Assessing environmental performance of agricultural practices in Europe using eco-system services: An environmental performance indicator approach

by K. Van Ruymbeke, J. Ferreira, V. Gkisakis, J. Kantelhardt, G. Manevska Tasevska, P. Matthews, A. Niedermayr, L. Schaller, K. Zawalinska, K. Mertens, and L. Vranken

Assessing environmental performance of agricultural practices in Europe using ecosystem services: An environmental performance indicator approach

Van Ruymbeke, K.[1]*, Ferreira, J.[2], Gkisakis, V.[3], Kantelhardt, J.[4], Manevska Tasevska, G.[5], Matthews, P.[6], Niedermayr, A.[4], Schaller, L.[4], Zawalinska, K.[7], Mertens, K.[1], Vranken, L.[1]

June 2021

*contact main author: kato.vanruymbeke@kuleuven.be

Abstract

Agroecosystems are one of the most important ecosystems for maintaining human wellbeing. Conventional farm management practices focus on maximizing production from these systems. Agroecologically-friendly farming, however, focuses more on reconciling production with regulating ecological processes and sociocultural identity through the provisioning of ecosystem service (ES), thereby maintaining both ecological and human wellbeing. Though many studies have evaluated the performance of (conventional and agroecological) agricultural practices against ES supply, this is mostly done by focussing on only a few practices simultaneously. We expand on this literature by incorporating 26 practices in an environmental performance assessment in order to draw more comprehensive conclusions on the delivery of ES from farm management practice in European agriculture.

A rapid evidence assessment of secondary literature was carried out evaluating the supply of 17 ES from 26 practices. Results were quantified by calculating indicators reflecting the potential supply of ES from management practices at the farm and territorial level. By incorporating a measure of evidence quantity and quality, the indicators also provides insight into the state of the currently available literature. Supplementary measures are reported alongside indicators to ensure transparency and to increase interpretability. Existing literature evaluates management practices most commonly against regulating and maintaining services, followed by provisioning services. However, this varied between the considered levels. At both farm and territorial level, the literature noticeably fails to evaluate cultural services. Disparities between the number of indicators calculated at both levels indicate a gap in the literature evaluating ES delivery from management practices at territorial level compared to farm level. Findings reflect not only the performance of a given management practice, but also the knowledge currently available in the literature, therefore knowledge gaps likely affect the estimated performance of the different practices.

[1] KU Leuven (Belgium), [2] SRUC (Scotland), [3] Demeter (Greece), [4] BOKU (Austria), [5] SLU (Sweden), [6] UNIKENT (England), [7] Irwin Pan (Poland)

1. Introduction

Agroecosystems are arguably one of the most important ecosystems to sustain human wellbeing. Not only do we rely on these systems for the provisioning of food and energy materials, we also derived many secondary benefits from them such as recreation, regulation of natural hazards and carbon sequestration. Historically, however, these systems have been primarily managed to sustain food production and other provisioning services (Sandhu et al., 2010; Swinton et al., 2007, 2006), with preservation of secondary benefits remaining largely on the backburner. That is not to say both primary and secondary benefits cannot be maintained simultaneously. Through well-planned and regulated farm management practices (both conventional and agroecological) we can manage agroecosystems to find a balance between meeting demands for productive output and maximizing environmental performance to ensure long-term sustainability (Bateman et al., 2009; Pretty, 2008a; Pretty and Bharucha, 2014; Wezel et al., 2017).

Conventional farm management practices focus on increasing productive output through the use of agrochemical inputs, traditional cropping rotations and limiting non-productive space. While such practices are effective at maintaining high yield, they may be detrimental to both environmental and human wellbeing when left unchecked (Swinton et al., 2007, 2006). Agroecological farm management practices, by contrast, attempt to improve environmental quality while maintaining agricultural production (Wezel et al., 2014). This is achieved by improving and maintaining energy flows, nutrient cycling, population-regulating mechanisms and overall system resilience (Pretty, 2008b). Practices such as the use of cover crops between crop cycles, conservation tillage and integrating various production systems (e.g. agroforestry) are only a few examples of such agroecological practices (Palomo-Campesino et al., 2018; Wezel et al., 2014). However, the implementation of these more environmentally-friendly management practices often comes paired with higher labour and energy demands (Wezel et al., 2014), challenges in overcoming cognitive and economic barriers (Jeanneaux, 2018; Lucas et al., 2018), as well as threats to yield from increased disease and pest damage (Wezel et al., 2014).

When evaluating the environmental performance of farm management practices, most studies adopt the ecosystem service (ES) concept (Turner and Daily, 2008). ES can be defined as the direct or indirect contribution of ecosystems to human well-being (Haines-Young and Potschin, 2018). The Common International Classification of Ecosystem Services (CICES) categorizes ES into three broad categories, i) regulating and maintaining services, which help maintain proper functioning of ecosystems (e.g. biodiversity), ii) provisioning services, which supply productive output that can be directly exploited (e.g. crop production), and iii) cultural services, which influence people's mental and physical wellbeing through non-material characteristics of an ecosystem (CICES, 2018). ES can also contain a spatial component, i.e. certain ES emerge only if a minimum scale threshold of specific service-providing processes/functions is met (Andersson et al., 2015). Relevant scale thresholds vary between ES from global (e.g. global climate mitigation) to plot level (e.g. pest control) (Andersson et al., 2015). The human dimension of the human-nature interactions captured by the ES concept is also spatially explicit, with demand for certain services often driven by socio-cultural and/or geographic conditions (Potschin and Haines-Young, 2011).

While the literature on environmental performance evaluation of farm management practices through the use of ES is very extensive (for examples see Laura et al., 2017; Toivonen et al., 2018; Van den Putte et al., 2010), only few studies incorporate more than a handful of practices and/or ES. Furthermore, many studies tend to focus on only a single spatial level. Despite this, there is ample evidence suggesting ES do not occur in isolation, nor does the impacts of farm management practices on ES ever truly follow a unidirectional path (Kragt and Robertson, 2014; Martín-López et al., 2014; Zhang et al., 2007; Zhou et al., 2019). In this paper we address this research gap by incorporating 26 farm management practices and 17 ES into an environmental performance assessment of farm management practices in the context of European agriculture. To achieve this we propose a novel approach by combining two commonly applied methodologies within the field of environmental performance assessments; a rapid evidence assessment (REA) and the composition of performance indicators.

The use of indicator-based methods to assess environmental performance of agriculture has risen in prevalence in recent decades (Bockstaller et al., 2008). Environmental performance indicators' popularity lies in their ability to provide an alternative to direct environmental performance measurement of farm management practices, which is often time-consuming, costly, and methodologically challenging (Bockstaller et al., 2008).

We attribute values to individual management practices based on their impacts on individual ES with evidence derived from the literature. Firstly we compile a database of synthesized results through an REA of secondary literature, delineating the impacts of 26 farm management practices on 17 ES across farming systems throughout Europe. Observations from this database are then used to quantify the impact of management practices on ES through the calculation of performance indicators at the farm and territorial level.

Through this exercise we provide what is to our knowledge, a first attempt at incorporating such two methodologies to gain a comprehensive overview of environmental performance of farm management practices in Europe. We address two main research questions: i) how do various farm management practices (both conventional and agroecological) impact the delivery of ES in agroecosystems across Europe?, and iii) how does environmental performance differ between farm and territorial level? Similarly to Rigby et al. (2001), we do not claim that the indicators presented in this paper are decisive of farm management practice impacts on ES. Rather we assert that the proposed indicator are valuable in that they provide a first attempt at summarizing the multitude of evidence available in the literature in a concise, intuitive and transparent manner. In this way we hope to gain a better understanding of the current state of affairs in the literature, identify where (and which) evidence is missing, and open up a discussion on how to go about utilizing the information we already have and filling the remaining information gaps.

2. Background

2.1. Ecosystem Services: supply and demand

The research objectives of this exercise are framed within an adapted version of the Ecosystem Service Cascade model as proposed by Potschin and Haines-Young (2011). The Cascade model presents a 'production chain' in which ecological structures and processes are linked through ES with human wellbeing through a five-tiered process (Figure 1). The model postulates that an ES may only be considered a service if human beneficiaries can be identified, thus making it evident that the derivation of benefits and values from ES is clearly a social construct dependent on the demand derived from contextual characteristics. On the other hand, the biophysical structures and functions that give rise to ES suggest an underlying ecological dimension to the model. Through incorporating these two dimensions, the Cascade model may be interpreted as a social-ecological system, in which humans are considered a part of – rather than separate from – nature (Folke, 2007). This interaction between the social and ecological dimensions is interpreted by Potschin and Haines-Young (2011, p582) as “some kind of supply-demand relationship”. Applying supply-demand relationships to ES is not a novel concept, and has been considered by various other authors (e.g. Burkhard et al., 2012; Burkhard and Maes, 2017; Nedkov and Burkhard, 2011). We therefore see a need to expand on the Cascade model by explicitly incorporating such supply and demand in its formulation.

As is illustrated in Figure 1, the ecological dimension of the Cascade model is commensurable with the concept of supply, whereby ecosystems – particularly the biophysical structures and functions underpinning ecosystems – supply ES. At the same time, the contextual characteristics specified in the social dimension of the model may be considered the underlying drivers determining demand for particular ES. These drivers make supply of and demand for particular ES spatially and temporally explicit (Potschin and Haines-Young, 2011). Certain services may only be supplied at certain spatial scales (e.g. on- vs off-farm services in the case of agroecosystems) or during certain times of the year (e.g. crop yield). Likewise, demand for ES will differ across different geographic regions, between different end-users, as well as through time (Potschin and Haines-Young, 2011). For example, in regions particularly prone to droughts, farmers will likely experience a greater demand for soil water retention during periods of drought to ensure stable crop yields. Simultaneously, the wider community in the same area may have a higher demand for water quality and the sufficient supply of drinking water. In non-drought prone regions, the demand for these ES would be much lower.

Using this understanding of how supply of and demand for ES may vary across space and time, we expand on Potschin and Haines-Young's (2011) core model-concept, and postulate that a service is delivered only if the demand for said service in the social dimension overlaps spatially and temporally with the supply of said service in the ecological dimension. Without this spatial and temporal overlap, ES may be supplied by the ecological dimension and may be demanded by the social dimension, but there will be no delivery. Identifying delivery of ES requires a clear delineation of both supply and demand levels. Supply of ES may be spatially delineated based on purely geographic levels (i.e. plot, regional, national and global). Delineating demand requires consideration of end-users (e.g. farmers vs wider society) and the geographical location at which the demand is exercised, i.e. locally (e.g. recreation) or non-locally (e.g. carbon sequestration).

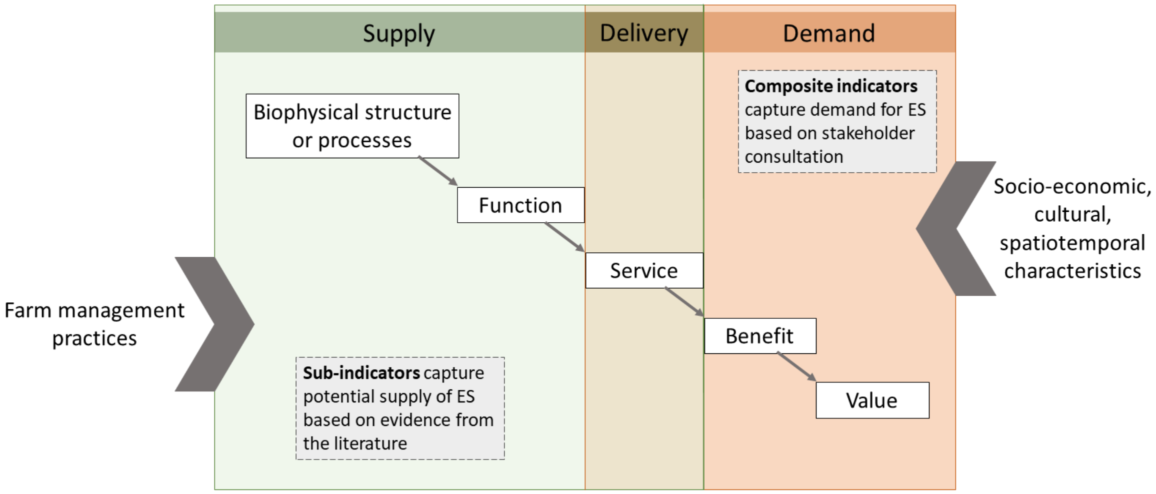

Supply
Delivery
Demand
Biophysical structure or processes
Function
Service
Benefit
Value
Composite indicators capture demand for ES based on stakeholder consultation
Sub-indicators capture potential supply of ES based on evidence from the literature
Farm management practices
Socio-economic, cultural, spatiotemporal characteristics

Figure 1. ES cascade model framework adapted from Potschin and Haines-Young (2011), specifying supply, demand and delivery of ES captured by sub- and composite indicators.

2.2. Environmental performance indicators

Performance indicators are widely used to represent the current state of affairs across many different domains, including the Human Development Index (HDI), the measure of GDP, and the Index for Biological Integrity (Gan et al., 2017; Oecd and JRC, 2008; United Nations Development Programme, 2013). Performance indicators are clearly useful tools in that they summarize complex, multi-dimensional concepts in a way that is easy to interpret (Gómez-Limón and Sanchez-Fernandez, 2010). In providing an alternative to direct impact measurement, indicator composition also provides an opportunity to substantially expand the scope of the assessment while at the same time overcoming methodological difficulties associated with practical (e.g. time and budget) and measurement constraints (Bockstaller et al., 2008). Nonetheless, performance indicators are often criticized for their subjectivity, for lacking transparency, and for being prone to mis-interpretation (Gómez-Limón and Sanchez-Fernandez, 2010). Increasing transparency and supplementing indicators with additional measures may help address some of these criticisms. For example, subjectivity often stems from the choice of weights in indicator aggregation. Increasing transparency in reporting these weights therefore not only addresses concerns about subjectivity, it also improves understandability of indicators and may reduce occurrences of mis-interpretation (Gómez-Limón and Sanchez-Fernandez, 2010).

Indicators can be either measured, estimated, modelled, or calculated by aggregation of data (Girardin et al., 1999). The approach adopted is often determined by the theoretical framework, as well as the aim and purpose of the indicators (Nardo et al., 2005). Through their ability to summarize complex information by incorporating a variety of components, performance indicators provide an opportunity to capture the various ecological (supply) and social (demand) drivers within an agroecological system as described in section 2.1. As policy makers become increasingly interested in using ES to inform policy decisions regarding biodiversity conservation, spatial planning and resource management (van Oudenhoven et al., 2018), performance indicators provide a unique opportunity to facilitate the decision making process.

3. Materials and Methods

In order to estimate the environmental performance of farm management practices on ES supply in European agriculture, we compose performance indicators per farm management practice – ES combination. Indicators are calculated through the aggregation of existing evidence in the secondary literature derived from an extensive REA.

3.1. Data collection: Rapid Evidence Assessment

3.1.1. Selection of papers and inclusion criteria

The comprehensive search string from which articles were derived was composed through an iterative process. This process consisted of formulating a search string for the individual management practices, combining these into a composite search string, and then evaluating the search string results against the inclusion in a set of pre-defined reference articles. The comprehensive search string (Table S2), the full list of management practices (Table S1), as well as the reference articles included in this assessment can be found in the supplementary materials.

Input provided by eight researchers across eight European countries was combined with the extensive list of European management practices identified by Rega et al. (2018) to select the most relevant practices for inclusion in this exercise. Rega et al. (2018) performed a literature review to identify which management practices were associated with which farming systems in Europe. It was opted to utilise the resulting 36 practices as the methodology adopted by Rega et al. (2018) was transparent, their focus on European agricultural systems matched our study focus, and the list of farm management practices they put forward can be validated against the literature (Altieri, 2000; Bourguet and Guillemaud, 2016; Wezel et al., 2015, 2014). Consultation with researchers did not identify any missing management practices.

Prior to carrying out the REA, a PICO with a clear ex ante delineation of Population, Intervention/Exposure, Comparator and Outcome, was established based on the research questions (Table 1). The PICO components were used as inclusion/exclusion criteria for sample selection in the REA. Reviewers consisted of 13 researchers from 9 research institutions across Europe. Through this process, a total of 647 articles were selected for inclusion based on title and abstract screening. Reviewers extracted meta-analytic data from the articles such as type of review, location of study, considered management practice(s), level of assessment (i.e. farm or territorial level), and ES assessed. Farm and territorial level of assessment were identified based on the spatial scale specified within the article, referring to the scale at which the ES were measured. A targeted selection of the 647 articles was conducted for full text screening. Targeted sampling consisted of, where possible, selecting five articles (of which one a meta-analysis) per management practice. This resulted in a total of 105 articles that were included in the final REA. At full text screening, 10 more articles were excluded based on exclusion criteria, resulting in a final corpus of 95 articles.

Table 1. PICO (Population, Intervention, Comparator, Outcome) used to establish inclusion criteria for the REA.

PICO	Component	Objective
Population	Quantitative or qualitative secondary literature	Robustly inform environmental assessment indicators using pre-existing literature reviews, quick scoping reviews, rapid evidence assessments, meta-analyses, systematic reviews and reviews of reviews; quantitative and qualitative data was selected to be input into indicators.

Population	European agricultural land	Use the most locally-relevant data on practices and their effects.
Intervention/exposure	Farm management practices	Cover the variety of practices to be included in the environmental assessment.
Comparator	Conventional agricultural practices	Compare conventional approaches to agriculture with more agroecological approaches (embedded within the literature reviews).
Outcome	Ecosystem service provision	Measure, through the use of indicators, proxies or qualitative data, the impact of adoption of management practices on ES supply.

3.1.2. Data extraction

The majority of the articles within the corpus were non-systematic literature reviews (70.65%). A further 21.74% were meta-analyses, 3.26% were systematic reviews and 4.35% were not considered a review of a specific type. Overall, the majority of articles reported global results (56.52%) of which only results relevant for Europe were extracted. Of the articles specifically considering European case studies, 34.38% considered Europe broadly, and 28.13% reported results from Northwestern Europe. Few articles reported results from South-eastern and Eastern Europe.

For each article, qualitative, semi-quantitative and quantitative data for the link between management practices and supply of an ES was extracted into an excel database. Semi-quantitative data was expressed as either a positive, inconclusive, or a negative relationship between the management practices evaluated and the ES assessed. As semi-quantitative data was extracted for all articles, but quantitative data was not, it was opted to use only the former for indicator calculation. As such, observations are henceforth defined as semi-quantitative observations reflecting the supply of an ES from a management practice, which was coded as 1 (negative), 2 (inconclusive) and 3 (positive). As the REA concerned secondary literature, multiple observations of the same management practice-ES link could be extracted from a single article.

During full-text screening, partners also evaluated the quality of each article across 26 standardized quality criteria adapted from Beillouin et al. (2019) and PRISMA (2015) (Table S3, supplementary materials). (Beillouin et al., 2019). The criteria reflect the quality of the articles across all steps of the review process, including literature search, data extraction, data analysis and interpretation (Beillouin et al., 2018). Reviewers were asked to indicate for each of the 26 criteria whether it was addressed (yes/no) in the article under consideration. A single final quality score, ranging between 0 and 1,was attribute to each article based on the performance across the 26 quality criteria. This score was calculated by weighting the different criteria for the degree to which they contribute to the quality of an article from 0 to 1. Weighted scores were then aggregated across the criteria to obtain a single quality score for each article.

3.2. Sub-Indicator calculation

Indicators reflect the potential supply of an ES from a single farm management practice in the context of European agriculture. Observations may take a value of 1, 2, or 3, respectively reflecting a potentially negative, inconclusive, or positive impact on supply. In order to compose indicators from these observations, a weighted arithmetic mean was calculated at farm level and territorial level separately, in which observations were weighted against the single quality score of the article from which they were derived. The integration of the observations with the quality criteria is illustrated in equation 4, and the full process of indicator composition is illustrated visually in Figure 2. In total, observations for 26 farm management practices and 17 ES were extracted during the REA. As such, if linkages between all 26 practices and 17 ES were to be derived during the REA, a total of 442 indicators could be calculated. However, as not all management practices impact the supply of all 17 ES, only 193 indicators were calculated in total. 133 indicators were calculated at farm level and 60 were calculated at territorial level.

Relying on semi-quantitative data derived from secondary literature, we are aware of a need to quantify our confidence in the conclusions put forward by the indicators composed here. Due to the nature of the semi-quantitative data, we are not able to incorporate traditional confidence measures such as confidence intervals. Instead, we rely on the quality as well as the quantity of the evidence to provide us with an indication of confidence. Similar to empirical research, sub-optimal research methodologies in secondary literature lead to reduced confidence in results (ROSES, 2021). As such, we attempt to quantify the quality of, and thus our confidence in, the evidence by incorporating a measure of research quality by following a standardized checklist of quality criteria evaluating research methodologies. Quality scores were calculated for each article included in the REA separately. Therefore quality of evidence is considered at the article level. All observations derived from the same article thus have the same article quality score.

Not only the quality of the literature, but also the quantity of evidence that can be derived reflects the confidence we may have in the results put forward by the secondary literature. Quantity of evidence is an important measure of confidence, as it illustrates the degree to which a certain management practice-ES link has been studied in the literature. Our confidence in conclusions drawn from 100 observations is naturally higher than in those drawn from only 5 observations. We consider quantity of evidence at the observation level (i.e. number of observations) rather than at the article level (i.e. number of articles) because observations were derived from secondary literature. As such, multiple observations from a single article in the REA reflect evidence from across various primary studies in the literature. By incorporating both quality and quantity of evidence into a single value, the correction factor serves to give us an indication of the confidence we may have in the conclusions put forward by the indicators. Equations 1 through 3 illustrate how the quantity and quality of evidence are incorporated into the correction factor.

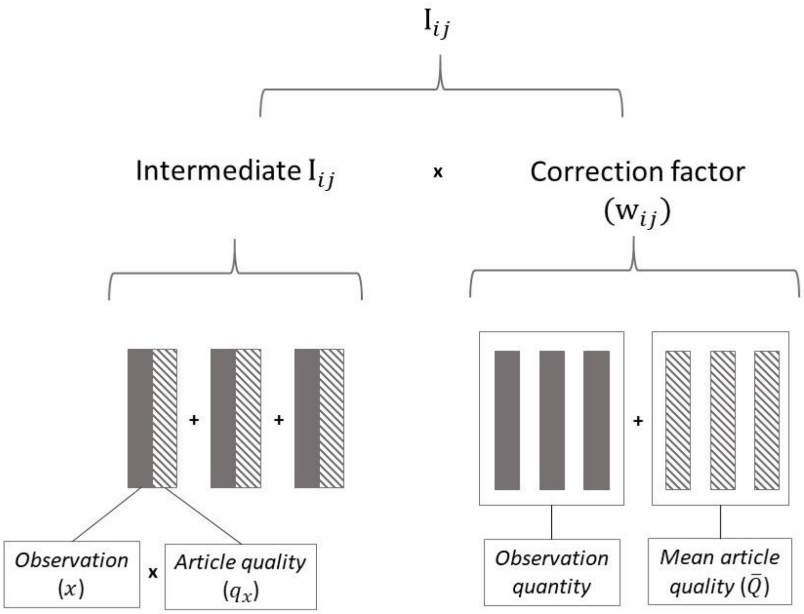


Figure 2. Visual representation of the indicator calculation process. The intermediate I_{ij} (the sum product across multiple observations (x) and their respective article quality score (q_x)) is multiplied by the correction factor (w_{ij}) to obtain I_{ij} for each management practice i linked to ecosystem service j. The correction factor is composed of a measure of the quantity of observations and the average article quality ($\bar{Q}$) across all articles included in I_{ij}.

As described above, the correction factor is composed of a measure of evidence quantity and quality. For each I_{ij}, the mean article quality ($\bar{Q}_{ij}$) across all articles evaluating the impact of management practice i on the supply of ES j was calculated as follows:

$$\bar{Q}_{ij} = \frac{\sum_{n=1}^{no_{ij}} q_n}{no_{ij}} \qquad (1)$$

Where q_n is the article quality associated with observation n, and no_{ij} is the total number of observations evaluating the supply of ES j from management practice i. $\bar{Q}_{ij}$ may take a value between 0 and 1.

Quantity of evidence was incorporated into the correction factor by evaluating no_{ij} per indicator. This was achieved by using the cumulative distribution function (CDF) for farm and territorial level observations separately. The CDF estimates the probability that each I_{ij} is based on exactly no_{ij} number of observations, considering the distribution of the number of observations across all indicators at the considered level. As such, we gain an understanding of how well a given ES-management practice link is studied in the literature, and may thus draw conclusions accordingly. Using the exponential distribution, probabilities were estimates as follows:

$$P\left(no_{ij} < \overline{no}\right) = 1 - e^{-\frac{no_{ij}}{\overline{no}}} \qquad (2)$$

Where $\overline{no}$ is the mean number of observations across all indicators ($\overline{no}_{farm} = 6.06, \overline{no}_{terr} = 2.17$) and no_{ij} is the number of observations linking management practice i to ES j. The CDF was calculated for each management practice i linked to ES j at farm and territorial level separately. The probabilities derived using the CDF were then incorporated with $\bar{Q}_{ij}$ to obtain a single value for the correction factor (w_{ij}) ranging between 0 and 1:

$$w_{ij} = \bar{Q}_{ij} + \frac{P(no_{ij}) - 0.5}{0.5} * r \tag{3}$$

Where w_{ij} is the correction factor for I_{ij} linking management practice i to ES j, calculated based on the mean article quality $\bar{Q}_{ij}$ across I_{ij}, the probability $P(no_{ij})$ associated with the number of observations in I_{ij}, as well as a constant r which reflects the trade-off made between the number of observations (no_{ij}) and the mean article quality ($\bar{Q}_{ij}$). Such a trade-off is considered because the quality and the quantity of evidence are related, but distinct measures influencing the indicator. r may take a value between 0 and 1, where any value closer to 0 results in more importance being placed on the quality of evidence, while any value closer to 1 results in more importance being placed on quantity of evidence.

By setting $r = 0.1$, we assume quality of evidence ($\bar{Q}_{ij}$) is more influential in determining our level of confidence in the indicator compared to the quantity of evidence (no_{ij}). The reasoning behind this is that we do not have the same level of confidence in a high number of low quality observations as we do in a low number of high quality observations. Not all secondary literature is created equal. If special care is not paid to the process of synthesizing evidence from primary literature, there is substantial risk of biased, misinterpreted and/or incorrect conclusions being drawn (Philibert et al., 2012). Therefore, by placing more importance on $\bar{Q}_{ij}$ we are able to correct our observations for these risks.

Finally, the correction factor was incorporated into the calculation of the indicator. The indicator was calculated using a weighted arithmetic mean as described in equation 4.

$$I_{ij} = \left(\left(\frac{\sum_{n=1}^{no_{ij}} x_n q_n}{\sum_{n=1}^{no_{ij}} q_n} \right) w_{ij} \right) - (2 * w_{ij}) \tag{4}$$

Where I_{ij} is the sub-indicator calculated for management practice i linked to ES j. x is the semi-quantitative value of observation n (which takes the value 1, 2 or 3), q_n is the article quality associated with observation n, and w_{ij} is the correction factor specific to the interaction between management practice i and ES j (derived in equation 3). Normalisation of the sub-indicator to a scale of -1 to +1 is achieved by $(2 * w_{ij})$.

This above-described process was carried out for the full set of data derived from the REA, and was repeated for observations at the farm and territorial level.

3.2.1. Supplementary measures

To increase transparency, supplementary measures are reported alongside the indicators. For each I_{ij} linking management practice i to ES j, supplementary measures include the correction factor (w_{ij}), the quantity of observations (no_{ij}), the quantity of articles (na_{ij}), as well as a consensus value (c). While the former two supplementary measure are also incorporated into the indicator calculation (as illustrated in section 3.2.1), the latter two are not. Instead, na_{ij} and c are merely reported alongside the indicators to increase transparency and to facilitate interpretation.

Consensus (c) quantifies the degree to which the various observations included in I_{ij} take the same value. In other words, the consensus value measures the amount of agreement amongst observations in terms of the reported impact of management practice i on the potential supply of ES j. Consensus is highly correlated to variance, but it is more suited to illustrate heterogeneity amongst ordinal observations, as it considers proximities of observations in ordinal scales more accurately (Tastle and Wierman, 2007).

Consensus was calculated according to the approach developed by Tastle and Wierman (2007):

$$c_{ij} = 1 + \sum_{x=1}^{3} p_x \log_2 \left(1 - \frac{|x - \mu_{ij}|}{d_x}\right) \tag{5}$$

Where p_x is the relative frequency of the semi-quantitative observation x (which takes the value of 1, 2 or 3), μ_{ij} is the mean value across all observations of x for management practice i linked to ES j as calculated according to equation 6, and d_x is the width across all observations of x calculated as $d_x = x_{max} - x_{min}$.

$$\mu_{ij} = \sum_{x=1}^{3} p_x x \tag{6}$$

A complete lack of consensus (i.e. observations taking opposing values) would result in a consensus value of $c = 0$. In contrast, if all observations took the same value there would be complete consensus and $c = 1$.

3.2.2. Sensitivity analysis

In this paper we calculate indicators for 26 management practices at farm and territorial level, holding the key assumptions outlined in section 3.2.1. We perform sensitivity analyses in which we relax the above-mentioned assumption and look at how this affects our indicators.

As observations were derived from articles which synthesized results from a variety of primary articles, and because many of the ES against which management practices were evaluated were quite broadly defined during data collection, we allowed for multiple observations to be derived for the same management practice-ES link derived from the same article. This way we were able to ensure that enough variation is captured by the indicator. Though output from the REA was thoroughly cleaned prior to indicator composition, we perform a sensitivity analysis to assess for double counting. To do this, we performed a separate calculation of the indicators, this time allowing for only one observation per management practice-ES link from a single article to be included.

Further, we compare the use of consensus (c) against the use of variance to evaluate the degree to which observations within a single indicator take the same value. We found a strong correlation between the two measures, $r(861) = -0.9506, p < 0.001$. Therefore, considering the consensus is more suited to reflect agreement amongst ordinal data, we opt to maintain the use of this measure in any further reporting.

Finally, and perhaps most importantly, we test the assumption of the increased importance of evidence quality over quantity made in the correction factor (w_{ij}). We do this by calculating indicators for each trade-off factor r ranging from $r = 0.1$ to $r = 0.9$, increasing r by 0.1 with each iteration.

4. Results and discussion

The total number of indicators calculated differed between the considered levels; significantly more indicators were calculated at farm level (133) than at territorial level (60). As a result, we also observe a higher mean number of indicators calculated per management practice at farm level (Table 2). Combined with the higher mean number of observations ($\overline{no}$) and articles ($\overline{na}$) at this level, we may conclude that management practice-ES linkages are more commonly studied at the farm level than at the territorial level.

Further, we observe that the quality of the evidence at farm level is slightly higher than at territorial level, reflected in the higher correction factor, though the difference is slight (Table 2). Unlike previously described trends, the mean consensus at farm level is lower than at territorial level (Table 2). This indicates that observations within indicators at the territorial level more frequently take the same values (positive, negative, or inconclusive). However, this may be attributed to the fewer number of observations per indicator at territorial level overall.

Significant differences were also found between the number of indicators calculated for each of the three CICES ES categories (CICES, 2018); more indicators were calculated for regulating and maintaining ES compared to provisioning and cultural ES at both farm and territorial level ($p < 0.001$). The number of indicators calculated per ES category was further found to differ significantly between farm and territorial level for regulating and maintaining services and for provisioning services, but not for cultural services. At farm level, significantly more indicators were calculated for provisioning services compared to territorial level ($p < 0.001$), while at territorial level significantly more indicators were calculated for regulating and maintaining services compared to farm level ($p < 0.001$). No significant differences were observed for the correction factors of the indicators between the ES classes, nor between the levels.

Table 2. Mean calculations describing the difference in indicator calculation between farm and territorial level.

	Farm level	**Territorial level**
Mean number of indicators per management practice	6	1.85
Mean number of observations ($\overline{no}$) per indicator	6	2.2
Mean number of articles ($\overline{na}$) per indicator	3.8	1.8
Mean correction factor ($\overline{w}$) across indicators	0.31	0.25
Mean consensus ($\bar{c}$) across indicators	0.75	0.91

The full set of indicators calculated for the 26 management practices linked to the 17 ES can be found in the supplementary materials in tabular format and in appendix in graphical format. Indicators and their respective supplementary measures (correction factor w, consensus c, number of observations no and number of articles na) are tabulated for both farm and territorial level.

Sensitivity analyses found no significant difference between the indicators calculated based on a multiple versus single observations. As such, all results are reported for indicators calculated based on the former. Further, we found a significant difference between the indicators composed using r values ranging from $r = 0.1$ to $r = 0.9$ at both farm ($p < 0.001$) and territorial ($p < 0.001$) level. This demonstrates that by changing the assumptions regarding the trade-off between quality and quantity of evidence to increasingly favour quantity, the magnitude of the indicators change. However, a ranking exercise demonstrated that the order of the indicators ranked from highest to lowest magnitude at both levels does not change with increasing r values. Therefore we maintain the assumption made, and favour evidence quality of quantity, positing that when considering secondary literature as a data source, evidence quality more accurately captures confidence than the number of times a given management practice-ES link is reviewed in the literature.

4.1. Interpretation

indicators delineate the impact of a management practice on the supply of an ES. Following the Cascade model framework described in section 2.1, we highlight that the supply quantified by the indicators does not beget delivery of an ES in practice. Delivery of an ES requires a spatial and temporal overlap with demand. The indicators presented here do not measure demand, and due to the nature of aggregation have lost nuance necessary to delineate spatial and temporal factors. Furthermore, indicators cannot be used to predict the increase or decrease in biophysical units of an ES related to a marginal increase in management practice. Rather, as indicators are dimensionless, they should be interpreted according to their directionality and the magnitude. In this sense, indicators illustrate the big picture of how management practices influence the potential supply of ES within the context of European agriculture.

The directionality of an indicator refers to the sign taken on by the indicator value, i.e. whether it is positive or negative. This directionality is determined by the value taken by the included observations. In the case of an indicator composed of only a single observation, if this observation is positive the directionality of the indicator will also be positive. Likewise, if the observation is negative the indicator will have a negative directionality. For example, the indicator composed for the link between cover crops and pollination at farm level is calculated based on a single positive observation, and therefore has a positive directionality ($I = 0.71, w = 0.71, c = 1, no = 1, na = 1$).

When multiple observations are included in a single indicator calculation, the link between observations and indicator directionality becomes more nuanced. The consensus value (c) sheds light on the distribution of values amongst observations. If $c = 1$, all observations take the same value. Therefore the directionality would be determined as above. In this case, the indicator will also take the same value as the correction factor (w). If not all observations take the same value, c will range between $c = 0$ and $c = 0.99$. In this case, the directionality of the indicator is determined in part by the most frequently taken observations, as well as the quality of the articles from which the observations were derived (see equation 4 in section 3.2.2). The following scenarios illustrate how observation value and article quality interact to determine indicator directionality.

If an even number of observations are included in an indicator calculation, half of which are positive and the other half negative, $c = 0$ because the included observations take opposing values. If article quality associated with the observations is the same, observations will carry equal weight in the indicator calculation, and directionality of the indicator will be neutral ($I = 0$). This scenario is only feasible if the indicator is calculated based on an even number of observations. In the indicator calculated for extensive livestock systems linked with biodiversity at territorial level ($I = 0, w = 0.77, c = 0, no = 2, na = 1$) we see that the indicator is calculated based on 2 observations, with $c = 0$; indicating that exactly half of the observations take a positive value while the remaining half take a negative value. If article quality associated with the observations is not equal, the observations derived from the article(s) of higher quality would be weighted more heavily within aggregation (equation 4, section 3.2.2) and would thus result in the indicator directionality shifting towards the value taken by this more heavily weighted observations. As both observations were derived from the same article ($na = 1$) in the indicator above, it is easy to see that the article quality associated with the 2 observations is the same.

In the indicator calculated for crop rotation linked with fresh water quality at farm level ($I = -0.01, w = 0.04, c = 0, no = 2, na = 2$) we once again see $c = 0$; indicating the two observations included in the indicator take opposing values. However, we also see that the indicator directionality is negative ($I = -0.01$). As described above, this is caused by the higher quality of the article from which the negative observation was derived, resulting in this observation being weighted more heavily in the final indicator composition.

The indicator magnitude refers to the size of the indicator value in the positive or negative direction. indicators close to +1 or -1 have a high magnitude, while indicators close to 0 have a low magnitude. In addition to the degree to which observations take the same values, indicator magnitude is also determined by w, which in turn is influenced by no, the quality of the articles from which observations were derived, and the trade-off between the two. We thus see that indicator magnitude is reflective of the current state of the literature, and may be interpreted as an indication of the level of confidence we have in the directionality of the indicator based on the available evidence. We assume that the combination of a high level of consensus ($c \rightarrow 1$) and a large no associated with an indicator implies that the link between a management practice and an ES is strong and easily observed. Under this assumption we may interpret the indicator magnitude also as a measure of the strength of the quantified management practice-ES impact.

When indicator magnitude is close to 0 (e.g. conservation tillage linked to disease and pest control at territorial level $I = 0.01, w = 0.21, c = 0.19, no = 4, na = 4$), we have little confidence in the directionality of the indicator. This because there may be a low level of consensus amongst observations ($c \rightarrow 0$), low quality articles are included in the indicator composition ($w \rightarrow 0$) implying questionable research methodologies/results, and/or few observations ($no < \overline{no}$) are included in the indicator composition implying a lack of evidence in the literature overall.

The degree to which the indicator magnitude deviates from 0 when $c \neq 1$ is jointly dependent on the quality of the articles from which the observations were derived as well as the number of positive, negative and/or inconclusive observations. Similarly to the directionality, if more positive/negative observations are included in the indicator compared to inconclusive observations, and/or if the positive/negative observations were derived from higher quality articles, the indicator magnitude will increase in the positive/negative direction. Alternatively, if more inconclusive observations are include, and/or if these inconclusive observations were derived from higher quality articles, the indicator magnitude will remain low and close to 0.

Reflecting on w and c may aid in the interpretation described above. Take the low magnitude indicator calculated for the use of organic fertilisers linked with decontamination and fixing processes at farm level ($I = -0.09, w = 0.31, c = 0.01, no = 7, na = 5$). We can see that the 7 observations included in this indicator share a low level of consensus ($c = 0.01$). The observations therefore take opposing values, placing the indicator magnitude close to. Directionality is determined here by the higher quality negative observations. Indicator magnitude, on the other hand, is determined by the low w, which in turn is being driven by the low mean article quality. The latter can be determined by evaluating the distance between $no = 7$ and $\overline{no}_{farm} = 6$. Due to the small distance, mean article quality is not greatly impacted by the number of observations in equation 3, and w therefore mainly reflects the mean article quality across observations. We may thus conclude that while the impact of organic fertilisers on the supply of decontamination and fixing processes at the farm level is negative, the lack of consensus and the low quality of evidence in the literature reduces our confidence in this conclusion. Instead, we contend that more high quality research is necessary to determine the exact linkage between the management practice and the ES.

4.2. Farm level performance

Table 3 summarises the five indicators with the highest positive and negative magnitudes calculated at the farm level. The ranking order is determined by the magnitude of the indicators, thus reflecting the strength of the impact as well as the confidence in the respective directionality based on the available evidence. Of all management practice-ES linkages at farm level, agri-environmental schemes was found to have the strongest positive impact with the highest degree of confidence on pollination ($I = 0.85, w = 0.85, c = 1, no = 4, na = 1$). Our confidence is derived from the high article quality ($w = 0.85$), the relatively high no, as well as the complete consensus amongst observations ($c = 1$). The highest negative impact was quantified for the link between the use of chemical pesticide inputs on soil formation and composition ($I = -0.71, w = 0.71, c = 1, no = 1, na = 1$). Here, our confidence is derived from the high article quality only ($w = 0.71$), as the indicator incorporates only one observation. In fact, we see that all except for one of the highest negative magnitude indicators listed in Table 3 are calculated based on a single observation. The exception to this is the indicator quantifying the impact of the use of organic fertiliser inputs on the regulation of fresh water quality ($I = -0.44, w = 0.44, c = 1, no = 4, na = 2$). Here, the complete consensus amongst the four observations implies agreement in the literature, though the low article quality across the included articles decreases our confidence.

Table 3. Five highest positive and negative magnitude indicators linking management practices and ecosystem services (ES) at the farm level. Results are displayed in descending order starting at highest positive indicator. Correction factor (w), consensus (c), no of observations (no) and number of articles (na) are included as supplementary measures to aid interpretation.

Management practice	ES	I	w	c	no	na
Five highest positive indicators						
Agri-environmental schemes	Pollination	0,85	0,85	1	4	1
Cover crops	Pollination	0,71	0,71	1	1	1
Extensive livestock systems	Biodiversity	0,68	0,68	1	1	1
Crop rotation	Carbon sequestration	0,65	0,65	1	8	1
Extensive livestock systems	Carbon sequestration	0,62	0,62	1	4	1
Five highest negative indicators						
Use of chemical pesticide inputs	Soil formation and composition	-0,71	0,71	1	1	1
Extensive livestock systems	Habitat creation/protection	-0,68	0,68	1	1	1
Low agrochemical pesticide input	Production	-0,50	0,50	1	1	1
Use of chemical fertiliser inputs	Regulation of fresh water quality	-0,44	0,44	1	4	2
Alternative weed management	Biodiversity	-0,37	0,37	1	1	1

Noteworthy as well is that extensive livestock systems appears twice amongst the practices with positive indicators for particular ES, and the use of chemical pesticide inputs appears twice amongst the practices with negative indicators. This indicates that of the 26 considered management practices, extensive livestock systems and chemical pesticide inputs seem to impact the potential supply of a wide variety of ES, but in opposite ways. Indeed, extensive livestock systems also appears once in the five highest negative indicators.

Figure 4 illustrates the full set of indicators calculated for extensive livestock systems and chemical pesticide inputs at the farm level. The plots provide a complete overview of the amount of ES evaluated for each management practice, as well as information on how the magnitudes and directionalities of the indicators differ between the ES. We see here that seven indicators were calculated for extensive livestock systems, while three were calculated for chemical pesticide inputs. Further, we see that the magnitudes of the indicators for extensive livestock systems are all relatively high, while for chemical pesticide inputs there is more discrepancy in indicator magnitudes across ES. Combined with the observation that no seems to be lower for negative indicators in Table 3, Figure 4 highlights a discrepancy between the amount of positive and negative indicators calculated at farm level.

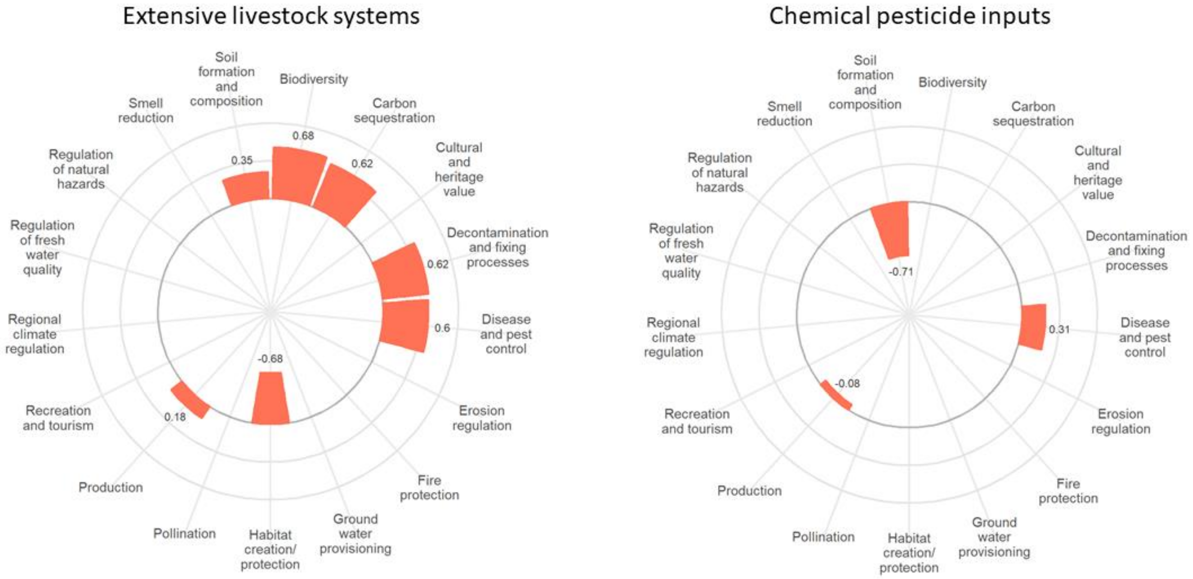


Figure 3. Plots illustrating the full set of indicators calculated for extensive livestock systems and chemical pesticide inputs at the farm level.

When considering the number of observations, we see that the indicators in Table 3 have widely ranging no values. At $no = 8$, the indicator for crop rotation linked to carbon sequestration was calculated based on the highest amount of observations. Six of the ten listen indicators were calculated based on only one observation. Of these six, four had a negative directionality. This may indicate a lack of evidence in the literature looking at the negative impacts of management practices on the supply of ES at farm level. Indeed, when consulting the full set of indicators included in the supplementary materials, we see that of all negative indicators, the highest recorded number of observations is $no = 7$. This is the case for the indicator linking the use of organic fertilisers to decontamination and fixing processes ($I = -0.08, w = 0.31, c = 01, no = 7, na = 5$). At the same time, $no = 68$ was the highest number of observations across all positive indicators at farm level; a great deal more than for the negative indicators. Furthermore, we were only able to calculate 30 negative indicators at farm level, as opposed to 101 positive indicators. While this may potentially be caused by our REA not fully capturing the evidence in the literature, we do not think this is the case. The selection procedure for articles included in the REA was rigorous (section 3.1), and post-data collection analysis of the REA database illustrates that our corpus is indeed representative of the literature.

Rather, we consider the cause of this discrepancy to be two-fold; on the one hand, the management practices that are reviewed here consist mainly of ecological farm management practices. These by definition are designed to reduce externalities and may thus result is less available evidence for their negative impact on the supply of ES. On the other hand, the discrepancy may be indicative of a gap in the literature, with more research focusing on the positive impacts of management practices on ES rather than the negative impacts. We consider both explanations feasible, and conclude that there is likely a combination of both occurring; with positive impacts being more widely reported in the literature (in a trend similar to publication bias), while at the same time having included an unbalanced selection of management practices in the analysis.

4.3. Territorial level performance

Table 4 summarizes the five indicators with the highest positive and negative magnitudes calculated at the territorial level. Here we see that extensive livestock systems has the strongest positive impact on the supply of habitat creation/protection at the territorial level ($I = 0.72, w = 0.72, c = 1, no = 1, na = 1$). The high indicator magnitude implies that our confidence in the positive impact of extensive livestock systems on the supply of habitat creation/protection is high. However, based on only a single observation, we cannot conclude that the magnitude of the indicators completely reflects the strength of the impact.

Cover crops linked to ground water provisioning ($I = 0.59, w = 0.59, c = 1, no = 3, na = 2$) and agri-environmental schemes linked to biodiversity ($I = 0.58, w = 0.78, c = 0.69, no = 4, na = 1$) are the only two indicators in Table 4 with $no > 1$. For the former, the combination of $c = 1$ and $no = 3$ implies a high level of consensus across a high quantity of evidence. However, the quality of the evidence is such that our confidence in the reported positive impact is somewhat diminished, reflected in a lower indicator magnitude. For the latter, the magnitude of the indicator reflects a decrease in confidence in the positive impact of agri-environmental schemes on the supply of biodiversity due to a lack of complete consensus in the literature. Nonetheless, the indicator is still ranked third highest across all positive indicators at territorial level, indicating that the quality of the evidence is high despite the disagreement amongst observations.

Across all indicators calculated at territorial level, agri-environmental schemes has the strongest negative impact on the supply of disease and pest control ($I = -0.18, w = 0.18, c = 1no = 1, na = 1$). Despite its ranking, the magnitude of the indicator remains low; especially when compared to the strongest negative indicators magnitudes at farm level. Indeed we see that indicator magnitude is low across all listed negative indicators in Table 5. Consulting the supplementary measures we see that all negative indicators at territorial level have low w values and are all calculated based on a single observations.

Table 4. Five highest positive and negative magnitude indicators linking management practices and ecosystem services (ES) at the territorial level. Results are displayed in descending order starting at highest positive indicator. Correction factor (w), consensus (c), no of observations (no) and number of articles (na) are included as supplementary measures to aid interpretation.

Management practice	ES	I	w	c	no	na
Five highest positive indicators						
Extensive livestock systems	Habitat creation/protection	0,72	0,72	1	1	1
Cover crops	Ground water provisioning	0,59	0,59	1	3	2
Agri-environmental schemes	Biodiversity	0,58	0,78	0,69	4	1

Semi-natural habitats	Disease and pest control	0,50	0,50	1	1	1
Semi-natural habitats	Erosion regulation	0,50	0,50	1	1	1
Five highest negative indicators						
Agri-environmental schemes	Disease and pest control	-0,18	0,18	1	1	1
Use of organic fertilisers	Regulation of fresh water quality	-0,13	0,13	1	1	1
Use of organic fertilisers	Soil formation and composition	-0,13	0,13	1	1	1
Crop rotation	Carbon sequestration	-0,11	0,11	1	1	1
Sustainable grazing	Erosion regulation	-0,11	0,11	1	1	1

Similarly to at farm level, in Table 4 we see a number of management practices appearing multiple times amongst the strongest positive and negative indicators. At territorial level we see that semi-natural habitats appears twice amongst the positive indicators and that organic fertiliser inputs appears twice amongst the negative indicators. In Figure 5 we observe that a lot of high magnitude, positive indicators have been calculated for semi-natural habitats. Consulting the supplementary measures, we see that for half of the indicators calculated for semi-natural habitats $no > \overline{no}_{terr}$. This indicates that semi-natural habitats is one of the most well studied management practices at the territorial level. While the quality of the evidence for this practice is also high, the high no values is a significant driver of the high indicator magnitudes.

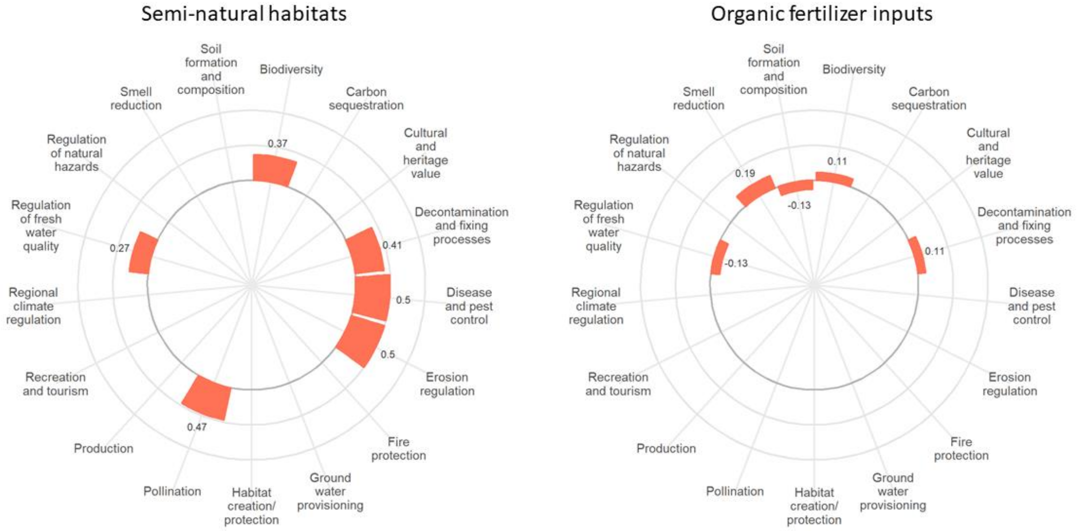


Figure 4. Plots illustrating the full set of indicators calculated for semi-natural habitats and organic fertiliser inputs at the territorial level.

Also in Figure 5 we see that, though an above average amount of indicators have been calculated for organic fertiliser inputs at the territorial level, the magnitudes of these indicators are all relatively low. Once again, consulting the supplementary measures sheds light on what is driving these low magnitude indicators. In this case, we see that both the quantity and quality of the evidence used to calculate the indicators linking organic fertilisers to ES is low. As the distance between no and $\overline{no}_{terr}$ across all indicators calculated for organic fertiliser inputs is small, the quantity of evidence is not sufficient to correct for the low quality of the evidence. This illustrates that though we can find evidence for both the positive and negative impact of organic fertilisers on ES in the literature, this is not of a high enough quality to confidently draw conclusions upon. Our indicators thus demonstrate that there is not only more research needed into the impact of organic fertilisers on ES at the territorial level, care should be taken to ensure that this research is of sufficiently high quality. The consistently high magnitude indicators calculated for semi-natural habitats provide an example of how more high quality research at the territorial level could influence the indicators, and provides insights into the value of such work in gaining a more substantiated understanding of the impact of management practices on ES supply at the territorial level.

4.4. Considerations and drawbacks

Based on results from the present exercise we are able to identify some noteworthy shortcomings in the literature on agricultural practices linked to ES. First, we see that there is a clear gap in the literature evaluating management practice performance at territorial level when compared to farm level. When comparing between the considered levels, quantity of evidence differed significantly between farm and territorial level, with $\overline{no}_{farm} > \overline{no}_{terr}$. Surprisingly though, no significant difference in the correction factor (w) between the two levels was found, though the average at territorial level was slightly lower than at farm level. While we are therefore not able to draw any conclusions on the difference in the quality of evidence between the levels, we are able to conclude that there is more research being done at farm level than at territorial level.

Second, we find a lack of evidence in the literature linking management practices to cultural ES at both levels (Figure 4). We speculate this may be caused by inherent difficulties in linking cultural ES to a single management practice as well as difficulties in quantifying particular cultural ES. These evidence gaps may illustrate a systematic trend in the secondary literature towards easily-synthesized results, i.e. small-scale studies of more readily quantified ES. Simultaneously, this evidence gap may illustrate an underlying bias within the primary literature towards such easily quantified and synthesized ES. This coincides with our findings where we see higher no values incorporated in positive indicators compared to negative indicators, as well as overall more positive indicators being calculated at both farm and territorial level. This may be indicative of an underlying tendency in the literature to report positive impacts of management practices on ES supply in a trend similar to publication bias.

A third shortcoming in the literature was identified through inconsistencies in defining a comparator when evaluating performance of agroecological farm management practices across articles. The PICO described in section 3.1.1 outlines the comparator (conventional management practices) that was adopted in the REA. The comparator was assumed to be embedded within the considered articles, and was therefore not explicitly defined in the exclusion/inclusion criteria. It was noticed, however, that this assumption was not self-evident. During data extraction there were some inconsistencies related to comparators amongst observations. While this was accounted for during data cleaning, this raises a caveat within our indicators and their ability to accurately reflect reality. Furthermore, this also identifies what seem to be systematic inadequacies in secondary literature in appropriately identifying comparators. From the REA, it was noted that comparators are not always clearly defined in the secondary literature, despite forming a major part of discussions and conclusions. While common guidelines for evidence syntheses (e.g. Collaboration for Environmental Evidence, 2013) may help mitigate this shortcoming, many reviews are able to be more flexible in their definition of comparators compared to primary studies, signalling a systematic drawback in secondary literature on environmental performance.

A drawback related to the proposed indicator framework rather than the underlying literature is the inability of the indicator to provide information on why a particular management practice may or may not have a particular impact on ES supply. The observed change in directionality of the indicator linking extensive livestock systems to habitat creation/protection between the farm and territorial level is a case in point. Understanding why this shift is observed requires consulting the article from which observations were derived, and cannot be readily interpreted from the indicators nor from the supplementary measures. This reflects an inherent drawback of performance indicators, in which information is lost during data synthesis and aggregation (Girardin et al., 1999). Indeed this is often a key critique regarding sustainability indicators more broadly (Bockstaller et al., 2008; Petkovová et al., 2020). However, we argue that in order to obtain a meaningful overview of the environmental performance of a large amount of management practices to potentially inform policy-making decisions, this is a necessary trade-off. Indeed, as was pointed out by Girardin et al. (1999), indicators are partial; inherently they will not be able to provide an exhaustive vision of reality, or complete information about the system. By being transparent in our methodology and providing access to all relevant supplementary materials we hope to minimize the impact of this trade-off significantly.

5. Conclusion and next steps

The work presented in this paper derives performance indicators quantifying the impact of farm management practices on potential ES supply within the context of European agricultural systems. Here, indicators shed light on both the potential of a management practice to supply a particular ES, as well as on the state of the literature evaluating said practice-ES in terms of quality and quantity of evidence. The indicators we propose give a first indication of causality, though they should not be used to estimate marginal effects of management practice implementation on potential ES supply. Rather, indicators should be used to quantify these impacts in an intuitive manner and to shed light on the current state of knowledge. We present the proposed approach as a framework which may be easily adopted in various contexts, potentially expanding on the types of management practices, ES, as well as the farming systems and/or geographic contexts considered.

We hope to inform policy makers by demonstrating which management practices, based on the available evidence, are most interesting to focus attentions on considering their potential impact(s) on the supply of the considered ES. We find that of the 26 considered ES, extensive livestock systems have a tendency to have the highest consistently positive impact on potential ES supply across various ES at farm level, ranking third highest in indicator magnitude after agri-environmental schemes and cover crops. At territorial level we see a shift towards semi-natural habitats having the highest consistently positive impact on the potential supply of ES, also ranking third highest in indicator magnitude after extensive livestock systems and cover crops. Simultaneously we hope to inform the scientific community by quantifying the current scientific landscape and illustrating where research gaps remain and where more work is needed. Here, we find that the literature could benefit greatly from an increase in high quality research at territorial level as well as for cultural ES.

The indicators presented in the paper fall within the ecological dimension of the ES cascade model presented in Figure 1 by quantifying the supply of ES from management practices. However, as was highlighted previously, indicators quantify only the potential supply of an ES. In order for an ES to be delivered, a consideration of the demand for ES in the social dimension is required. In succeeding work we propose to incorporate this demand-side dimension in the ES cascade model by consulting with local stakeholders across three case study areas in Belgium and England. Using insights gained from this consultation we will calculate composite indicators with the indicators presented here, which consider the realised supply (i.e. delivery) of an ES to obtain an indicator of overall environmental performance of the 26 considered management practices. In doing so we illustrate how the proposed indicators may be applied to evaluate overall environmental performance of management practices considering nuanced spatiotemporal characteristics.

References

Altieri, M.A., 2000. Agroecology: principles and strategies for designing sustainable farming systems. Univ. California, Berkeley.< http//www. cnr. … 1–5. https://doi.org/10.1017/S0889189300008559

Andersson, E., McPhearson, T., Kremer, P., Gomez-Baggethun, E., Haase, D., Tuvendal, M., Wurster, D., 2015. Scale and context dependence of ecosystem service providing units. Ecosyst. Serv. 12, 157–164. https://doi.org/10.1016/j.ecoser.2014.08.001

Bateman, I., Brouwer, R., Cranford, M., 2009. Valuing Environmental Impacts: Practical Guidelines for the Use of Value Transfer in Policy and Project Appraisal. Value Transf. Guidel. … 44, 0–23.

Beillouin, D., Ben-ari, T., Makoswki, D., 2018. Assessing the quality and results of meta-analyses on crop diversification Protocol for systematic review and evidence map.

Beillouin, D., Ben-Ari, T., Makowski, D., 2019. A dataset of meta-analyses on crop diversification at the global scale. Data Br. 24. https://doi.org/10.1016/j.dib.2019.103898

Bockstaller, C., Guichard, L., Makowski, D., Aveline, A., Girardin, P., Plantureux, S., 2008. Agri-environmental indicators to assess cropping and farming systems. A review. Agron. Sustain. Dev. https://doi.org/10.1051/agro:2007052

Bourguet, D., Guillemaud, T., 2016. Sustainable Agriculture Reviews. Sustain. Agric. Rev. 19, 35–120. https://doi.org/10.1007/978-94-007-5449-2

Burkhard, B., Kroll, F., Nedkov, S., Müller, F., 2012. Mapping ecosystem service supply, demand and budgets. Ecol. Indic. 21, 17–29. https://doi.org/10.1016/j.ecolind.2011.06.019

Burkhard, B., Maes, J., 2017. Mapping Ecosystem Services project Mapping and assessment of ecosystem services to improve resource management and human wellbeing in data-scarce peri-urban ecosystems View project.

CICES, 2018. Revision Highlights [WWW Document]. URL https://cices.eu/revision-highlights/ (accessed 12.23.20).

Collaboration for Environmental Evidence, 2013. Guidelines for Systematic Review and Evidence Synthesis in Environmental Management, Environmental Evidence.

Gan, X., Fernandez, I.C., Guo, J., Wilson, M., Zhao, Y., Zhou, B., Wu, J., 2017. When to use what: Methods for weighting and aggregating sustainability indicators. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2017.05.068

Girardin, P., Bockstaller, C., Werf, H. Van der, 1999. Indicators: Tools to evaluate the environmental impacts of farming systems. J. Sustain. Agric. 13, 5–21. https://doi.org/10.1300/J064v13n04_03

Gómez-Limón, J.A., Sanchez-Fernandez, G., 2010. Empirical evaluation of agricultural sustainability using composite indicators. Ecol. Econ. 69, 1062–1075. https://doi.org/10.1016/j.ecolecon.2009.11.027

Haines-Young, R., Potschin, M., 2018. CICES V5. 1. Guidance on the Application of the Revised Structure. Fabis Consult. 53.

Jeanneaux, P., 2018. PerfCuma : A framework to manage the sustainable development of small cooperatives. Int. J. Agric. Manag. 7, 54–65. https://doi.org/10.5836/ijam/2018-07-68

Kragt, M.E., Robertson, M.J., 2014. Quantifying ecosystem services trade-offs from agricultural practices. Ecol. Econ. 102, 147–157. https://doi.org/10.1016/j.ecolecon.2014.04.001

Lucas, V., Gasselin, P., Douwe, J., Der Ploeg, V., 2018. Agroecology and Sustainable Food Systems Local inter-farm cooperation: A hidden potential for the agroecological transition in northern agricultures 43, 145–179. https://doi.org/10.1080/21683565.2018.1509168

Martín-López, B., Gómez-Baggethun, E., García-Llorente, M., Montes, C., 2014. Trade-offs across value-domains in ecosystem services assessment. Ecol. Indic. 37, 220–228. https://doi.org/10.1016/j.ecolind.2013.03.003

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E., 2005. Handbook on constructing composite indicators, OECD Statistics Working Papers.

Nedkov, S., Burkhard, B., 2011. Flood regulating ecosystem services-Mapping supply and demand, in the Etropole municipality, Bulgaria. Ecol. Indic. 21, 67–79. https://doi.org/10.1016/j.ecolind.2011.06.022

Oecd, JRC, 2008. Handbook on Constructing Composite Indicators: Methodology and User Guide, Handbook on Constructing Composite Indicators: Methodology and User Guide. https://doi.org/10.1787/9789264043466-en

Palomo-Campesino, S., González, J.A., García-Llorente, M., 2018. Exploring the connections between agroecological practices and ecosystem services: A systematic literature review. Sustain. https://doi.org/10.3390/su10124339

Petkovová, L., Hartman, D., Pavelka, T., 2020. Problems of aggregation of sustainable development indicators at the regional level. Sustain. 12, 7156. https://doi.org/10.3390/su12177156

Philibert, A., Loyce, C., Makowski, D., 2012. Assessment of the quality of meta-analysis in agronomy. Agric. Ecosyst. Environ. 148, 72–82. https://doi.org/10.1016/j.agee.2011.12.003

Potschin, M.B., Haines-Young, R.H., 2011. Ecosystem services: Exploring a geographical perspective. Prog. Phys. Geogr. https://doi.org/10.1177/0309133311423172

Pretty, J., 2008a. Agricultural sustainability: Concepts, principles and evidence. Philos. Trans. R. Soc. B Biol. Sci. 363, 447–465. https://doi.org/10.1098/rstb.2007.2163

Pretty, J., 2008b. Agricultural sustainability: Concepts, principles and evidence. Philos. Trans. R. Soc. B Biol. Sci. https://doi.org/10.1098/rstb.2007.2163

Pretty, J., Bharucha, Z.P., 2014. Sustainable intensification in agricultural systems. Ann. Bot. https://doi.org/10.1093/aob/mcu205

PRISMA, 2015. PRISMA: TRANSPARENT REPORTING of SYSTEMATIC REVIEWS and META-ANALYSES [WWW Document]. URL http://www.prisma-statement.org/ (accessed 3.3.21).

Rega, C., Paracchini, M.L., Mccraken, D., Saba, A., Zavalloni, M., Raggi, M., Viaggi, D., Britz, W., Frappier, L., 2018. LIFT-Deliverable D1.1 Review of the definitions of the existing ecological approaches.

Rigby, D., Woodhouse, P., Young, T., Burton, M., 2001. Constructing a farm level indicator of sustainable agricultural practice. Ecol. Econ. 39, 463–478. https://doi.org/10.1016/S0921-8009(01)00245-2

ROSES, 2021. About ROSES [WWW Document]. URL https://www.roses-reporting.com/about-roses (accessed 3.3.21).

Sandhu, H.S., Wratten, S.D., Cullen, R., 2010. Organic agriculture and ecosystem services. Environ. Sci. Policy. https://doi.org/10.1016/j.envsci.2009.11.002

Swinton, S.M., Lupi, F., Robertson, G.P., Hamilton, S.K., 2007. Ecosystem services and agriculture: Cultivating agricultural ecosystems for diverse benefits. Ecol. Econ. 64, 245–252. https://doi.org/10.1016/j.ecolecon.2007.09.020

Swinton, S.M., Lupi, F., Robertson, G.P., Landis, D.A., 2006. Ecosystem services from agriculture: Looking beyond the usual suspects, in: American Journal of Agricultural Economics. Narnia, pp. 1160–1166. https://doi.org/10.1111/j.1467-8276.2006.00927.x

Tastle, W.J., Wierman, M.J., 2007. Consensus and dissention: A measure of ordinal dispersion. Int. J. Approx. Reason. 45, 531–545. https://doi.org/10.1016/j.ijar.2006.06.024

Toivonen, M., Huusela-Veistola, E., Herzon, I., 2018. Perennial fallow strips support biological pest control in spring cereal in Northern Europe. Biol. Control 121, 109–118. https://doi.org/10.1016/j.biocontrol.2018.02.015

Turner, R.K., Daily, G.C., 2008. The ecosystem services framework and natural capital conservation, in: Environmental and Resource Economics. Springer, pp. 25–35. https://doi.org/10.1007/s10640-007-9176-6

United Nations Development Programme, 2013. Human Development Index (HDI) [WWW Document]. https://doi.org/10.1016/j.ecolind.2012.12.025

Van den Putte, A., Govers, G., Diels, J., Gillijns, K., Demuzere, M., 2010. Assessing the effect of soil tillage on crop growth: A meta-regression analysis on European crop yields under conservation agriculture. Eur. J. Agron. 33, 231–241. https://doi.org/10.1016/j.eja.2010.05.008

Van Der Werf, H.M.G., Petit, J., 2002. Evaluation of the environmental impact of agriculture at the farm level: A comparison and analysis of 12 indicator-based methods. Agric. Ecosyst. Environ. 93, 131–145. https://doi.org/10.1016/S0167-8809(01)00354-1

van Oudenhoven, A.P.E., Schröter, M., Drakou, E.G., Geijzendorffer, I.R., Jacobs, S., van Bodegom, P.M., Chazee, L., Czúcz, B., Grunewald, K., Lillebø, A.I., Mononen, L., Nogueira, A.J.A., Pacheco-Romero, M., Perennou, C., Remme, R.P., Rova, S., Syrbe, R.U., Tratalos, J.A., Vallejos, M., Albert, C., 2018. Key criteria for developing ecosystem service indicators to inform decision making. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2018.06.020

Van Vooren, L., Reubens, B., Broekx, S., De Frenne, P., Nelissen, V., Pardon, P., Verheyen, K., 2017. Ecosystem service delivery of agri-environment measures: A synthesis for hedgerows and grass strips on arable land. Agric. Ecosyst. Environ. 244, 32–51. https://doi.org/10.1016/j.agee.2017.04.015

Wezel, A., Casagrande, M., Celette, F., Vian, J.F., Ferrer, A., Peigné, J., 2014. Agroecological practices for sustainable agriculture. A review. Agron. Sustain. Dev. https://doi.org/10.1007/s13593-013-0180-7

Wezel, A., Nicholls, C.I., Altieri, M.A., Vazquez, L., 2017. Agroecological Principles for the Conversion of Farming Systems, in: Agroecological Practices for Sustainable Agriculture. WORLD SCIENTIFIC (EUROPE), pp. 1–18. https://doi.org/10.1142/9781786343062_0001

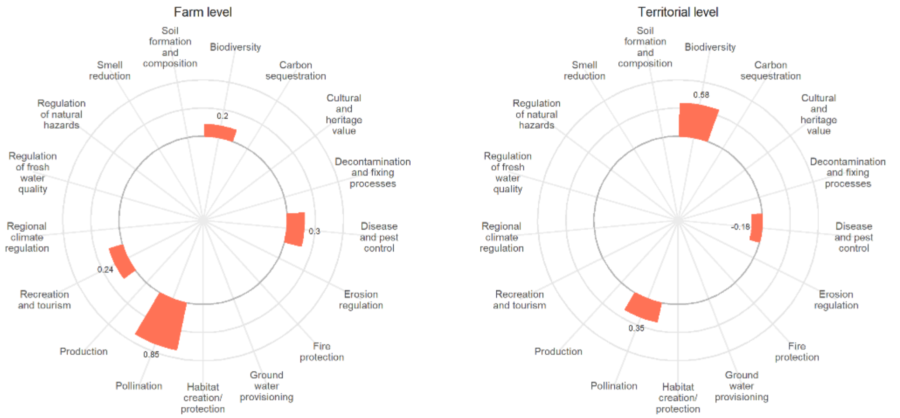
Wezel, A., Soboksa, G., McClelland, S., Delespesse, F., Boissau, A., 2015. The blurred boundaries of ecological, sustainable, and agroecological intensification: a review. Agron. Sustain. Dev. 35, 1283–1295. https://doi.org/10.1007/s13593-015-0333-y

Zhang, W., Ricketts, T.H., Kremen, C., Carney, K., Swinton, S.M., 2007. Ecosystem services and dis-services to agriculture. Ecol. Econ. 64, 253–260. https://doi.org/10.1016/j.ecolecon.2007.02.024

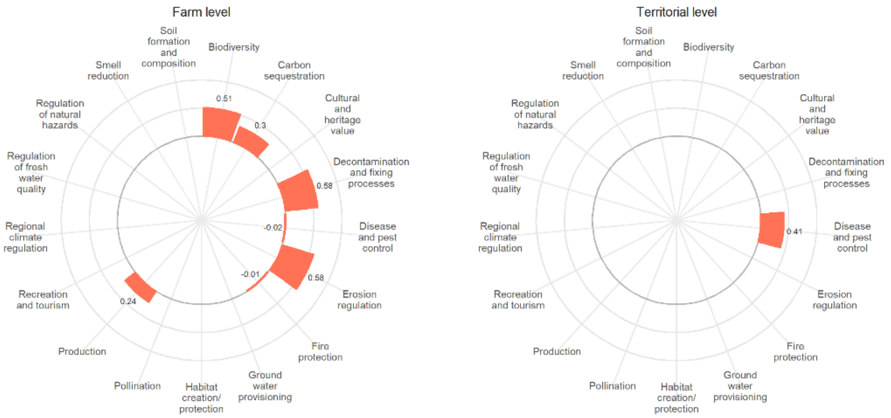
Zhou, Z., Robinson, G.M., Song, B., 2019. Experimental research on trade-offs in ecosystem services: The agro-ecosystem functional spectrum. Ecol. Indic. 106, 105536. https://doi.org/10.1016/j.ecolind.2019.105536

Appendix A: indicators at farm and territorial level for full set of farm management practices (spider diagrams)

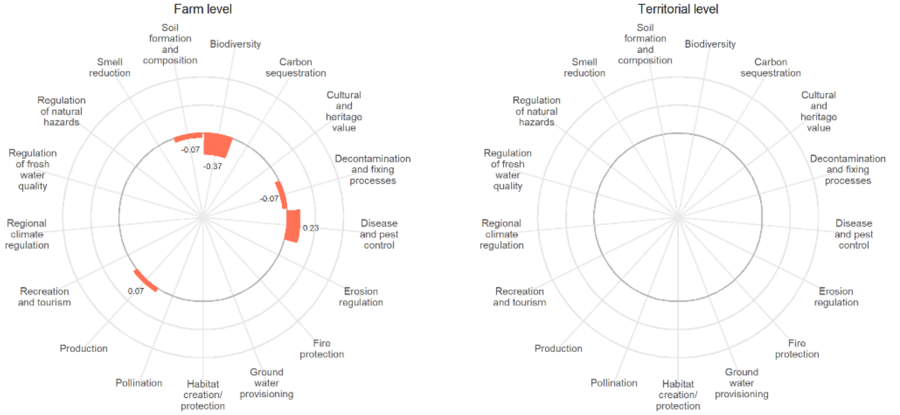
Agri-environmental schemes



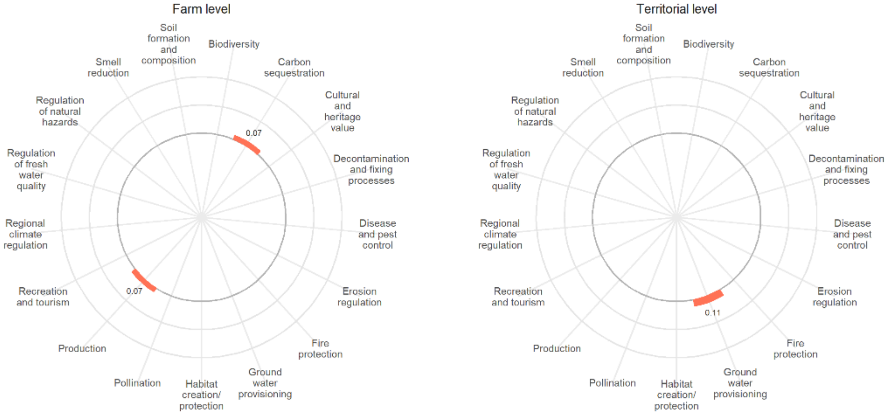
Agroforestry



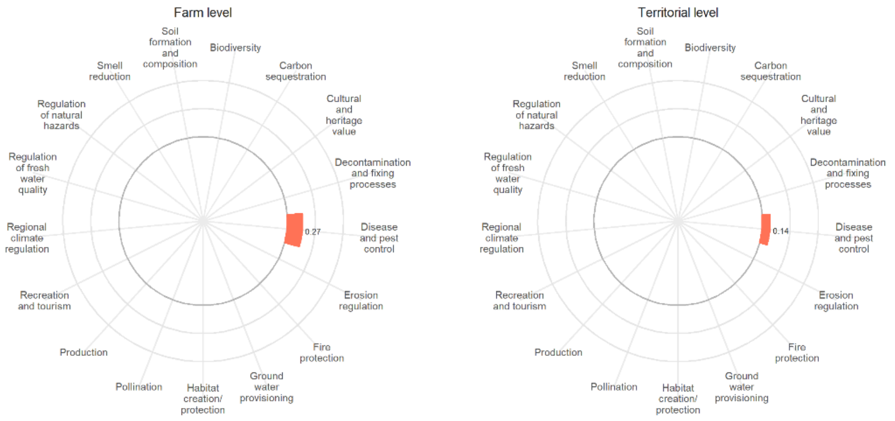
Alternative weed management



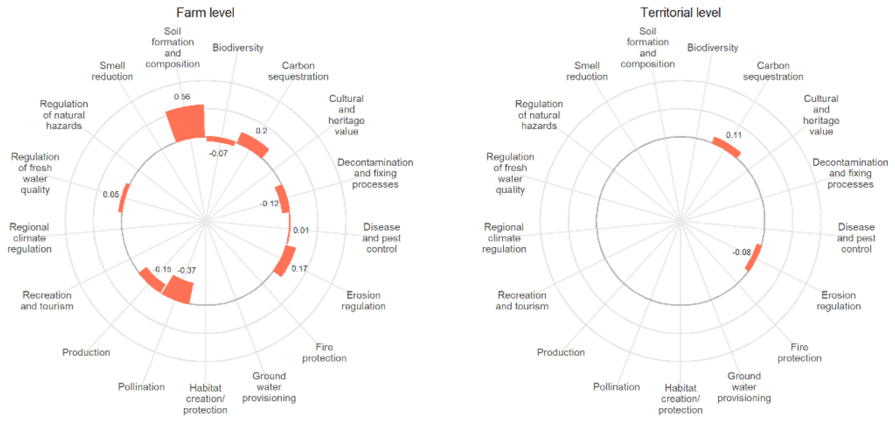
Biological N fixation



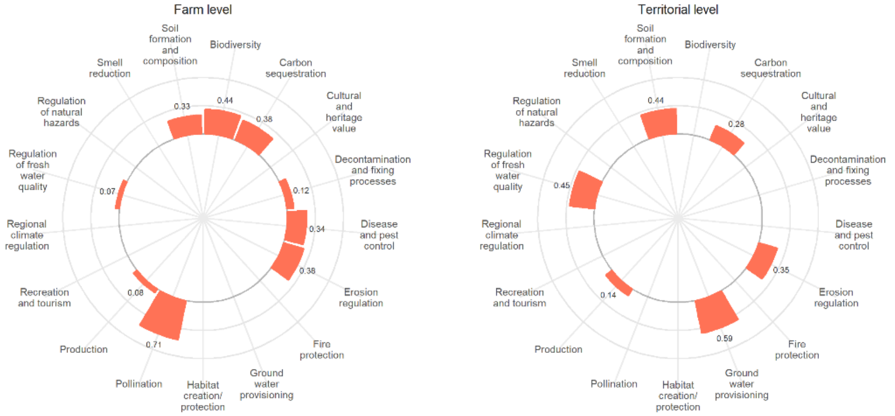
Biological pest control



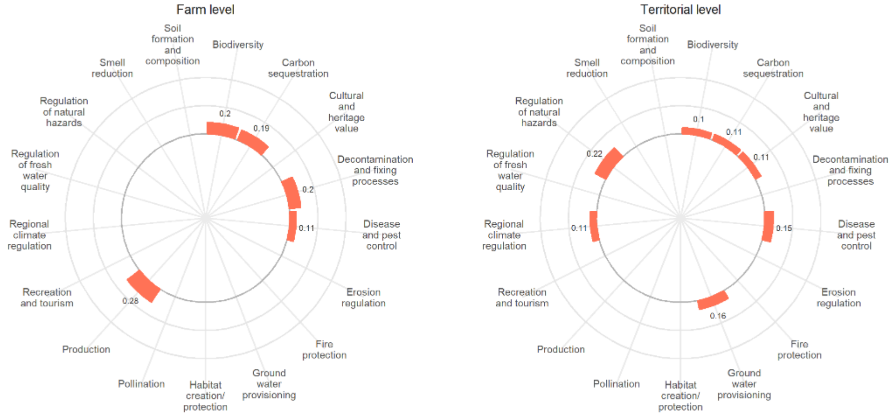
Conservation tillage



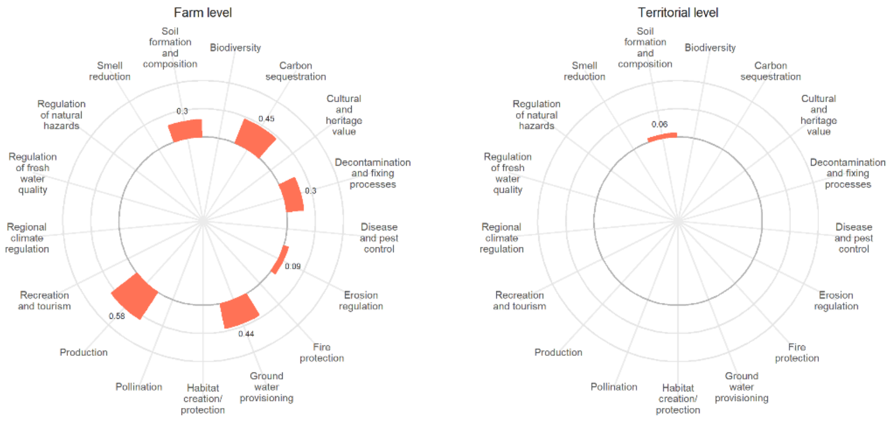
Cover crops



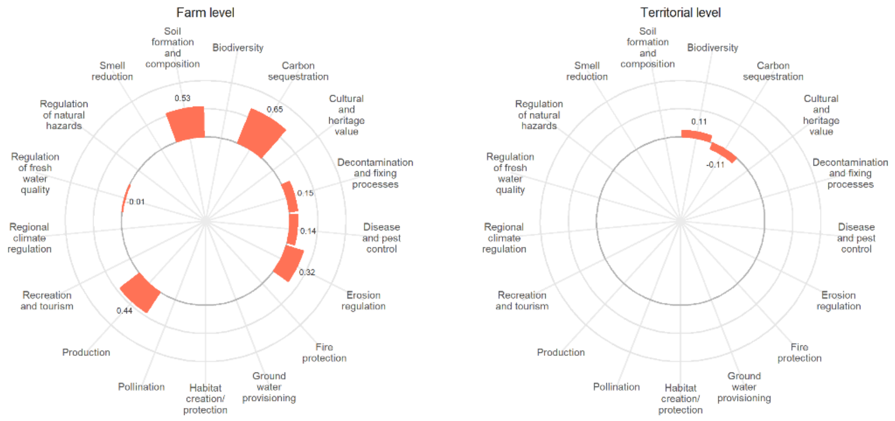
Crop livestock integration



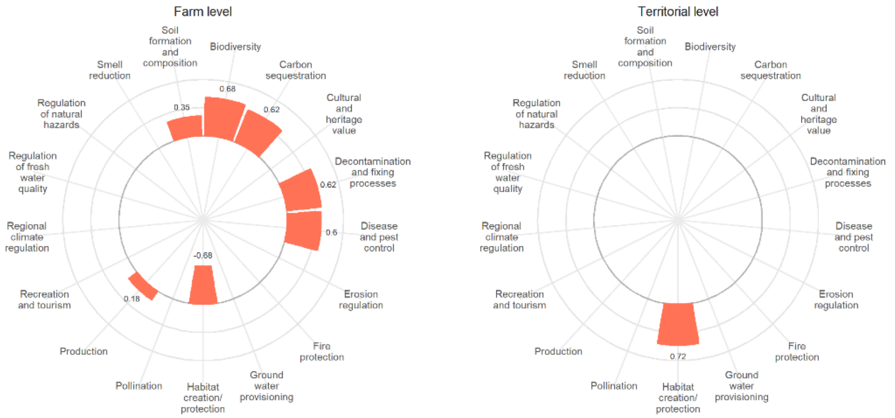
Crop residue management



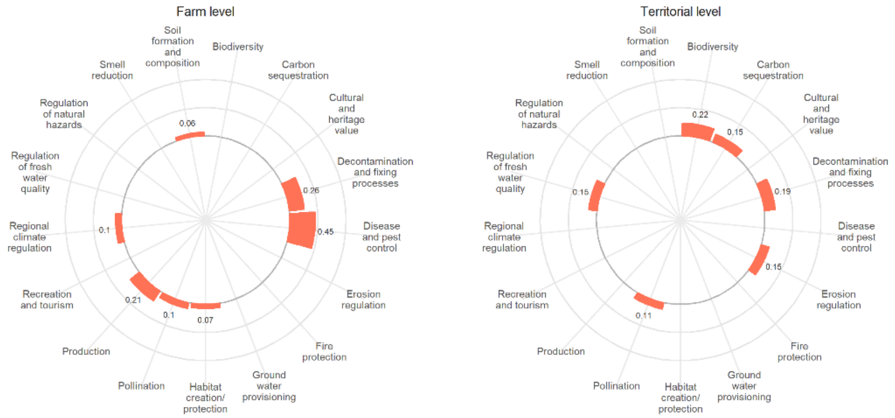
Crop rotation



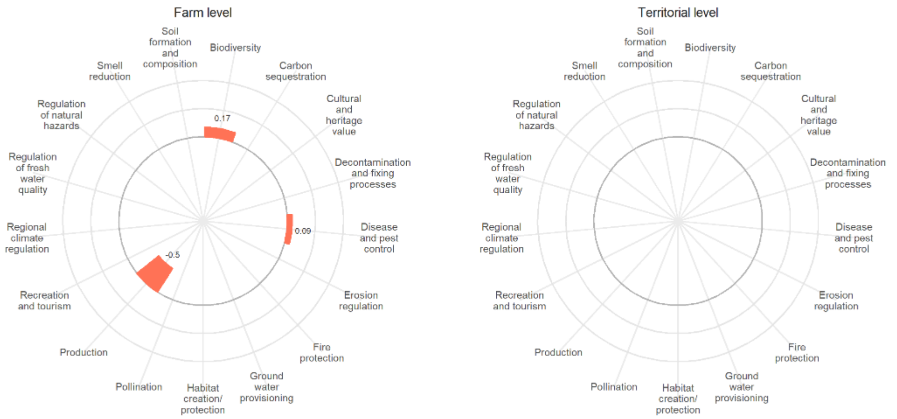
Extensive livestock systems



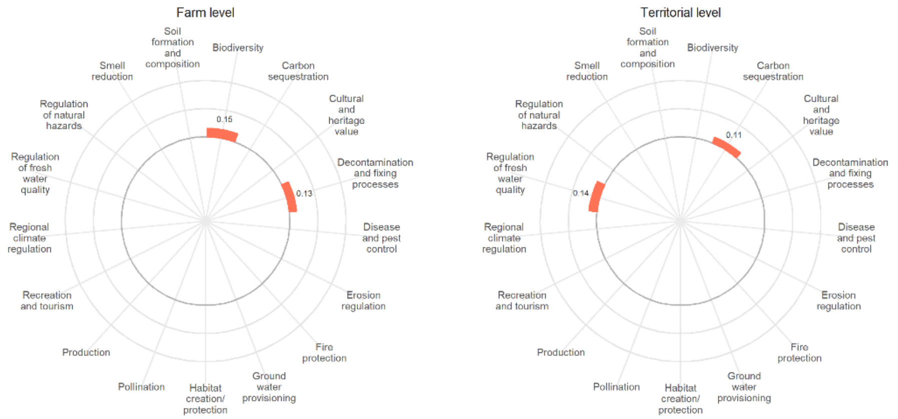
Intercropping



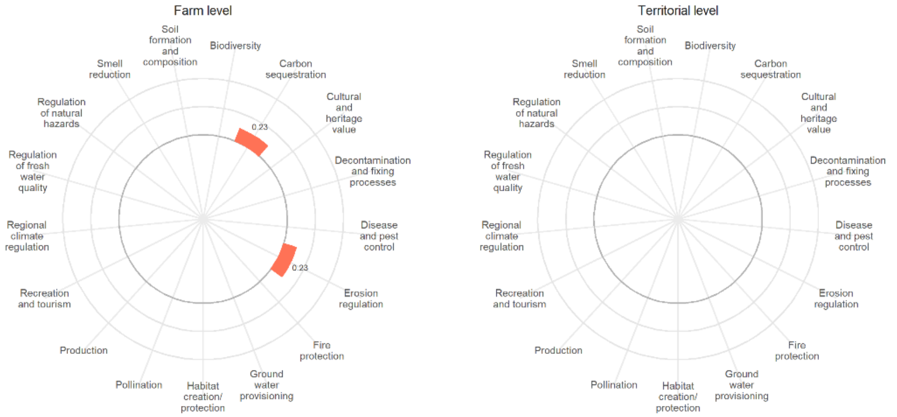
Low agrochemical pesticide input



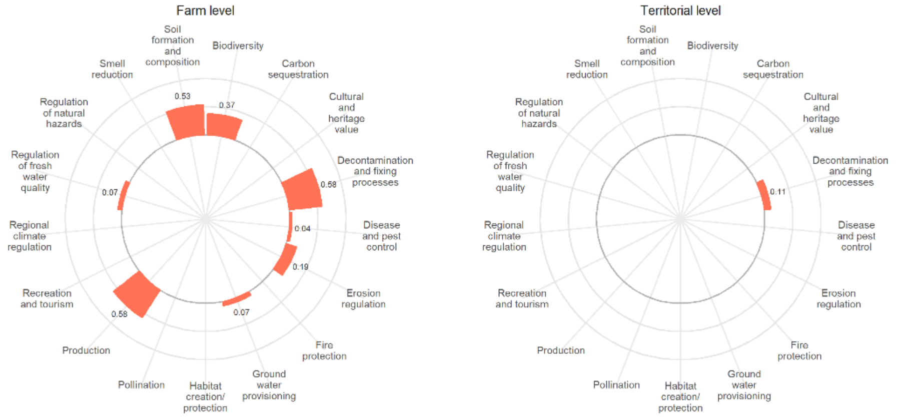
Low fertiliser input

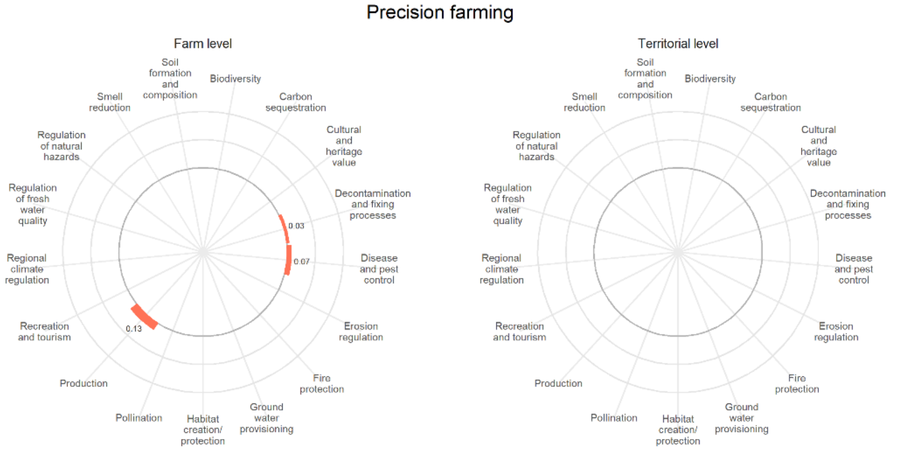


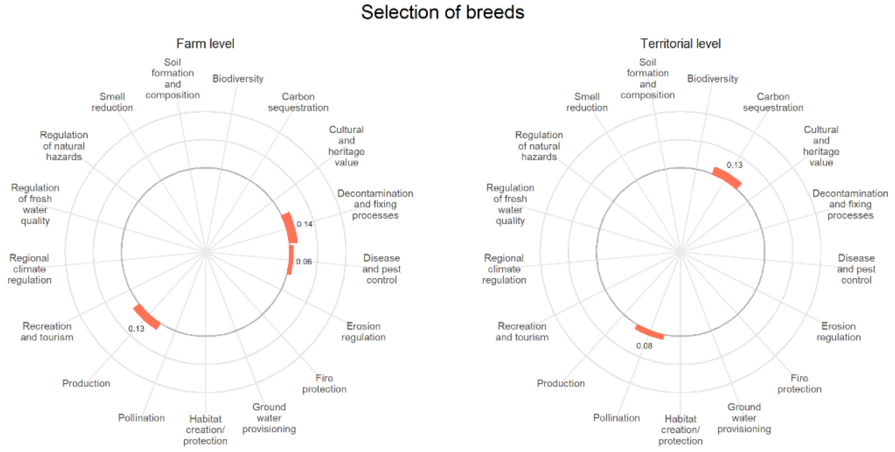
Low mechanisation

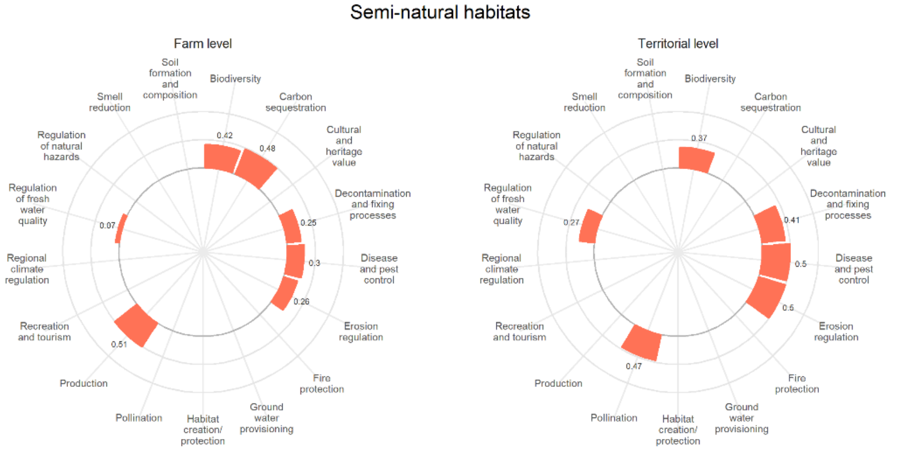


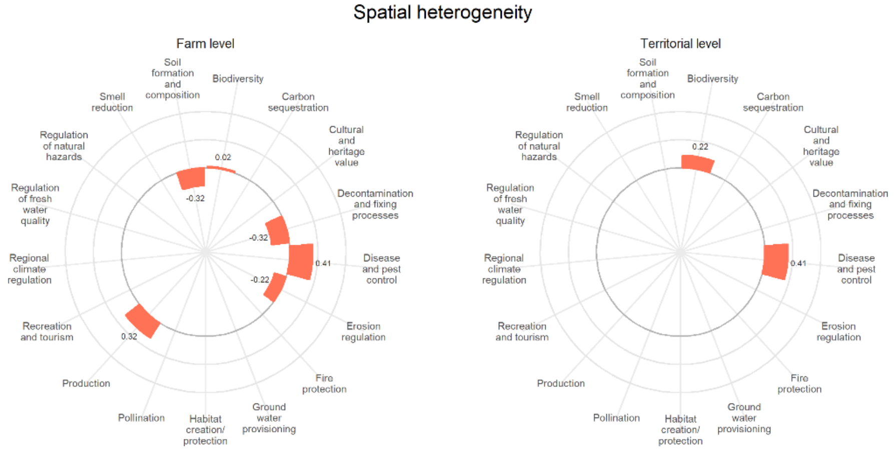
Mulching



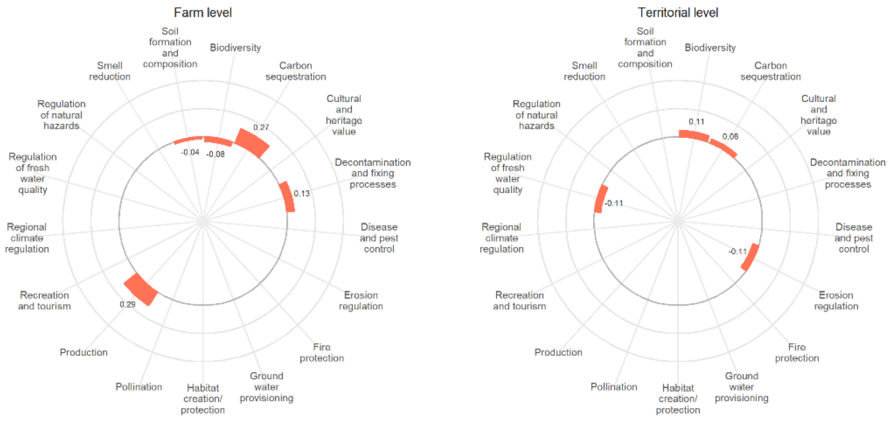

Precision farming
Farm level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
0.03
0.07
0.13
Territorial level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction


Selection of breeds
Farm level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
0.14
0.06
0.13
Territorial level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
0.13
0.08

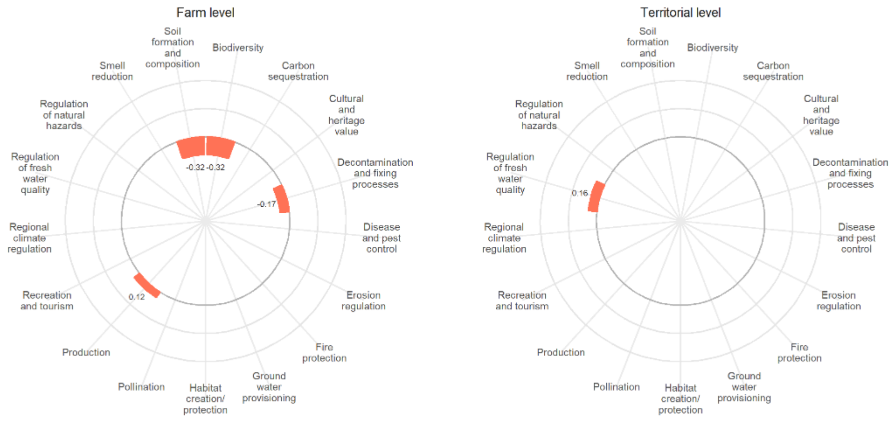

Semi-natural habitats
Farm level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
0.42
0.48
0.25
0.3
0.26
0.51
0.07
Territorial level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
0.37
0.41
0.5
0.5
0.47
0.27


Spatial heterogeneity
Farm level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
0.02
-0.32
-0.32
0.41
-0.22
0.32
Territorial level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
0.22
0.41

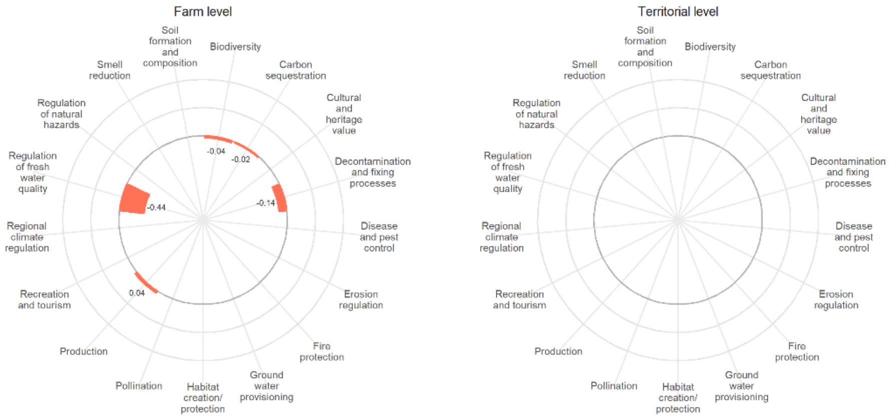
Sustainable grazing



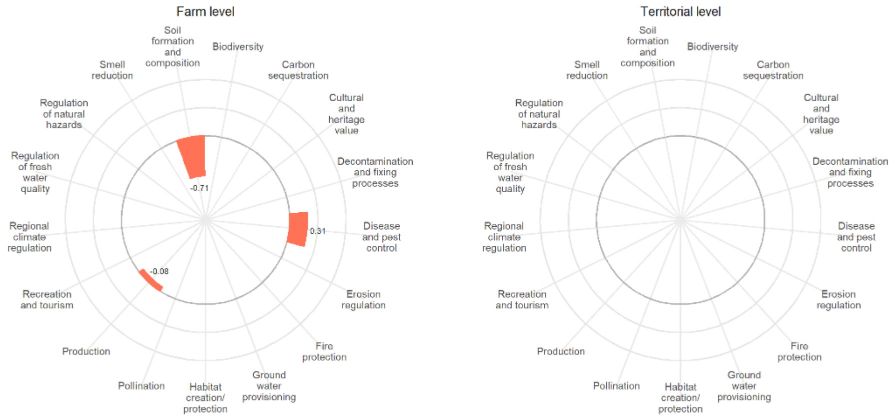
Sustainable water management



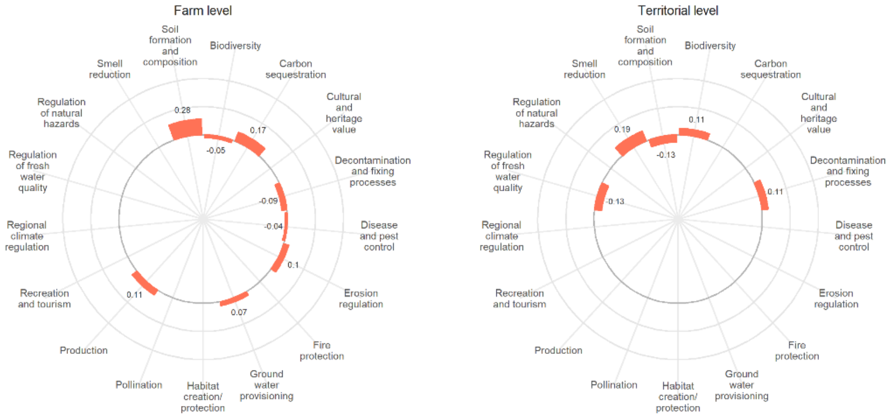
Use of chemical fertiliser inputs



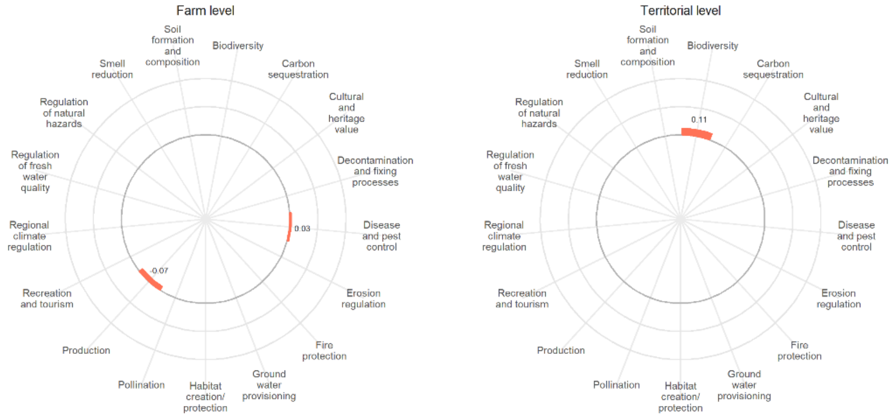
Use of chemical pesticide inputs



Use of organic fertilisers



Use of organic pesticides


Farm level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction
Territorial level
Soil formation and composition
Biodiversity
Carbon sequestration
Cultural and heritage value
Decontamination and fixing processes
Disease and pest control
Erosion regulation
Fire protection
Ground water provisioning
Habitat creation/ protection
Pollination
Production
Recreation and tourism
Regional climate regulation
Regulation of fresh water quality
Regulation of natural hazards
Smell reduction