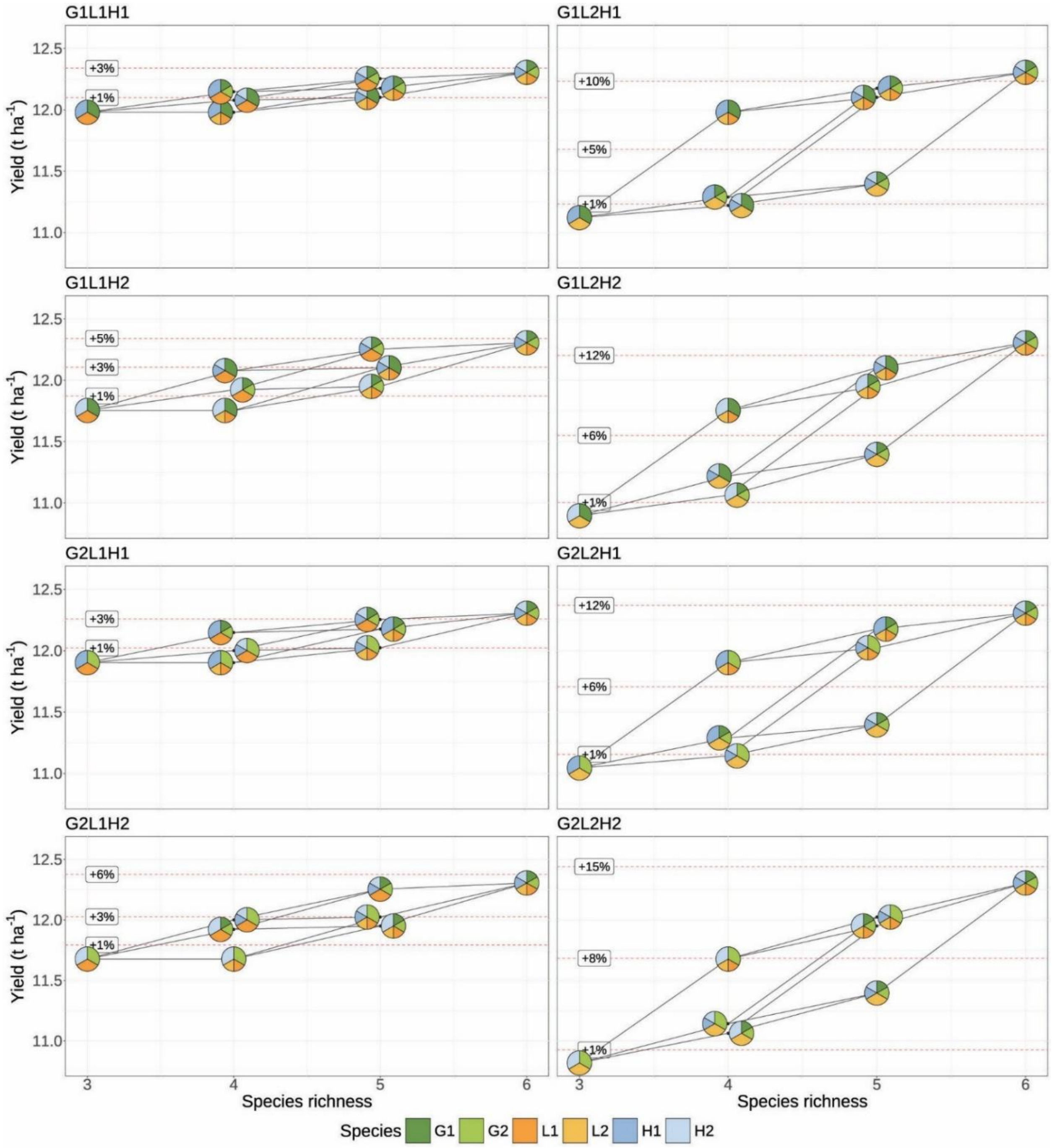


Fig S4 Adding one species at a time generally increased yield across the richness gradient for mixtures that each contained at least one grass, one legume, and one herb, with the magnitude depending on species identity. Each panel shows the predicted yield of three-, four-, five-, and six-species mixtures that each contained at least one grass, one legume, and one herb versus richness, with pie-glyphs illustrating the sown species proportions. For each mixture, the three sown functional group proportions are always equal to 1/3 each, and when there are two species within a functional group, their proportions are both equal to 1/6. The eight panels are distinguished by the combination of species sown in the three-species mixture on the left-hand

side (see panel title). Dotted red lines in each panel indicate percentage level increases relative to the three-species mixture on the left-hand side. Moving from left to right within each panel, solid black lines trace the change in predicted yield when species richness is increased by adding one extra species to each functional group at a time. Moving from right to left illustrates the effects of species removals on yields (i.e., indicative extinction pathways). A jitter has been introduced on the pie-glyphs to prevent overplotting.

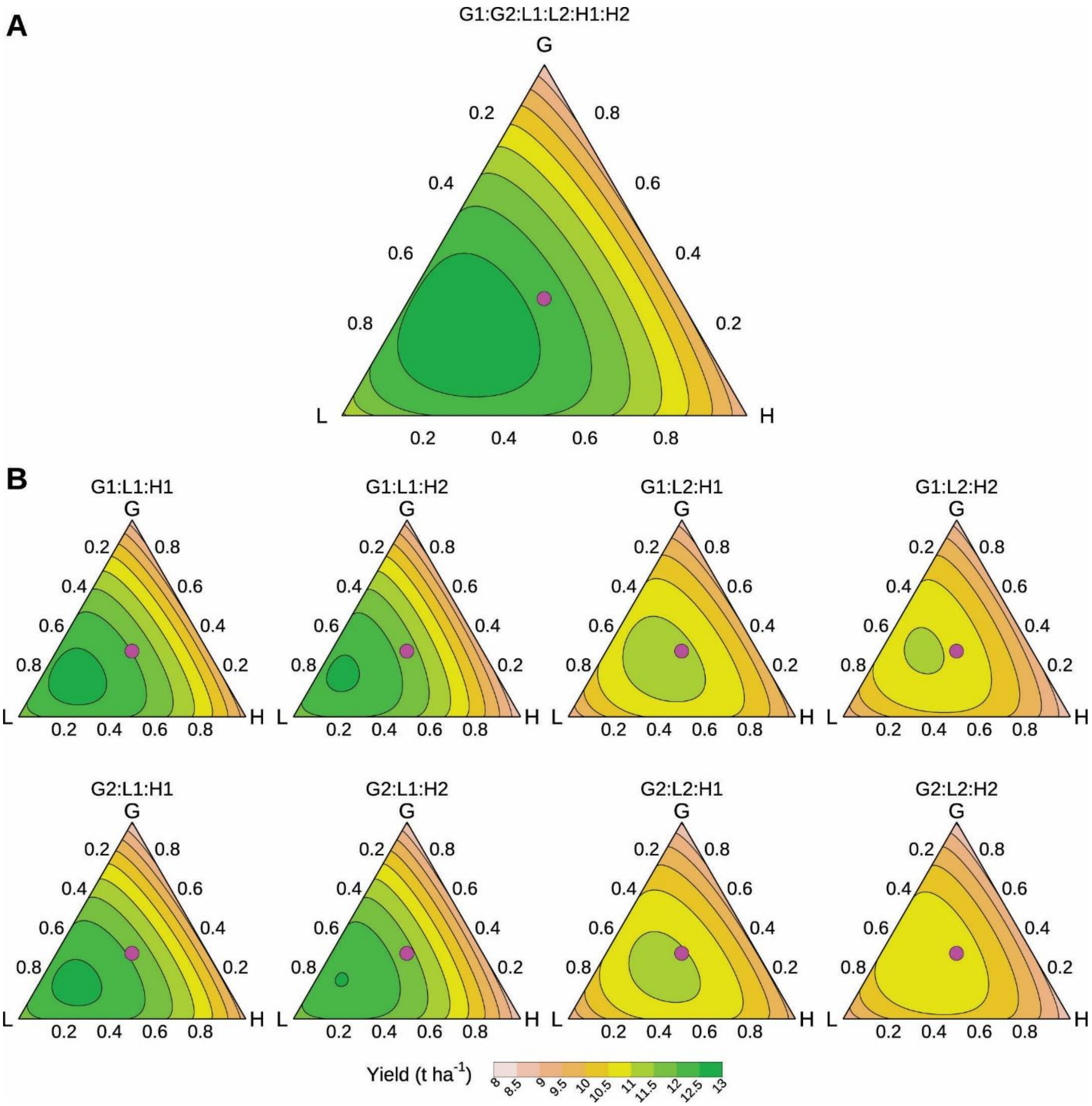


Fig. S5. Sowing two species, rather than one, within each functional group leads to higher or comparable yields. Each ternary diagram shows the predicted yield of communities as the sown proportion of the three functional groups varies. In **(A)** two species are sown in equal proportion to each other within the three functional groups, G = grasses, L = legumes, H = herbs. In **(B),** only one species is sown in each of the three functional groups. The title of each ternary diagram indicates the three species that were sown (G1 to H2). In each ternary, the equi-proportional mixture is shown as a magenta circle (this is a six-species mixture in **A** and a three-species mixture in each ternary in **B**)

.

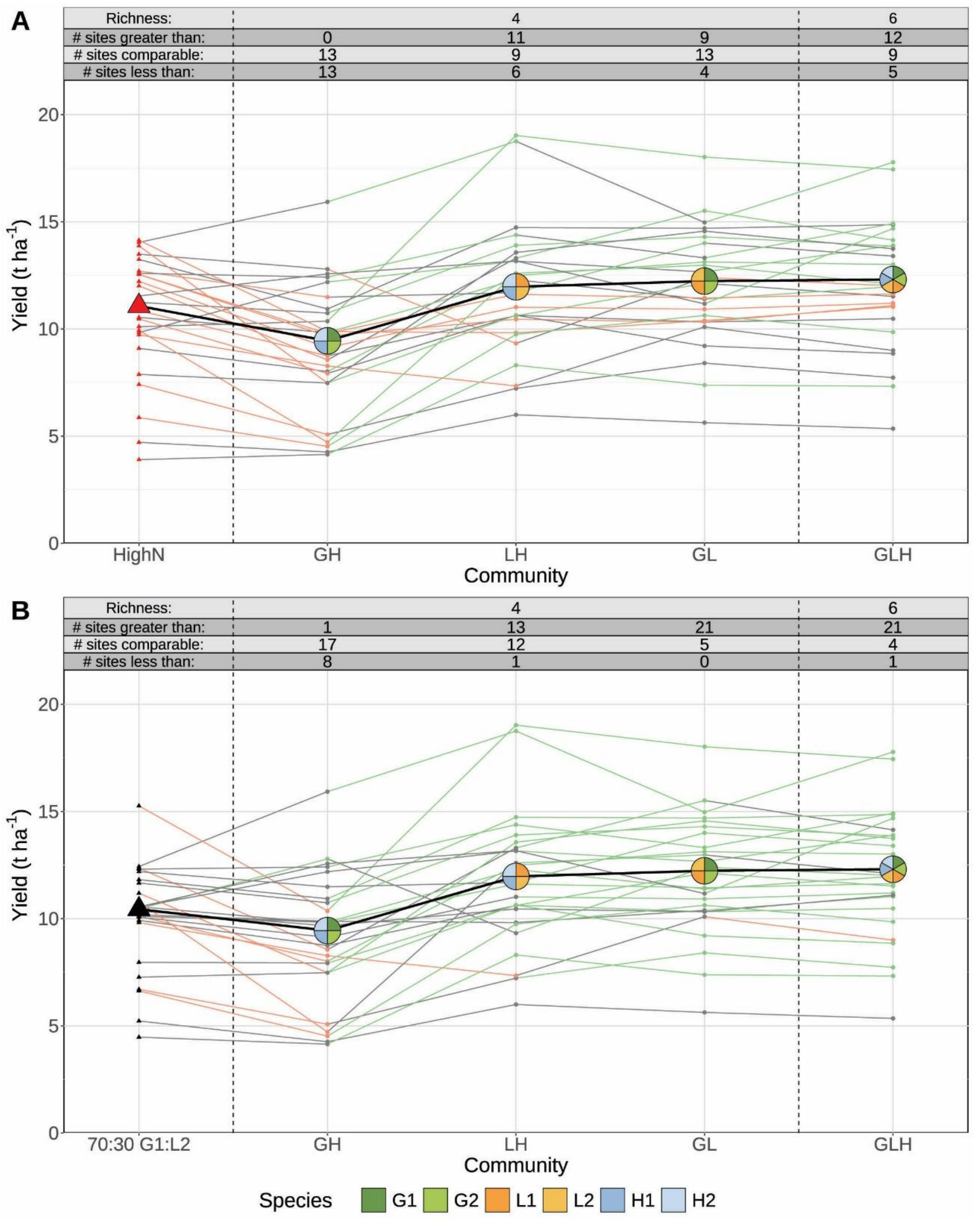


Fig. S6. Yields of four- and six-species equi-proportional mixtures were generally greater than, or comparable to, the high N grass and the 70:30 G1:L2 community at individual sites, despite high site-to-site variability. Predicted yields of selected communities across sites (black line and pie-glyphs) and at each site (colored lines and smaller points): the high N grass monoculture (red triangle), the 70:30 G1:L2 community (black triangle); the four-species equi-proportional grass-herb (GH), legume-herb (LH) and grass-legume (GL) mixtures; and the six-species equi-proportional GLH mixture (1/6:1/6: 1/6:1/6: 1/6:1/6 for G1:G2: L1:L2: H1:H2). The

four species communities are arranged in order of increasing predicted yield across all sites (calculated from estimates in Table S2). The communities are joined by lines to show the predictions specific to each site (values in Table S4); both the color of a site prediction and the line to the left of it are green if the prediction was more than 2 SEs greater than the reference community (left-hand side of each panel), grey if it was comparable (within ± 2SE), and orange if more than 2 SEs lower. The predicted comparison community is represented by a triangle: (**A**) red for the high N grass monoculture, and (**B**) black for the 70:30 G1:L2 two-species community over all sites (large) and for each individual site (small), where G1 was *L. perenne* and L2 was *T. repens* at the majority of sites. The number of sites where the comparison is greater than, comparable to, or less than the reference community is listed along the top row for each mixture community. Site-level predictions were made using the DI model fitted to data from each site separately.

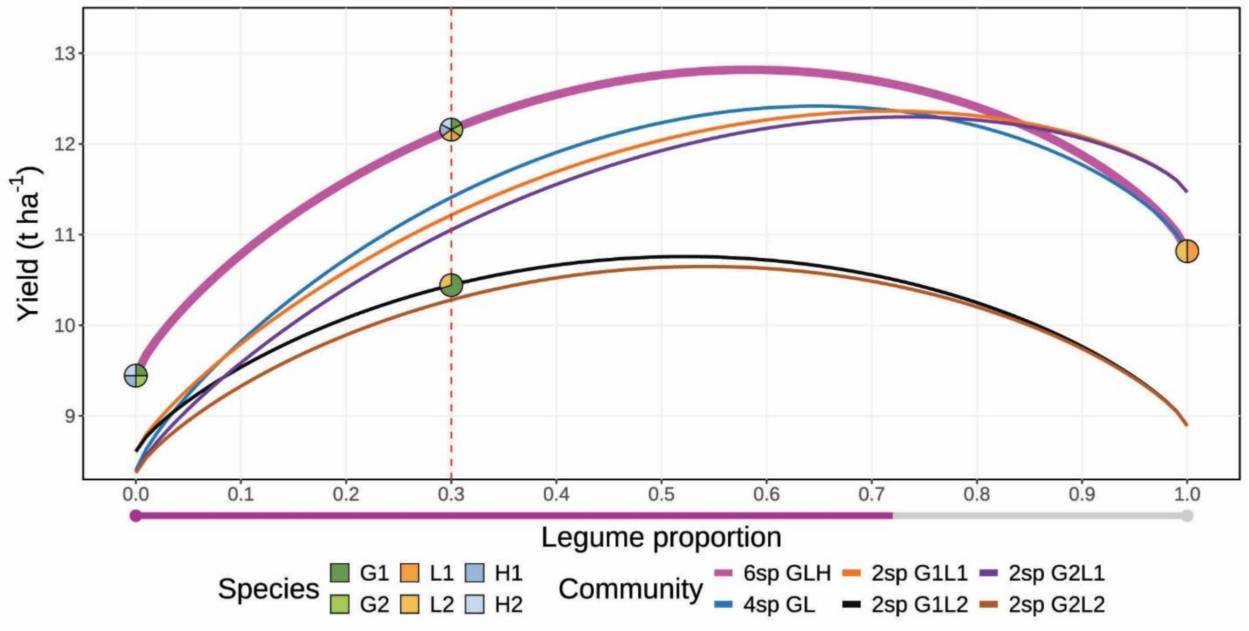


Fig. S7 At the same proportion of legumes, six-species mixtures of grasses, legumes, and herbs performed better than two- and four-species grass-legume communities with sown legume proportions < 0.7. Curves show the predicted yield of the six-species grass-legume-herb communities (bold magenta curve) and two- and four-species grass-legume communities versus sown legume functional group proportions (varying from 0 to 1). For the four- and six-species communities, species within each functional group have equal sown proportions, e.g., for sown legume proportion of 0.4, these communities would be: 0.3: 0.3: 0.2: 0.2: 0: 0 and 0.15: 0.15: 0.2: 0.2: 0.15: 0.15 respectively. The dashed red vertical line indicates the sown legume proportion = 0.3, with pie-glyphs highlighting the six-species mixture 0.175: 0.175: 0.15: 0.15: 0.175: 0.175 and the two-species 0.7:0.3 G1:L2 mixture. The pie-glyphs on the magenta line illustrate how the sown species proportions vary across the curve (noting that the lower end has legume = 0 and is a grass-herb equi-proportional four-species mixture, while the upper end has legume = 1 and is a 50:50 L1:L2 mix, while all other points on the curve are six-species mixtures). The magenta line underneath the x-axis indicates the range of sown legume proportion values for which the six-species curve performed better than all other curves (tested at each point vertically along the sown legume proportion gradient). While not shown, the six-species mixtures curve is significantly higher than the G1L2 and G2L2 curves at all sown legume proportions.

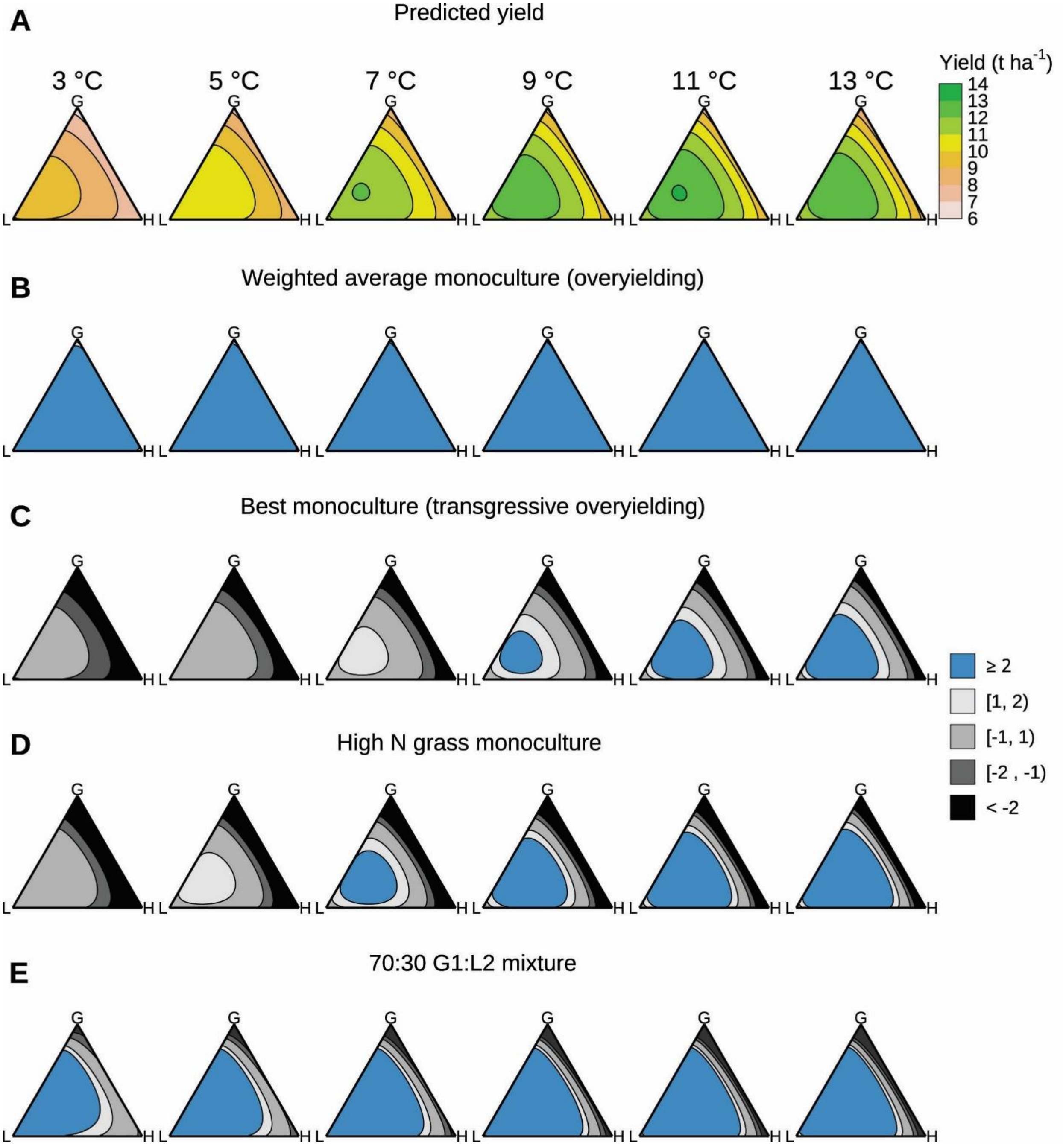


Fig. S8. Across the temperature gradient, mixture communities performed strongly in comparison to selected communities. Across the daily temperature gradient from 3ºC to 13ºC, the ternary diagrams show the predicted yield (row A) and the regions where mixtures performed significantly better than the weighted average monoculture (row B, overyielding), the best-performing monoculture (row C, transgressive overyielding), the high N grass monoculture (row D), and the 70:30 G1:L2 community (row E).

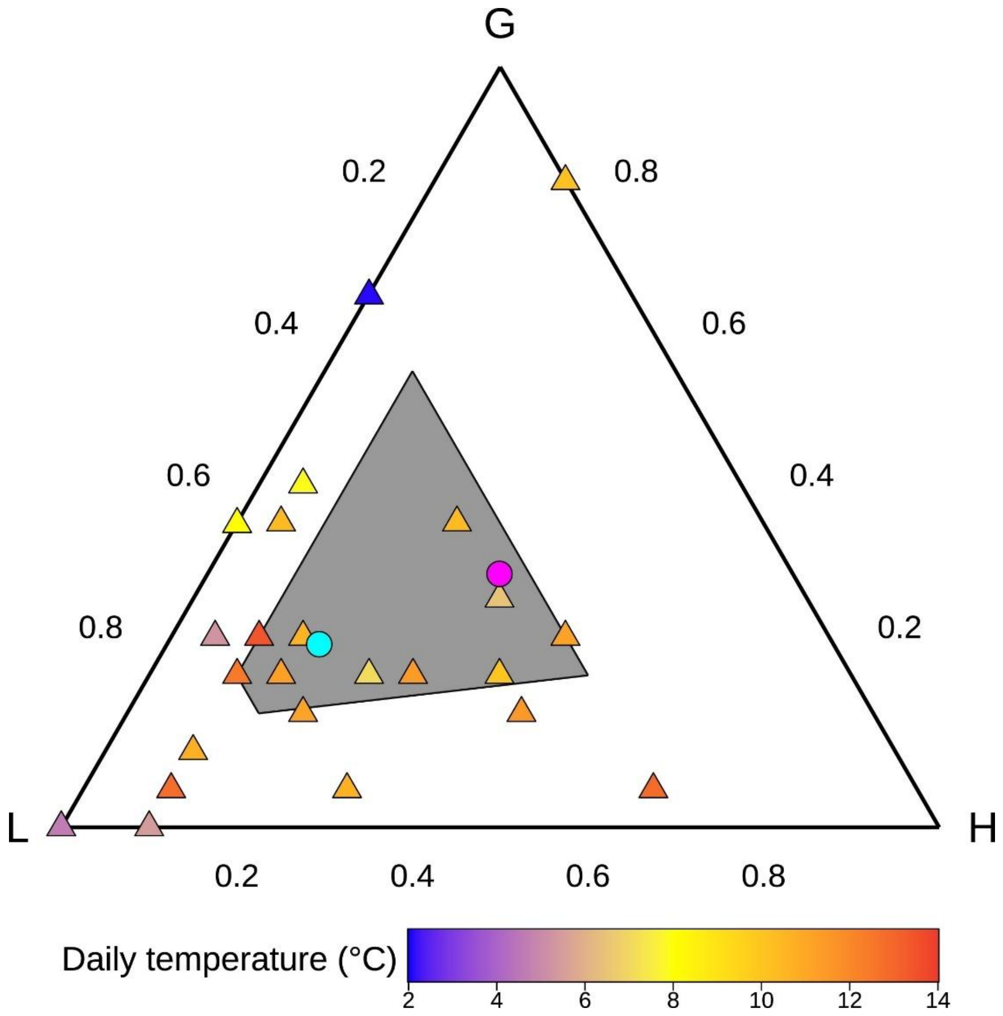


Fig S9. The best-performing mixtures at each site lie within, or close to, the optimal region. The sown functional group proportions for the 'best mixture' from each site are represented by the location of triangle symbols, with each triangle colored according to the average daily temperature of the site. The cyan and magenta circles highlight the sown functional group proportions (G:L:H = 0.24:0.59:17) that gave the highest predicted yield (12.83 t ha^{-1}) and the six-species equi-proportional mixture (12.31 t ha^{-1}), respectively.

Comm	Reps	N	G	L	H	G1	G2	L1	L2	H1	H2
1	3	Moderate	1	0	0	1	0	0	0	0	0
2	3	Moderate	1	0	0	0	1	0	0	0	0
3	3	Moderate	0	1	0	0	0	1	0	0	0
4	3	Moderate	0	1	0	0	0	0	1	0	0
5	3	Moderate	0	0	1	0	0	0	0	1	0
6	3	Moderate	0	0	1	0	0	0	0	0	1
7	1	Moderate	1	0	0	0.5	0.5	0	0	0	0
8	1	Moderate	0	1	0	0	0	0.5	0.5	0	0
9	1	Moderate	0	0	1	0	0	0	0	0.5	0.5
10	1	Moderate	0.5	0.5	0	0.5	0	0.5	0	0	0
11	1	Moderate	0.5	0.5	0	0.5	0	0	0.5	0	0
12	1	Moderate	0.5	0	0.5	0.5	0	0	0	0.5	0
13	1	Moderate	0.5	0	0.5	0.5	0	0	0	0	0.5
14	1	Moderate	0.5	0.5	0	0	0.5	0.5	0	0	0
15	1	Moderate	0.5	0.5	0	0	0.5	0	0.5	0	0
16	1	Moderate	0.5	0	0.5	0	0.5	0	0	0.5	0
17	1	Moderate	0.5	0	0.5	0	0.5	0	0	0	0.5
18	1	Moderate	0	0.5	0.5	0	0	0.5	0	0.5	0
19	1	Moderate	0	0.5	0.5	0	0	0.5	0	0	0.5
20	1	Moderate	0	0.5	0.5	0	0	0	0.5	0.5	0
21	1	Moderate	0	0.5	0.5	0	0	0	0.5	0	0.5
22	1	Moderate	0.33	0.33	0.33	0.33	0	0.33	0	0.33	0
23	1	Moderate	0.33	0.33	0.33	0.33	0	0.33	0	0	0.33
24	1	Moderate	0.33	0.33	0.33	0.33	0	0	0.33	0.33	0
25	1	Moderate	0.33	0.33	0.33	0.33	0	0	0.33	0	0.33
26	1	Moderate	0.33	0.33	0.33	0	0.33	0.33	0	0.33	0
27	1	Moderate	0.33	0.33	0.33	0	0.33	0.33	0	0	0.33
28	1	Moderate	0.33	0.33	0.33	0	0.33	0	0.33	0.33	0
29	1	Moderate	0.33	0.33	0.33	0	0.33	0	0.33	0	0.33
30	1	Moderate	0.5	0.5	0	0.25	0.25	0.25	0.25	0	0
31	1	Moderate	0.5	0	0.5	0.25	0.25	0	0	0.25	0.25
32	1	Moderate	0	0.5	0.5	0	0	0.25	0.25	0.25	0.25
33	3	Moderate	0.33	0.33	0.33	0.17	0.17	0.17	0.17	0.17	0.17
34	5	High	1	0	0	1	0	0	0	0	0

Table S1.

Experimental design of the grassland communities established in the LegacyNet grassland phase. The variables are: a community identifier for each set of unique sown species proportions by N fertilizer level (comm), the number of replicate plots sown per community (reps); amount of N fertilizer applied (N: moderate or high); the sown proportions of each functional group (G: grasses, L: legumes, and H: herbs); and the sown proportions of each species (G1 to H2).

	Parameter	Variable	Fixed effect	Std. Error	p-value
Identity effects	β_1	Grass 1	8.61	0.568	
	β_2	Grass 2	8.38	0.554	
	β_3	Legume 1	11.47	0.759	
	β_4	Legume 2	8.89	0.807	
	β_5	Herb 1	9.19	0.672	
	β_6	Herb 2	8.51	0.596	
Interaction effects	δ_{GL}	Grass-legume	5.92	0.563	**< 0.0001**
	δ_{GH}	Grass-herb	1.67	0.343	**0.0002**
	δ_{HL}	Herb-legume	4.80	0.410	**< 0.0001**
	δ_{GG}	Grass-grass	-0.27	0.498	0.5895
	δ_{LL}	Legume-legume	1.88	0.510	**0.0002**
	δ_{HH}	Herb-herb	0.35	0.500	0.4798
	θ	Non-linear parameter	0.7816041	(Estimated using profile likelihood during the model selection process)	
Climate effects	γ	Centered mean daily temperature squared	-0.049	0.0395	**0.2283**
	τ_{GL}	Grasses-legumes × centered mean daily temperature	0.36	0.188	**0.0486**
	τ_{HL}	Herbs-legumes × centered mean daily temperature	0.45	0.128	**0.0045**
	α	High nitrogen grass	11.07	0.620	

Table S2.

Diversity-Interactions model fitted to yield across sites. Fixed effect estimates, standard errors, and p-values from the final model. The model selection process showed that the fixed interaction effects were dictated by functional group membership, and that the theta parameter, θ, was significantly different from one (and thus incorporated a non-linear form of the species interactions, see Fig. S1). The parameters are grouped by species identity effects (beta values, β), species interaction effects (delta values, δ, plus theta, θ), climate effects (gamma, γ, and tau, τ), and the high N grass (alpha, α).

A

Effect	G1	G2	L1	L2	H1	H2	GH	GL	HL	N High
G1	5.6273	4.6230	4.6230	4.6230	4.6230	4.6230	-0.0139	-0.0139	-0.0139	4.9099
G2	4.6230	5.1997	4.6230	4.6230	4.6230	4.6230	-0.0139	-0.0139	-0.0139	4.9099
L1	4.6230	4.6230	12.1812	4.6230	4.6230	4.6230	-0.0139	-0.0139	-0.0139	4.9099
L2	4.6230	4.6230	4.6230	14.1401	4.6230	4.6230	-0.0139	-0.0139	-0.0139	4.9099
H1	4.6230	4.6230	4.6230	4.6230	8.9513	4.6230	-0.0139	-0.0139	-0.0139	4.9099
H2	4.6230	4.6230	4.6230	4.6230	4.6230	6.4500	-0.0139	-0.0139	-0.0139	4.9099
GH	-0.0139	-0.0139	-0.0139	-0.0139	-0.0139	-0.0139	1.5450	1.3532	1.3532	0.9957
GL	-0.0139	-0.0139	-0.0139	-0.0139	-0.0139	-0.0139	1.3532	6.3869	1.3532	0.9957
HL	-0.0139	-0.0139	-0.0139	-0.0139	-0.0139	-0.0139	1.3532	1.3532	2.6589	0.9957
N High	4.9099	4.9099	4.9099	4.9099	4.9099	4.9099	0.9957	0.9957	0.9957	7.2578

B

Site ID	Residual variance	Site ID	Residual variance
CA1	0.7202	IE3	1.3209
CA2	0.5067	IE4	2.9323
CH1	3.0575	IT2	1.2997
CN1	4.4206	NL1	0.5132
CN2	1.4423	NL2	0.3704
CN3	4.4031	NO1	0.7990
CZ1	1.8650	NO2	1.9568
DE1	0.7898	NO3	0.4991
DE2	3.1946	NZ1	1.6897
DK1	0.5641	PL1	0.6404
FR1	1.2467	UK1	3.5194
IE1	0.3669	US1	0.4218
IE2	0.6038	US2	0.3689

Table S3.

Variance parameter estimates for (A) the random coefficients and (B) the within-site error terms. (**A**) The variance-covariance parameter estimates for the random site-to-site effects. (**B**) The estimated within-site residual error variance for each site.

	Conventional practices		Mixtures			
Site	High N	70:30 G1:L2	GH	LH	GL	GLH
Across sites	11.07	10.44	9.45	11.97	12.23	12.31
CA1	7.41	6.70	5.08	7.23	8.41	7.73
CA2	9.73	9.82	8.27	7.34	10.10	9.00
CH1	12.01	12.36	8.56	13.3	15.51	14.14
CN1	14.04	12.43	15.93	18.75	14.97	17.78
CN2	4.71	5.23	4.26	6.00	5.63	5.35
CN3	10.11	15.26	10.36	19.03	18.02	17.44
CZ1	13.49	10.45	12.79	9.33	12.40	12.01
DE1	14.14	10.18	9.88	9.82	10.39	11.03
DE2	13.24	11.82	10.93	14.73	14.69	14.86
DK1	10.86	9.92	8.75	10.62	12.13	11.52
FR1	3.90	4.47	4.15	8.31	7.38	7.33
IE1	10.54	10.12	9.83	12.29	11.39	11.98
IE2	9.10	10.12	8.02	12.49	13.13	13.01
IE3	10.47	7.97	7.93	10.64	10.33	10.48
IE4	9.83	10.55	12.19	13.16	12.66	13.89
IT2	11.24	11.66	10.74	13.90	14.30	13.86
NL1	12.22	10.07	9.13	11.02	10.92	11.20
NL2	12.54	10.56	9.79	11.62	11.44	11.63
NO1	12.54	10.36	9.67	10.45	10.32	11.10
NO2	9.93	10.77	4.71	12.60	12.98	12.09
NO3	12.70	12.20	11.49	11.69	14.01	13.40
NZ1	13.88	11.19	7.48	13.57	14.56	13.74
PL1	12.60	12.31	12.42	14.38	13.32	14.92
UK1	11.54	10.55	12.57	13.17	11.17	14.69
US1	7.88	7.27	7.48	10.64	9.21	8.85
US2	5.87	6.63	4.52	9.75	10.63	9.86

Table S4.

Average and site-level predicted yields (t ha^{-1}) for selected communities. The high N grass and the two-species 70:30 G1:L2 communities reflect management practices that are widespread in temperate sown grasslands. The predictions across sites (first row) are calculated using the fixed effects estimates from the final model (Table S2), for example, the predicted yield of the four-species equi-proportional grass-legume mixture (GL) is:

$$12.23 = 8.61\left(\frac{1}{4}\right) + 8.38\left(\frac{1}{4}\right) + 11.47\left(\frac{1}{4}\right) + 8.89\left(\frac{1}{4}\right) + 5.92\left[4 \times \left(\frac{1}{4} \times \frac{1}{4}\right)^{0.7816041}\right] - 0.27\left(\frac{1}{4} \times \frac{1}{4}\right)^{0.7816041} + 1.88\left(\frac{1}{4} \times \frac{1}{4}\right)^{0.7816041}.$$

Similarly, we can predict from the model for any combination of sown proportions of G1, G2, L1, L2, H1 and H2, with moderate N fertilizer management. The predicted yield of the high N grass across sites is 11.07; note that we cannot predict yield for any other monoculture or mixture at high N, however, we can construct contrasts using our model to compare any monoculture or mixture at moderate N to the high N grass. Predictions at the site level are calculated by fitting the final model (excluding climate variables) to the data from each individual site.

"OMalleyetal_Data_S1.csv" is publicly available on Dryad (*51*)

Data S1.

Site information data. For each site the following information is provided: unique site identification code (site ID), country, latitude, longitude, the names of the six species sown at the site (G1, G2, L1, L2, H1, H2), the date of sowing of the experiment, the grassland phase duration, the rates of moderate and high nitrogen fertilizer, the average daily temperature and precipitation at the site over the experimental period. Short-hand notation for the sown species: *L. perenne - Lolium perenne, P. pratense - Phleum pratense, B. inermis - Bromus inermis, F. arundinacea - Festuca arundinacea, E. dahuricus - Elymus dahuricus, F. krasanii - Festulolium krasanii, D. glomerata - Dactylis glomerata, T. pratense - Trifolium pratense, T. repens - Trifolium repens, O. viciifolia - Onobrychis viciifolia, M. sativa - Medicago sativa, L. corniculatus - Lotus corniculatus, C. intybus - Cichorium intybus, P. lanceolata - Plantago lanceolata, S. divaricata - Saposhnikovia divaricata, P. asiatica – Plantago asiatica.*