

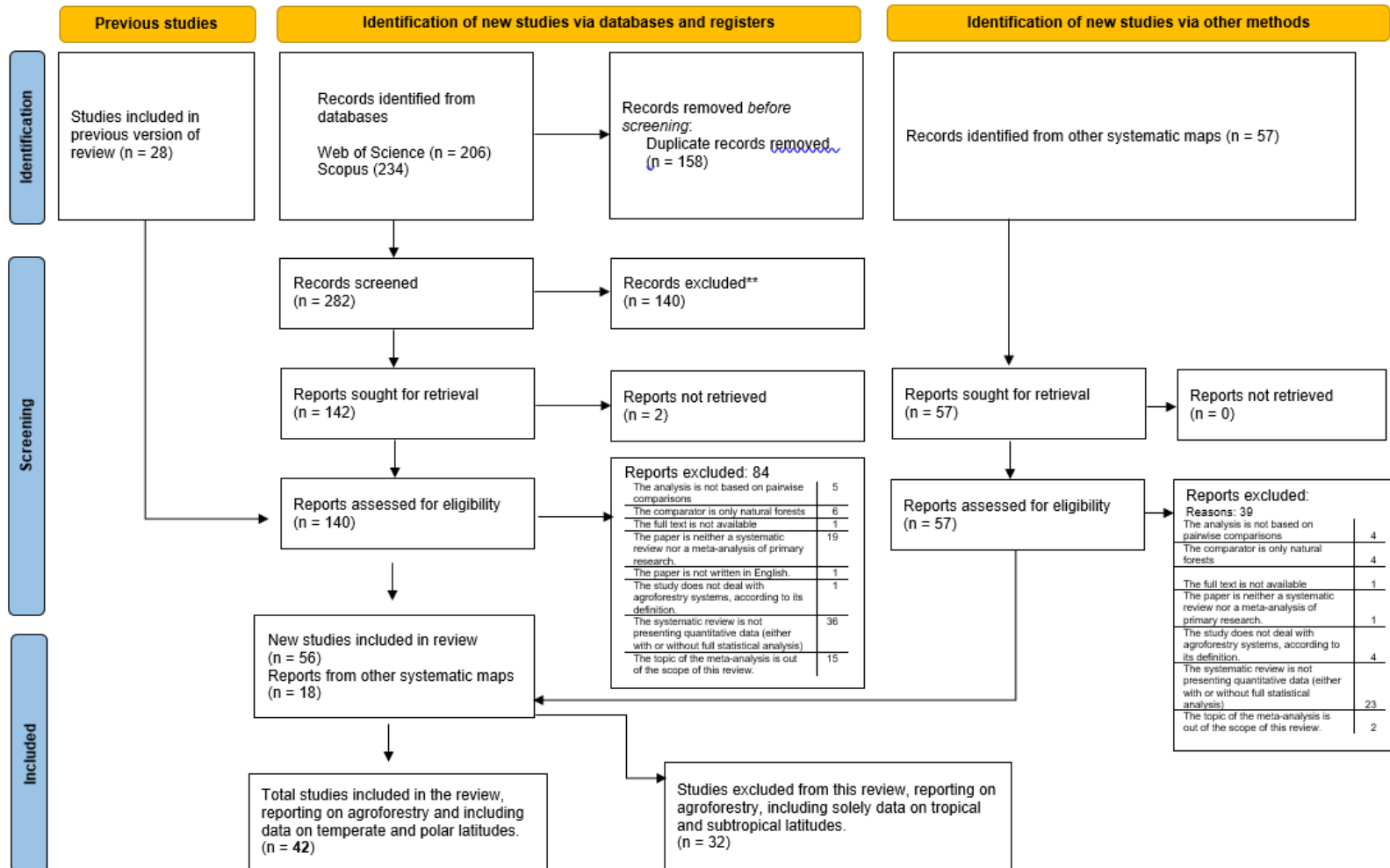
## Supplementary Materials

<b>Supplementary Figures.....</b>	<b>2</b>
Supplementary Figure 1.....	3
Supplementary Figure 2.....	4
Supplementary Figure 3.....	5
Supplementary Figure 4.....	6
Supplementary Figure 5.....	7
Supplementary Figure 6.....	9
<b>Supplementary tables (see spreadsheet attached).....</b>	<b>10</b>
Supplementary Table S 1.....	10
Supplementary Table S 2.....	10
Supplementary Table S 3.....	10
Supplementary Table S 4.....	10
Supplementary Table S 5.....	10
Supplementary Table S 6.....	10
Authors's contributions table.....	10
<b>Statistical Tests.....</b>	<b>11</b>
Bayesian ordinal model for comparing pairs of donuts (Temperate vs main results).....	11
Statistical test on Quality scores (Figure S5-A).....	12

## Supplementary Figures

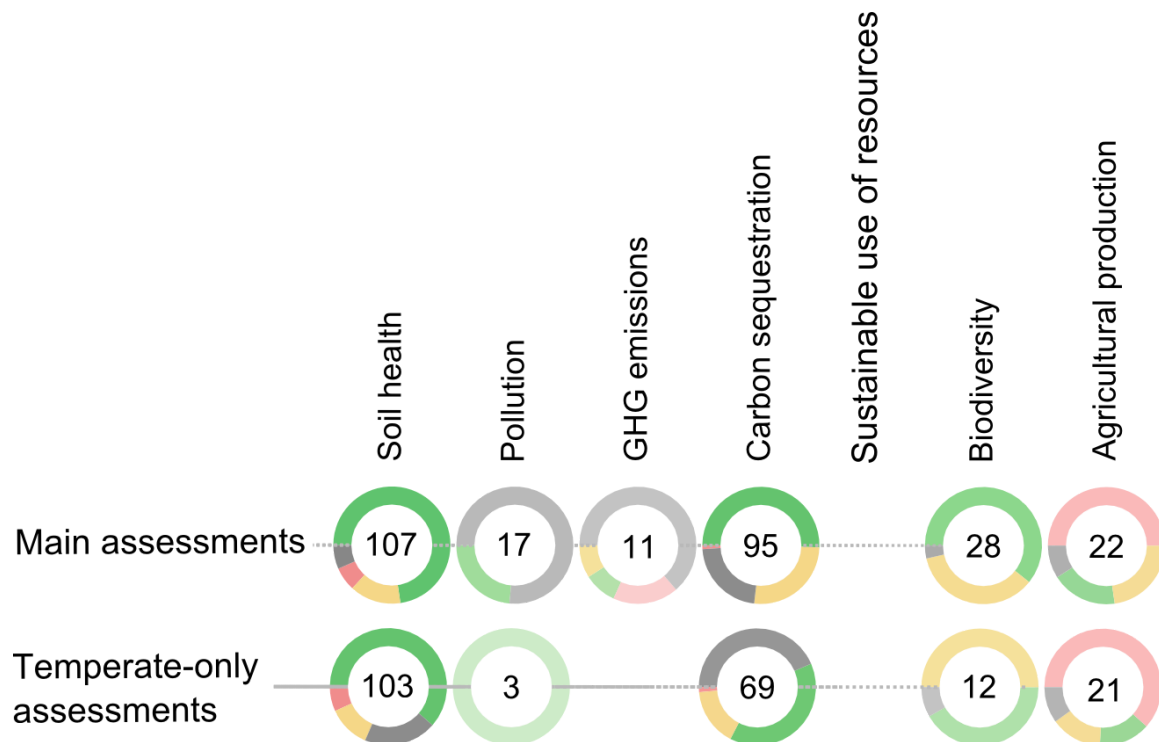
# Supplementary Figure 1

## PRISMA statement diagram



## Supplementary Figure 2

Overview of the main results. Results (i.e. mean effect sizes extracted from the select MAs) were extracted for main assessments, as well as for temperate-only subgroups. Metrics used in empirical studies (as reported by MAs) are grouped across main “sustainability outcomes”, in accordance to the classification reported in the JRC-Farming practices Evidence Library ([European Commission. Joint Research Centre., 2025](#)). Donut plots show the share of results showing significant positive (green) or negative effects (red), non-significant effect (yellow) or non-statistically-tested results (grey), as presented by the selected MAs. The numbers report the total count of effect sizes reported by the selected MAs. The full PRISMA statement diagram (Page et al., 2021) is reported in Supplementary Figure 1. Detailed classification of specific metrics belonging to each class of outcomes is available in Supplementary Table 4.

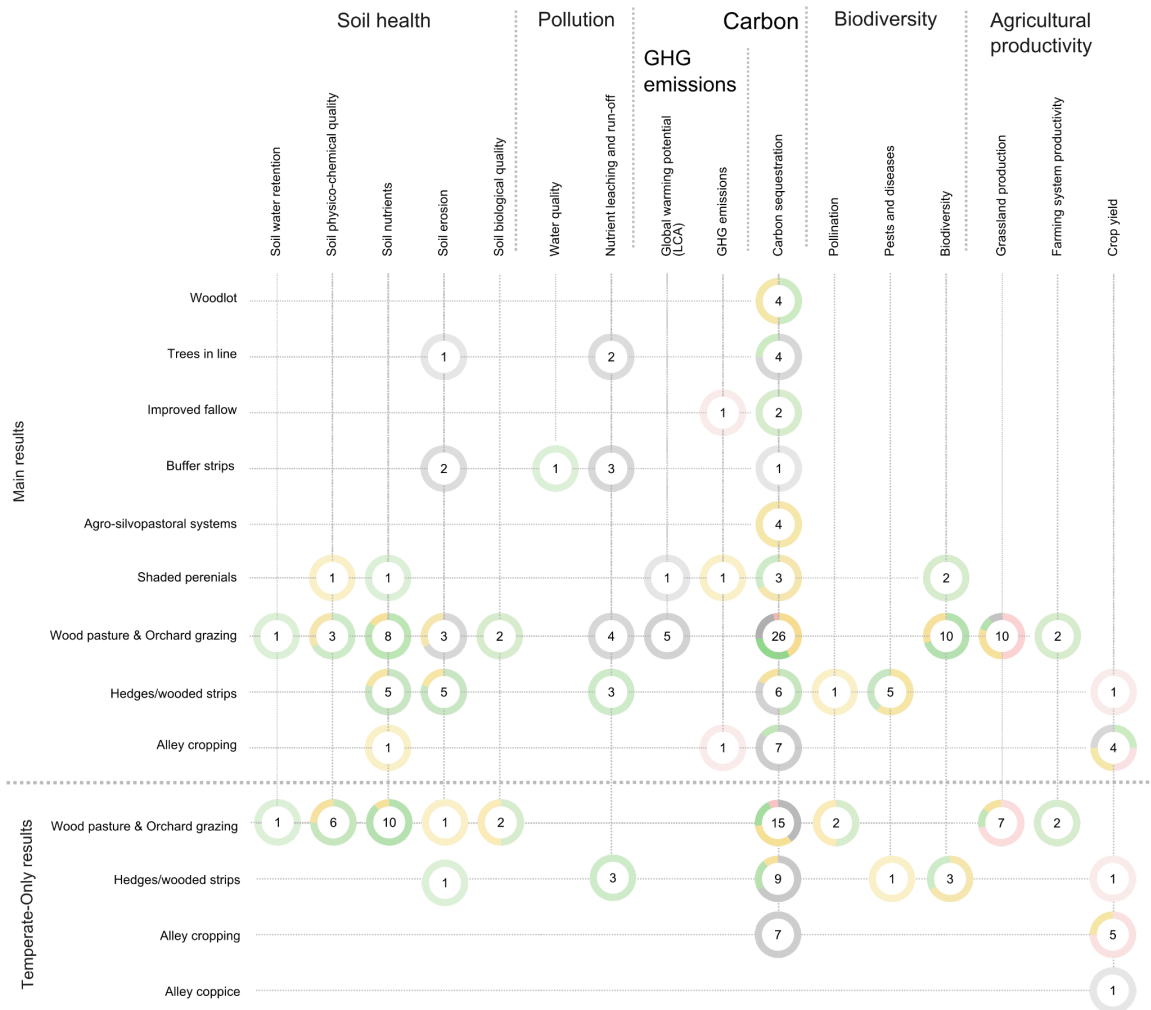


## Supplementary Figure 3

Evidence map of main classes of agroforestry practices vs ecosystems services classes (4 digits). Results (i.e. mean effect sizes extracted from the select MAs) were extracted for main assessments, as well as for temperate-only subgroups.



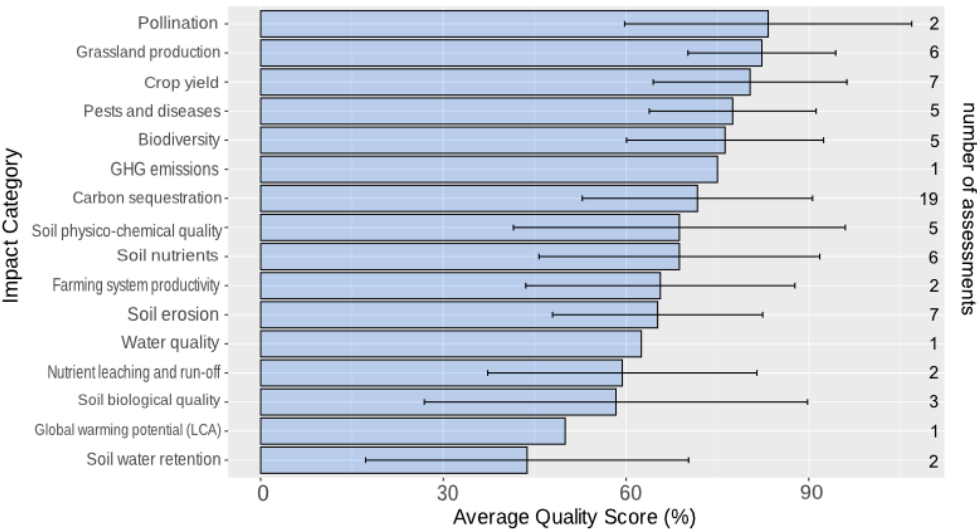
Agroforestry systems/practices (level 3) vs Ecosystem services class. Results (i.e. mean effect sizes extracted from the select MAs) were extracted for main assessments, as well as for temperate-only subgroups.



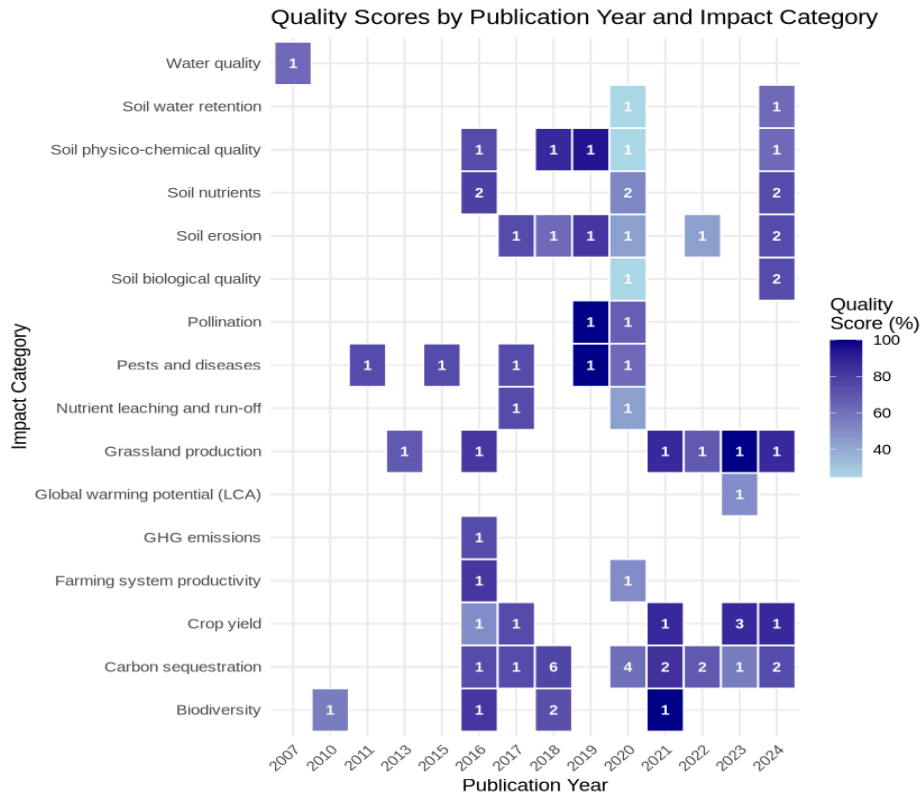
Supplementary Figure 5

Distribution of average (and standard deviation) quality scores across different impact categories (A) and over time (publication year) (B). Digits in both plots indicate the number of assessments available for each impact category.

A



B





## Supplementary Figure 6

Distribution of coordinates retrieved in primary studies across Whittaker biomes.



## Supplementary tables (see spreadsheet attached)

### Supplementary Table S 1

Search equations used in Web of Science and Scopus engines for retrieval of potentially relevant meta-analyses and systematic reviews reporting the effects of agroforestry.

### Supplementary Table S 2

List of selection (inclusion and exclusion) criteria used during the systematic screening of records.

### Supplementary Table S 3

Full critical appraisal document, reporting the details of the systematic screening and selection of relevant documents.

### Supplementary Table S 4

Classification of empirical metrics and their matching to the CICES classification

### Supplementary Table S 5

List and rationale of the 16 quality criteria used for the assessment of the selected meta-analyses.

### Supplementary Table S 6

Sections, divisions, groups and classes of the CICES classification covered with results by the selected MAs (blue shade).

### Authors's contributions table

## Statistical Tests

Bayesian ordinal model for comparing pairs of donuts (Temperate vs main results)

[https://github.com/dbeillouin/Ordinal\\_models](https://github.com/dbeillouin/Ordinal_models)

For “Main results” Dataset: (for CICES classes with enough data)

```
# A tibble: 3 × 6
  CICESgroupdivision mean_diff ci_low ci_high p_pos_gt_neg conclusion
  <chr>               <dbl> <dbl> <dbl> <dbl> <chr>
1 Atmospheric composition and conditions 0.213 -0.532 0.833 0.77 Moderate evidence Positive > Nega...
2 Cultivated terrestrial plants for nutrition, materials or energy -0.373 -0.815 0.273 0.0937 Evidence Negative >= Positive
3 Regulation of soil quality 0.688 0.538 0.830 1 Strong evidence Positive > Negati...
```

For *Atmospheric composition and conditions*, the mean probability difference was 0.21 (95% CI –0.53 to 0.83, posterior probability 0.77), providing moderate evidence that positive effects are somewhat more likely than negative effects. For *Cultivated terrestrial plants for nutrition, materials or energy*, the mean probability difference was –0.37 (95% CI –0.82 to 0.27, posterior probability 0.094), indicating that negative effects are slightly more likely than positive effects. For *Regulation of soil quality*, the mean probability difference was 0.69 (95% CI 0.54 to 0.83, posterior probability 1), meaning that, on average, the probability of a positive effect is 69 percentage points higher than that of a negative effect, providing strong evidence of beneficial outcomes.

For TempOnly: (for CICES classes with enough data)

```
# A tibble: 2 × 6
  CICESgroupdivision mean_diff ci_low ci_high p_pos_gt_neg conclusion
  <chr>               <dbl> <dbl> <dbl> <dbl> <chr>
1 Cultivated terrestrial plants for nutrition, materials or energy 0.0160 -0.342 0.408 0.532 No evidence of difference
2 Regulation of soil quality 0.478 -0.0308 0.826 0.970 Strong evidence Positive > Negat...
```

For *Cultivated terrestrial plants for nutrition, materials or energy*, there was no evidence of a difference between positive and negative effects (mean probability difference = 0.016, 95% CI –0.342 to 0.408, posterior probability 0.53). For *Regulation of soil quality*, the mean probability difference was 0.48 (95% CI –0.031 to 0.83, posterior probability 0.97), indicating that, on average, the probability of a positive effect is 48 percentage points higher than that of a negative effect. This provides strong evidence that interventions in soil quality are more likely to be beneficial than detrimental.

Comparison between Temperate-only and Main results (for CICES cat with enough data)

```
> df_compare_groups
# A tibble: 2 × 6
  CICESgroupdivision mean_diff ci_low ci_high p_full_gt_temp conclusion
  <chr>               <dbl> <dbl> <dbl> <dbl> <chr>
1 Cultivated terrestrial plants for nutrition, materials or energy -0.389 -0.960 0.327 0.115 No strong evidence
2 Regulation of soil quality 0.210 -0.172 0.734 0.825 No strong evidence
```

Comparison between the full and temperate-only datasets revealed no strong differences in predicted effect probabilities. *Regulation of soil quality* showed a modest, non-significant increase when including all regions (mean difference = 0.21, 95% CI –0.17–0.73), whereas *Cultivated terrestrial plants* showed a slight, non-significant decrease (mean difference = –0.39, 95% CI –0.96–0.33), indicating that restricting to temperate zones minimally alters overall patterns.

### Statistical test on Quality scores (Figure S5-A)

This ANOVA output suggests that there is no statistically significant difference in the means of the response variable across the different categories of Impact, as indicated by the high p-value (0.566) and the relatively low F value (0.902). Both tests indicate that there are no statistically significant differences in quality scores across impact categories ( $p > 0.05$ ).

*Table xx . Anova and Kruskal-Wallis test applied to test the significance of differences between the average quality scores across impact categories (Figure 6 A).*

Anova	Df	Sum Sq	Mean Sq	F value	p-value
Impacts	15	5155	343.7	0.902	0.566
Residuals	58	22111	381.2		
Kruskal-Wallis test	15	chi-squared = 12.799			0.6179