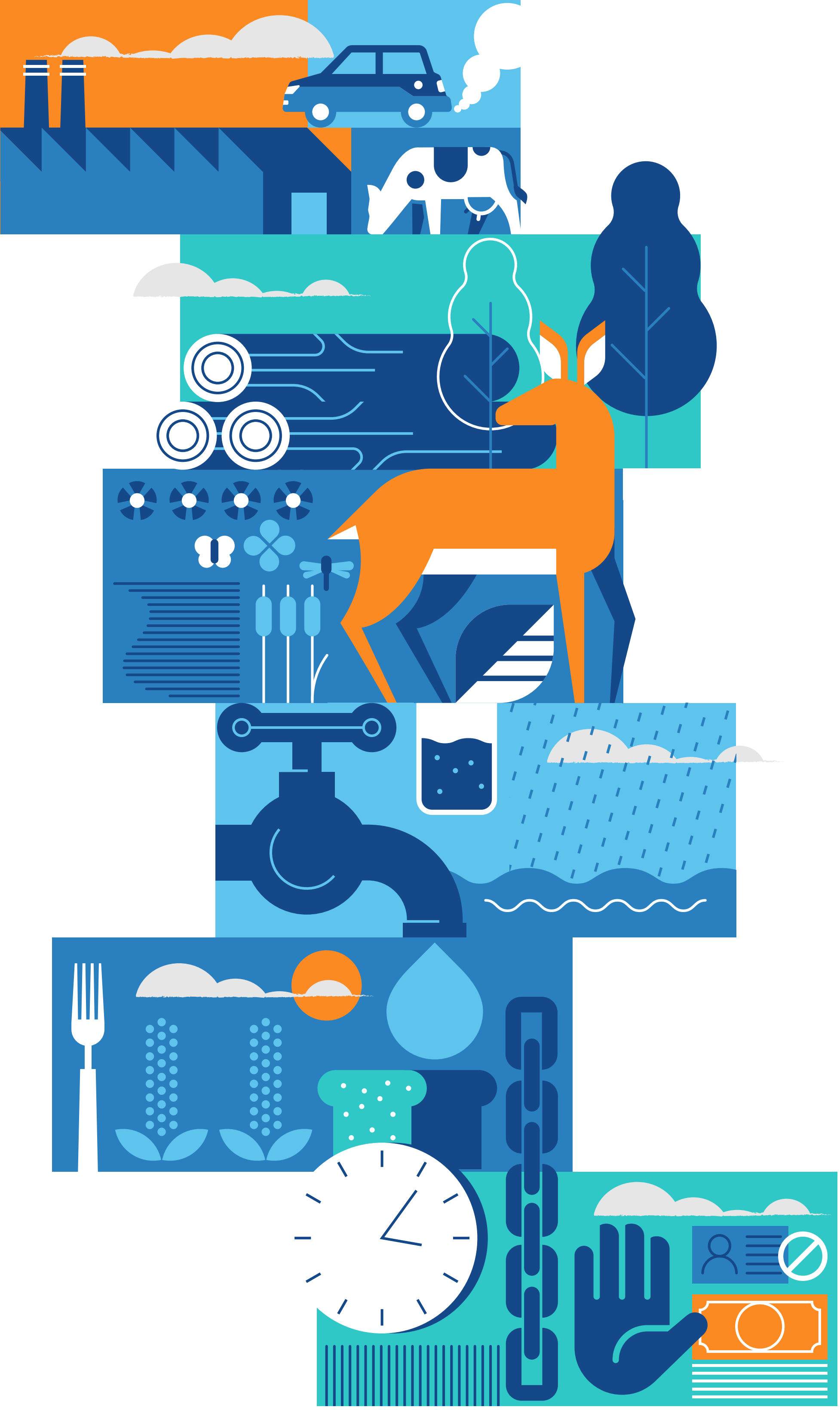


August 2020

Performance metrics for key sustainability issues

Authors:
Dr Steve Jennings, Caitlin McCormack & Richard Sheane

Executive Summary

Voluntary Sustainability Standards (VSS) have grown rapidly in number and importance in global commodity markets in recent years. The standards community has also made progress in researching its own impacts.

However, the metrics generated and reported by standards systems have traditionally been designed to demonstrate the reach of the standard (e.g., number of clients, the total land area certified). As stakeholders evolve their sustainability commitments, they also seek to ensure that standards are providing the benefits they require in a cost-effective way. This increases the demand for good outcome data.

This work is intended to support standard systems and other sustainability initiatives to move towards more data-driven outcome claims, whilst providing insight into the process of considering and understanding the suitability of certain metrics.

The report summarises an assessment of a range of leading metrics that can be used to credibly measure and report on performance over time and across multiple spatial scales. The research is focused on six critical sustainability issues: deforestation, biodiversity, water use, forced labour, poverty, and Greenhouse Gas emissions. The intention is to focus on the applicability of various metrics and data sources, best practices, and limitations and trade-offs associated with their use, rather than to build a definitive inventory of metrics, or to recommend a 'best' metric for each issue. In total, we considered eighty-seven metrics.

Each metric was assessed against a total of eighteen criteria, bundled into three key areas: considerations around defining the sustainability issue; considerations around performance monitoring; and considerations around data use, collection, sources and quality. The full assessment, including explanatory notes and links to further sources of information is presented in an Excel database separately to this document. In this report, we highlight the key characteristics of the metrics against the criteria, and the key choices and trade-offs that an organisation has to make when deciding which metrics to use.

The overarching findings – focusing on the key decisions that an organisation has to make when choosing which metrics to use – include:

- **Aligning the definition of the sustainability issue with your organisational needs.** Only two of the sustainability issues included within this study have relatively uncontested definitions: GHG emissions and forced labour. Within biodiversity, water use, poverty and deforestation, multiple working definitions and theoretical constructs exist, with each having spawned a range of supporting metrics. We summarise the main differences and the practical consequences of the competing definitions and suggest that VSS consider which best aligns to their organisational goals, ethos and to the expectations of their stakeholders, either individually or collectively under the ISEAL umbrella.
- **Trade-offs.** The most common trade-offs are between the cost of data collection, analysis, management and communication and a range of other attributes. The metrics assessed show one common characteristic that is not confined to sustainability: high quality information that reflects the outcomes of management practices and which is responsive to context generally requires significant effort to collect, analyse, manage and report. Metrics that report outputs and inputs are generally cheaper to collect but typically do not provide a good indication of change. Metrics are that are based on secondary data and modelling, and hence often require less investment to use, tend (with some exceptions) to be insufficiently granular to reflect the work of VSS clients, which by definition is often outlying from

the modelled normative practice. Similarly, relatively few metrics were found to have an external baseline against which VSS can demonstrate progress.

One way of managing these trade-offs is to employ a hierarchical approach to metrics, using risk or output measurements at a broad scale to prioritise where digging deeper into outcomes is necessary. There also would seem to be opportunities for shared effort to begin to address some of these issues. For example, agreeing a set of outcome measures for impact evaluation might allow consistent reporting, greater shared learning, and in some cases cost saving.

- **Scaling up.** The emerging emphasis on landscape-level sustainability approaches requires metrics that can be used across different spatial scales. It is possible to aggregate the many of the producer-level metrics to make claims at greater scales (e.g., 'deforestation was 10% less amongst scheme members than amongst similar non-certified organisations'). However, difference at scales larger than a producer does not equate to describing an *outcome* at a larger scale: the sustainability issue may be worsening within a landscape even if the certified entities within it are bucking the trend. Conversely, those metrics designed to operate at a level greater than the individual producer (landscapes or watersheds) rarely provide data with sufficient spatial granularity to reflect the management practices of individual certified entities. Nonetheless, there may be value in aligning producer-level metrics to those that are produced at a larger spatial scale, so that performance can be monitored against external, context-specific baselines.
- **Glimpsing the future.** Technology is changing the way that information is collected (e.g., rapid mobile phone surveys), the scales at which data can be collected (e.g., big data analytics) and the uses to which data can be put (e.g., remote sensing based information being both a performance measure and a compliance tool). Technological advances will continue to open up new possibilities for VSS performance monitoring. VSS can enable the use of many existing and future technology-enabled metrics for most (if not all) of the sustainability issues covered in this report is by ensuring that they have accurate and granular location data on their clients.

Table of Contents

1 Introduction

1.1 Background

Voluntary Sustainability Standards (VSS) have grown rapidly in number and importance in global commodity markets in recent years. Alongside this spread in reach, the standards community – responsible for setting and evaluating standards – has made progress in researching its own impacts, investing in increasingly robust data collection, monitoring and evaluation systems.

However, the metrics generated and reported by standards systems have traditionally been designed to demonstrate the reach of the standard (e.g., number of clients, area certified) and compliance, and have mainly been focused at the level of certified site. As businesses in particular evolve their public policy commitments and seek to ensure that standards are providing the benefits they require in a cost-effective way, the demands for good outcome data increase. They seek to know whether standards systems are reducing Greenhouse Gas (GHG) emissions, and by how much, whether they are alleviating poverty and reducing deforestation, etc.

For sustainability standards, this means taking into account the following trends:

- The development of outcome-focused standards with quantified reporting of achievement, rather than the more traditional process-oriented standard, which defines what processes need to be complied with;
- Demands for better information about performance, with stakeholders wanting to have access to data and insights about performance against key metrics;
- For some sustainability issues, meaningful performance and impact needs to be measured and reported on at a scale larger than individual certified sites.

The shift towards more outcome-focused standards and measurements will support standards to credibly communicate how the results of their interventions and activities impact on the issues people care about.

This report summarises the findings of research commissioned by the ISEAL Alliance, the global membership association for credible sustainability standards.

1.2 Purpose and scope

The research is focused on six critical sustainability issues: deforestation, biodiversity, water use, forced labour, poverty, and GHG emissions. The aim of the research is to document a range of leading metrics and commensurate data sources that can be used to credibly measure and report on performance over time and across multiple spatial scales for each sustainability issue.

The intent of the research is not to build a definitive inventory of the best metrics and data sources but rather to focus on the applicability of various metrics and data sources, and best practices, limitations and trade-offs associated with their use.

In particular, the following three key considerations were identified by ISEAL to frame the research:

- *Considerations around defining the sustainability issue*. Is there global consensus on how to define each issue, or does it remain contested? If contested, what metrics are used to reflect the changes ISEAL members want to deliver, and how can progress be measured? Are new partnerships or initiatives forming that are set to deliver consensus around definitions and metrics?
- *Considerations around performance monitoring*. What are the trade-offs to reconcile in terms of data quality and accuracy vs what is feasible to collect and analyse at larger scales and across multiple actors? What is an appropriate spatial scale and timescale at which to measure each sustainability issue?
- *Considerations around data use, collection, sources and quality*. How feasible is it to collect the data and implement the metric, given that standards are all working across unique operational contexts? Are there baseline data available to help put performance measurement in context, and at what spatial scale are the baselines reported? Are there examples of good strategies for extrapolating data from individual sites up to landscapes? What is on the horizon for improving performance measurement for each issue in the future? Are new tools/software, data sources becoming available to support data collection and analysis?

1.3 About this report

The report summarises the methods used, followed by a chapter on each of the six sustainability issues that summarises the main decision points and trade-offs associated with metrics of its performance. The overall findings are summarised in a short conclusions chapter.

The report is intended to be useful as a stand-alone product, but readers that wish to understand more details of any metric are encouraged to consult the accompanying tables that are available as a separate document. These tables include the assessment of each metric against the considerations described in the following section, with explanatory notes and links to important further sources of information.

The research is intended to further contribute to ISEAL's thinking and member support around the implementation of indicators and metrics for improved performance measurement at appropriate spatial scales. The work was funded by the German Federal Ministry for Economic Cooperation and Development (BMZ) through GIZ.

1.4 How to use this report

This report is intended to provide information to support decisions about selecting sustainability metrics. It can be read in its entirety, or readers can navigate to relevant sections depending on their needs:

- For a brief overview of the current state of sustainability metrics and measurement, readers should see the **report introduction**. This is particularly useful to those who are new to sustainability metrics, or who are reviewing their measurement approach.
- Six sustainability issues are then presented in six stand-alone chapters. If a reader is interested in only one sustainability issue, they can navigate directly to the relevant chapter:

Greenhouse Gas Emissions

Deforestation

Biodiversity

Water Use

Poverty

Forced Labour

- At the start of each of these chapter, there is a Background section providing background to the sustainability topic. This is aimed at a non-expert audience and is particularly useful for readers who are new to the sustainability issue.
- Readers who are already well-informed about the sustainability issue may choose to skip the context section and navigate straight to the **Overview of metrics assessed**. This section focuses on the differences in definitions of the sustainability issue that the group of metrics reflect, and the practical trade-offs that need to be considered. We also highlight any future developments in metrics for the sustainability issue. A summary table detailed the metrics assessed for each sustainability issue. The metrics in these tables are presented in the order impact-outcome-output-input to facilitate quick like-with-like comparison. Full tables with more detail about the metrics and the assessment are available as a separate Excel database.
- Each chapter also contains a box with a **Case Study** from an organisation about their experiences using a metric to measure the sustainability issue in question.
- Those who are interested in how the metrics were selected and assessed for this report should see the **Methods** chapter.

Each topic chapter ends with a short summary section. An overall summary of the main conclusions of the report is given at the end.

2 Glossary

The following is a short glossary of some of the key terms used in this report.

Impact, Outcome, Output and Input
These group of terms describe what part of a theory of change is being measured by a metric:

Impact - Long-term effects (direct or indirect, intended or unintended) on the sustainability issue produced as the result of an intervention.

Outcome - The short-term and medium-term effects of an intervention on the sustainability issue in question.

Output - The products, capital goods and services which result from an intervention; may also include changes resulting from the intervention which are relevant to the achievement of outcomes.

Input - The financial, human, and material resources used for the intervention.

Baselines and benchmarks
These terms relate to data against which to compare information gathered for a metric in order to measure performance on an issue:

Benchmark - a reference required to convert data into a metric e.g., net household income data only becomes a poverty metric when compared to a benchmark such as the Living Wage Benchmark.

Baseline - a comparator in space and/or time against which to measure performance. In this report we distinguish between internal baselines (i.e., data collected repeatedly from a specific cite, or from selected comparison populations in the same area) and external baselines (i.e., external data that provides a broader contextual comparison for a metric).

Metric and indicator
These terms refer to data.

Metric - a system or standard of measurement (see Section 3.1).

Indicator - in this report we use the term to mean a type of data that only indirectly measures the issue in question (e.g., the presence or absence of rare species may indicate the broader levels of biodiversity). Note that the term is sometimes used in the literature to describe how metrics are used (e.g., the proportion of households above the poverty line is an ‘indicator’ of the success or otherwise of an intervention).

3 Methods

3.1 Definition of a metric

We use a broad definition of a 'metric', namely a system or standard of measurement. We therefore included:

- metrics that have an entire codified system of data capture, analysis and reporting (e.g., the Cool Farm Tool for on-farm GHG emissions),
- metrics that define how to collect data but are agnostic on its analysis and reporting (e.g., the Household Economy Approach for poverty),
- metrics that define how to analyse and report data of a particular type without specifying exactly how that data should be collected (e.g., the Shannon Index for biodiversity),
- frameworks that provide guidance on what characteristics of a system need to be measured but for which the precise data capture, analysis and reporting can be adapted to user requirements (e.g., the Sustainable Livelihoods Framework for poverty).

3.2 Criteria for evaluation

A number of key considerations are important when making decisions on which performance metrics to use. The considerations are organised under three themes: the definition of the sustainability issue; characteristics with respect to performance monitoring; and data use. These were further defined as 19 key topics in a framework which was used to analyse each metric in turn (Table 1).

3.3 Research framework

A framework was created to capture and analyse data on each metric in a consistent, structured and systematic manner. The framework included three types of information:

- Descriptive information (e.g., name of metric, unit of measurement);
- Analysis of the metric's characteristics against each consideration interest (e.g., the degree of consensus that the metric measures the sustainability issue); and
- Explanation of the analysis and key trade-offs, including references where appropriate.

The research framework was used to inform this report and to produce summary tables of the main characteristics of metrics under each sustainability issue, allowing readers to see some of the main strengths, weaknesses and trade-offs without needing to read all of details captured in the framework. It is available as a separate document (spreadsheet) for readers who would like to explore the characteristics of the metrics in more detail.

Table 1:

Key considerations against which metrics were evaluated

Scheme	Schemes included	Examples
Sustainability definition	Degree of consensus that the metric measures the sustainability issue	The metric measures either the whole of the sustainability issue or the main driver of it
	Degree that the metric is underpinned by evidence	The metric should have a firm basis in science or social science theory (peer reviewed literature), good data and/or has significant credible application
	Extent to which the metric is widely used	The metric is used by organisations with global reach
	Is the metric readily communicable?	Does the metric make sense to engaged people who are not necessarily experts? Links to degree of consensus and how widespread it is
	Addresses material impacts	The metric captures the main sources of the impact or main impacts
	Emerging initiatives	(descriptive)
Performance monitoring	Practical to collect	Refers to primary capturing of the data
	Practical to implement	Refers to the collation, management, analysis and communication of the data by a VSS or ISEAL
	Context sensitive	The metric should be relevant (material) to all sectors and geographies
	Outcome orientated	Metrics need to be as closely linked to environmental impacts as possible (e.g. reductions in greenhouse gas emissions), rather than proxy indicators (e.g. proportion of suppliers with a carbon reduction target). However, some process/input metrics can be helpful and have been include. The categories here are: input, output, outcome and impact and are defined in the Glossary.
	Responsive	The metric should be able to detect change resulting from an intervention
	Scalable	In relation to the 'scale of application': ideally it would be possible to aggregate data to higher levels (e.g., from farm to landscape)
	Based on primary data	The metric needs to relate to the activities of ISEAL members/certificate holders, and not be heavily reliant on out-of-date secondary sources of information or on estimates

3.4 Desk review

The main body of research was a desk review of metrics used for each sustainability issue. Principal sources of information were academic publications, websites and reports of companies, NGOs and metric developers and communities of practice, work done by ISEAL on the metrics used by its members, and in-house expertise.

As described in the previous section, the main task of this research was to identify a range of leading metrics and indicators and commensurate data sources that can be used to credibly measure and report on performance across the six sustainability issues. The intent was not to build a definitive inventory, but rather to focus on the applicability of various types of metrics and data sources, best practices, limitations and trade-offs associated with their use. We therefore focused on identifying and assessing a range of metrics with different characteristics (in particular; the scope, measurement unit and whether the metric is an outcome, output or input measure) rather than creating a list of near-identical metrics. Typically, we identified up to fifty metrics under each sustainability issue before discounting at least half as being duplicates, near duplicates or because they were not appropriate to ISEAL members' needs.

3.5 Interviews

Insights into implementing metrics – that is, embedding the use of a metric within an organisation – were conducted once the desk review had been completed. The purpose of the interviews was two-fold: to sense-check our analysis of the metric and similar metrics used by the informant, and to give insight into the processes of embedding performance metrics within complex organisations.

A total of six semi-structured interviews were held, focused on a single metric from each of the sustainability issues. Interview questions were prepared in advance and covered five topics:

- Why that particular metric was chosen over others to measure this aspect of sustainability;
- What the challenges of using the metric have been;
- What the benefits or successes of using this metric have been;
- Specific use examples of the metric; and
- Major learnings from implementing the metric.

Questions were designed to open up interesting and potentially unexpected insights that could then be followed up with further questioning. Informants were selected based on their use of a leading metric within each issue and on their organisation being likely to encounter similar issues in embedding metrics as would ISEAL members. The findings from these interviews are presented in case study boxes in the relevant chapters.

3.6 Limitations of this study

The range of sustainability metrics in existence is not equal across sectors or sustainability issues, and this is reflected in our search. In general, metrics designed for use in the agricultural and forestry sectors were well represented. The majority of metrics found for two of the sustainability issues – biodiversity and deforestation – may be less applicable to some of the sectors in which some ISEAL members operate, which include, marine fisheries, aquaculture, aluminium, jewellery, and golf. This is perhaps less true for GHGs and water use, which include metrics that are designed for different types of production sites, farms and supply chains and are therefore more universally applicable. Nevertheless, VSS in the agricultural and forestry sectors will generally have a wider choice of potential metrics to consider.

A further limitation of the study is that we were only able to speak to a small number of metric users. Whilst we hope that these use case studies provide some general lessons that will resonate with all VSS, we recognise that this small sample of organisations means that some of the specific contexts of ISEAL members may not be represented.

Finally, the analysis of the metrics, although systematic and consistent, is still essentially subjective. For example, in answer to the question 'How practical is it to collect the data?' we have judged each metric as high, medium or low. Ideally, answers to questions such as this could have been quantified, in terms of effort and/or cost. However, this information is rarely available in the literature, varies hugely according to the specific use to which the metric is being put, and compiling quantitative information for multiple questions across over 100 metrics was beyond the scope of the research. We would encourage anyone who has an interest in a particular metric to directly contact organisations that developed or are using them in order to find out more detail. For this reason, have included numerous references and links in the Excel database that accompanies this report.

4 Greenhouse Gas Emissions

4.1 Background

A large amount of research has been undertaken and published on climate change metrics, driven by the very significant global scientific and policy focus on this issue. The science of climate change – including mitigation and impact measurement – is assessed by a UN body, The Intergovernmental Panel on Climate Change (IPCC)[1]. Given the history and importance of this institution, corporate and public sector climate change accounting and reporting approaches are heavily influenced by the methods published in its 'Assessment Reports', the sixth of which is due in 2021. Climate change metrics can be broadly bundled into three categories:

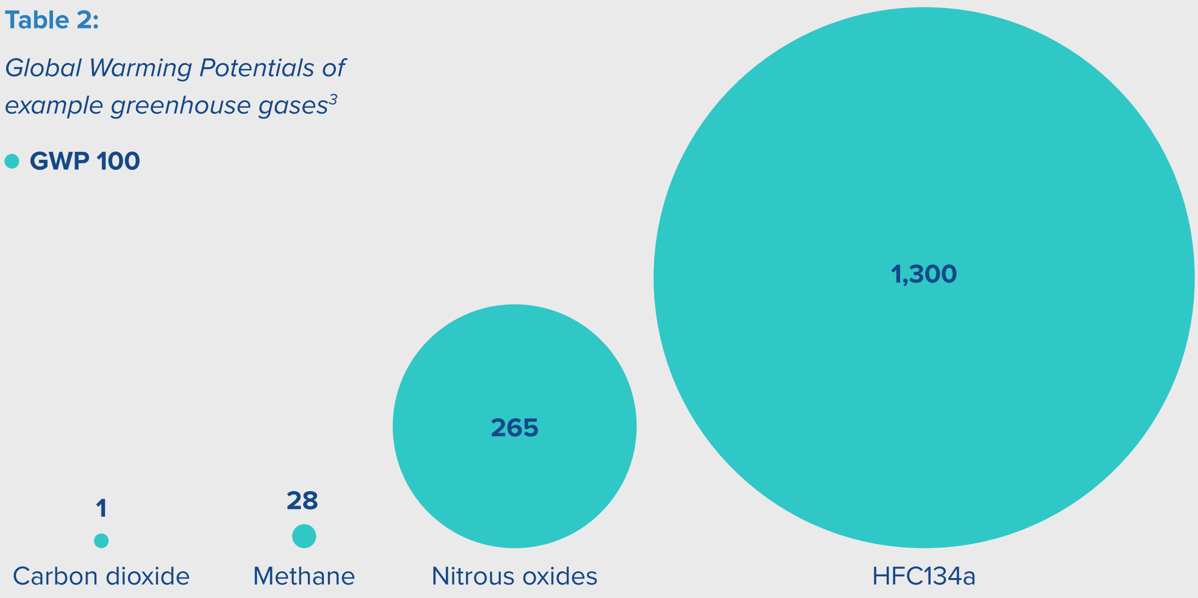
- **Quantification of emissions and/or carbon dioxide removals using Global Warming Potential (GWP)** and expressed as quantity of carbon dioxide equivalents (CO_2e), each with variations in terms of scope of emissions covered (e.g., *Direct greenhouse gas emissions ('Scope 1 & 2'), Value chain greenhouse gas emissions ('Scope 1, 2 & 3'), Product carbon footprints,* see Table 4). These metrics generally perform well across the range of considerations assessed here. In particular, they are well aligned to the consensus definition of climate change, underpinned by a wealth of scientific evidence, practical to implement, scaleable, and have comprehensive support and guidance available for implementation. An introduction to the concept of Global Warming Potential is given below.
- **Scale of investment in or uptake of emissions mitigation (low carbon) activities** e.g. Percentage of energy from renewable energy, hectares of soil management or habitat protection (e.g., *Investment in greenhouse gas mitigation activities*, see Table 4). In general, these metrics are context sensitive, scalable, relatively easy to collect and implement, but are often input metrics (therefore several steps away from actual changes in global temperatures) and may have lower credibility
- **Efficiency metrics linked to inputs that have a significant climate impact of production or use** e.g. fertilizers, energy, animal feed (see *Carbon dioxide emissions, Greenhouse gas emissions from the production and use of fertilizers per tonne of production and Greenhouse gas emissions from fertilizers per hectare of production*, see Table 4). These metrics are typically responsive to changes in producers' management practices, scalable, have guidance for implementation and are underpinned by evidence, but are less context sensitive, address only a subset material impacts, and typically lack an external baseline against which to measure performance. These efficiency metrics have the benefit of being often more clearly linked to business bottom-line cost efficiencies.

4.1.1 Global Warming Potential (GWP)– the gold standard climate metric

Due to the international research and policy focus on climate change, the emissions measure Global Warming Potential (GWP) has developed as the gold standard for quantifying and comparing the contribution to climate change of an organisation, activity, region, etc. Global Warming Potential has also become the default measure for expressing emissions of different gases, such as methane and nitrous oxide, on a common scale; so-called, carbon dioxide equivalents or CO_2e. As the reference gas, carbon dioxide has a GWP of 1 (see Table 2)[2]. By comparison, some substances such as the refrigerant HFC134a have a global warming potential many times greater i.e. 1kg of HFC134a is equivalent to 1,300kg of CO_2.

Table 2:

Global Warming Potentials of example greenhouse gases[3]



4.1.2 Measurement vs modelling emissions

It is important to note that emissions are very rarely ‘measured’ in a physical sense, for example through the use of meters to quantify volumes of gases that come from a process. Rather they are normally *modelled* by combining input or activity data (e.g. litres of diesel combusted) with published ‘conversion factors’ or ‘emissions factors’. For example, in the United Kingdom, the government publish an annual update to a database that describes average emissions for a range of activities and different technologies[4] including the consumption of a kWh of grid electricity or a kilometre travelled in an average car. This inevitably introduces a degree of uncertainty into calculations, which varies depending on the system being modelled. For example, there is relatively low uncertainty when estimating quantities of carbon dioxide from burning a litre of diesel compared to relatively high uncertainties for estimating the quantities of nitrous oxides that result from the application of nitrogen fertilizers to soils.

Emissions factors for converting input data into emissions are frequently sourced from licensed databases (such as Ecoinvent[5]), peer-reviewed articles (e.g. the International Journal of Life Cycle Assessment[6]), from industry reports and research and from other non-governmental organisations. Examples of emissions calculation tools and databases are available to download from the GHG Protocol website.[7]

Complex models are often used for quantifying emissions associated with agricultural processes, for example nitrous oxide emissions from soils or methane emissions from ruminant livestock. These are frequently derived from IPCC methodologies and implemented in user-friendly excel or web tools, such as Cool Farm Tool (see Table 3).[8] These models take the form of complex equations featuring a range of parameters and constants that need to be adapted depending on the location and production.[9]

4.1.3 Alternative emissions metrics

The downsides of the GWP method are well known - no emissions metric is perfect - as the IPCC 5th Assessment Report acknowledges: "*The most appropriate metric will depend on which aspects of climate change are most important to a particular application, and different climate policy goals may lead to different conclusions about what is the most suitable metric with which to implement that policy.*"

A recent example of alternative emissions metrics includes proposals from researchers at the Oxford Martin School at the University of Oxford[10] who have highlighted the need for metrics that better serves policy decision making on short term pollutants, such as methane. As we highlighted above, each molecule of CO_2 released into the atmosphere today is adding to existing CO_2 in the atmosphere and continuing to increase warming, even if the absolute amount of CO_2 emissions decrease year on year. On the other hand, if methane emissions levelled off now, no *additional* warming would occur – and in fact some cooling could occur if methane emissions fell. The implications of this are most important for setting policies and targets on climate mitigation. The researchers have developed a new measure (called GWP*) that is more useful for assessing future warming trajectories and, they argue, is therefore more helpful in assessing the merits of different options for mitigation (i.e. comparing carbon taxes on fossil fuel combustion vs. methane emissions from cattle).

However, until alternative emissions measures are rubber-stamped by the IPCC and integrated into international reporting under the UNFCCC, it is unlikely to gain wide-spread adoption under corporate reporting frameworks, such as the GHG Protocol.

4.2 Overview of the metrics assessed

We assessed 19 climate change metrics (see Table 3). These included twelve outcome metrics (i.e., estimation of GHG emissions); three output metrics (e.g., tallying commitments or plans to reduce emissions); and four input metrics (e.g., the quantity of energy used in a production process,

Table 3:

Description of GHG metrics

Metric	**Metric description**
Sectoral greenhouse	(Many) government collects data on economy-wide GHG emissions as part of international climate change agreements
Direct greenhouse gas emissions ('Scope 1 & 2')	Metric principally covers emissions that occur directly from an organisation and its use of purchased electricity. For agricultural businesses, Scope 1 emissions also include more complex emissions from livestock, soils, and manures. Conversion factors and models convert energy and process data to emissions estimates. The most frequently used indicator is Global Warming Potential 100 expressed as carbon dioxide equivalents (CO_2e). Multiple accounting and reporting standards, guidance and tools exist e.g. The GHG Protocol.

Table 3 cont:

Description of GHG metrics

Metric	Metric description
Value chain greenhouse gas emissions ('Scope 1, 2 & 3')	This metric can include all direct and indirect greenhouse gas emissions attributable to an organisation: both those that occur on-site but also those that occur upstream (e.g. through production of purchased goods) and downstream (e.g. consumer use and disposal impacts). The most frequently used indicator used is Global Warming Potential 100 expressed as carbon dioxide equivalents (CO_2e). Multiple accounting and reporting standards, guidance and tools exist e.g. The GHG Protocol.
Direct greenhouse gas emissions per unit of production ('Scope 1 & 2')	Most frequently used metric is Global Warming Potential 100 (GWP100) expressed as carbon dioxide equivalents. Multiple accounting and reporting standards, guidance and tools exist e.g. The GHG Protocol. This metric expresses emissions per tonne of production (also known as an 'GHG intensity' measure)
Greenhouse gas emissions from the production and use of fertilizers per tonne of production	Fertilizer type and information on the quantity and methods for application is combined with secondary data and tools to calculate total emissions. Multiple methods exist for calculating emissions from soils - many derived from IPCC methods used in national GHG inventories. Total emissions are then divided by quantity of production for the same time period e.g. 1000 tonnes of crop.
Greenhouse gas emissions from fertilizers per hectare of production	Fertilizer type and information on the quantity and methods for application is combined with secondary data and tools to calculate total emissions. Multiple methods exist for calculating emissions from soils - many derived from IPCC methods used in national GHG inventories. Total emissions are then divided by area of production for the same time period e.g. 1000 hectares of crop
Value chain greenhouse gas emissions per unit of production ('Scope 1, 2 & 3')	This metric can include all direct and indirect greenhouse gas emissions attributable to an organisation: both those that ocurr on-site but also those that occur upstream (e.g. through production of purchased goods) and downstream (e.g. consumer use and disposal impacts). The most frequently used indicator used is Global Warming Potential 100 as carbon dioxide equivalents (CO_2e), expressed per unit of production (i.e. GHG intensity). Multiple accounting and reporting standards, guidance and tools exist e.g. The GHG Protocol.
Carbon dioxide emissions (total or per unit output)	This greenhouse gas metric focuses purely on carbon dioxide - rather than including all greenhouse gases. It tends to focus on direct emissions from the combustion of fossil fuels and/or use of electricity.
Product carbon footprint	This metric can include all direct and indirect greenhouse gas emissions attributable to a single product: both those that occur on-site but also those that occur upstream (e.g. through production of purchased goods) and downstream (e.g. consumer use and disposal impacts). The most frequently used indicator used is Global Warming Potential 100 as carbon dioxide equivalents (CO_2e) and expressed per unit of production (i.e. GHG intensity). Multiple accounting and reporting standards, guidance and tools exist e.g. The GHG Protocol.

Table 3 cont:

Description of GHG metrics

Metric	**Metric description**
Net greenhouse gas emissions	Net' greenhouse emissions are typically calculated by combining greenhouse gas emissions from an organisation and netting off any greenhouse gas removals or avoidance attributable to the company's activities or investments. For instance, carbon removals from forestry, the purchasing of carbon offsets, or avoided emissions through the export of renewable energy. There is not standard definition for establishing the scope of emissions included i.e. this could cover direct (Scope 1 and 2) emissions only.
Emissions per unit area (remote sensing)	Remote sensed satellite data is used to estimate greenhouse gas fluxes from regions
Carbon dioxide removals	This metric measures the amount of carbon dioxide removed from the atmosphere, in CO_2e. This is typically through natural processes e.g. afforestation, soils management, peatland restoration. Currently there is a very wide range of methodologies used - many derived from IPCC national inventory methods. The GHG Protocol is currently developing new guidance on accounting and reporting for carbon removals.
% of producers with energy efficiency policy or plan	% (or number) of farms with an action plan
Percentage of producers with credible greenhouse gas reduction targets	% of organisations that have set greenhouse gas reduction targets in line with climate science (e.g. Science-based targets). Targets adopted by companies to reduce greenhouse gas (GHG) emissions are considered "science-based" if they are in line with what the latest climate science says is necessary to meet the goals of the Paris Agreement – to limit global warming to well-below 2°C above pre-industrial levels and pursue efforts to limit warming to 1.5°C.
Percentage of producers adopting GHG mitigation practices or technologies	There are a wide range of farm practices that can reduce GHG emissions from agriculture. Monitoring the uptake of these mitigation methods provides an indicator of progress towards achieving a reduction in agricultural production emissions. Example applications: Fairtrade; UK Government GHG Indicators for agriculture.
Fossil energy use per unit production (direct)	Metric focuses on quantifying the quantity of fossil energy used in the production of goods and services. This can be direct energy use only or include the 'embodied' energy or purchased goods and services used by the organisation.
% of raw materials from local sources	This metric measures the extent to which raw materials used by a business have needed to be transported
% of energy from renewable sources	Renewable energy use divided by total energy use on site or by company. Requires decision on whether this only includes on-site generation or purchasing of renewable energy
Investment in greenhouse gas mitigation activities	Amount of money invested in projects, operational changes, new assets, etc. that deliver greenhouse gas reductions across a single or multiple years

4.3 Applicability and trade-offs

4.3.1 Key technical decision points in measuring GHGs

Scope of measurement

The most significant variation between metrics is the scope of what is included. This ranges from very specific processes (e.g. emissions from nitrogen fertilizer, see *Greenhouse gas emissions from the production and use of fertilizers per tonne of production,* Table 4), through to 'total carbon footprint' type calculations that attempt to quantify the total emissions associated with the full life cycle of a product, from raw material extractions through to consumer use and disposal at end-of-life (e.g., *Product carbon footprint,* Table 4)

The scope of GHG emissions is often broken down into three categories:

- **Scope 1 – All Direct Emissions** from the activities of an organisation or under their control, including fuel combustion on site such as diesel used in vehicles, gas boilers. An example of a metric that considers Scope 1 emissions only is *Carbon dioxide emissions* (see Table 5)
- **Scope 2 – Indirect Emissions** from electricity purchased and used by the organisation. Several of the metrics assessed include both Scope 1 and 2 within the same protocol, for example, *Direct greenhouse gas emissions ('Scope 1 & 2')*, and *Direct greenhouse gas emissions per unit of production ('Scope 1 & 2')*, both defined by the GHG Protocol[11] (see Table 4). As a general rule, organisations should be encouraged to measure Scopes 1 and 2, as these are the emissions that are within their direct control.
- **Scope 3 – All Other Indirect Emissions** from activities of the organisation, occurring from sources that they do not own or control, covering emissions associated with extraction and production of purchased materials; transportation of purchased fuels; and emissions associated with purchased services. Examples of metrics that include Scopes 1, 2 and 3 include *Value chain greenhouse gas emissions ('Scope 1, 2 & 3')* from the GHG Protocol[12] (see Table 4). For some organisations, particularly those at the top of supply chains, Scope 3 emissions can be the largest share of their overall emissions footprint. However, measuring and reducing Scope 3 emissions requires collaboration throughout the supply chain, which can be challenging.
- Some organisations, such as the Sustainable Food Lab, additionally argue in favour of a further landscape-level category of emissions, 'Scope 4'.[13] **Scope 4** encompasses all of the enterprises or farms in a watershed, and so is necessary to address water stewardship or biodiversity. An example of a metric that could be used to estimate Scope 4 is *Emissions per unit area (remote sensing)*[14], see Table 4). Note that there is little consensus on Scope 4, and in fact other radically different definitions exist for what an additional scope might be.

While the IPCC define methods for accounting and reporting national-level emissions inventories under the UNFCCC, multiple global, regional, and national accounting and reporting standards and guidance materials have developed to support more consistent approaches to quantifying emissions from organisations, product and projects. These can be voluntary or as part of local climate regulation. The leading international approaches are published by the Greenhouse Gas (GHG) Protocol and also by ISO (e.g. ISO14064). These frameworks typically reference IPCC concepts, methods and definitions.

While these standards typically require that GWP100 is used, they can be more or less prescriptive on the scope and boundaries of an inventory i.e. which processes and sources of emissions of a value chain to include. Generic corporate reporting frameworks, such as the GHG Protocol Corporate Standard, tend to focus on ensuring transparency of methods over highly prescriptive requirements on exactly which sources to include and how to estimate them (e.g. which secondary sources of emissions factors should be used for calculations). This means that most voluntary reporting standards do not deliver comparability. To achieve this, much tighter restrictions and guidance is needed within schemes to assure alignment on key assumptions and background data sources.

In general, many consider it to be critical that organisations get good quality data on, and deliver reductions in, Scope 1 & 2 emissions (i.e. direct emissions and purchased electricity). This is where organisations have greatest influence and where reductions can most credibly be quantified. Best practice beyond that is to quantify and look for options to reduce Scope 3 emissions in collaboration with key suppliers. In primary industries such as agriculture, forestry and fisheries, a significant proportion of their emissions are likely to be Scope 1 & 2. This is may not be the case in other sectors, and for actors higher in the supply chain.

Benchmarked performance or targets?

Climate action has increasingly been framed not around benchmark performance but around delivering 'science-based' emissions reductions; the reductions needed to limit global temperature increases to 2°C or 1.5°C. More recently, targets to reach 'Net Zero' emissions by 2050, or earlier, have also grown in popularity (see Table 4).

With the increasing adoption of 'Net Zero' and 'science-based' targets commitments, the focus of organisational climate action activity has been commitments that lead to absolute reductions in GHG emissions (as opposed to *relative* commitments to improve GHG efficiency but which may still lead to growth in tonnes of emissions). There are different routes to achieving overall reductions in emissions at a sectoral, national and ultimately global level. Two basic approaches exist: either all organisations decarbonise at the same rate, or, some organisations decarbonise faster than others but overall carbon budgets are met in aggregate (see Table 4). Currently, Sectoral Decarbonisation Approaches (SDAs) have been developed for sectors including Power, Pulp & Paper, Aluminium, Iron & Steel, Cement[15].

Figure 1:

Greenhouse gas removals and the concept of Net Zero[16]

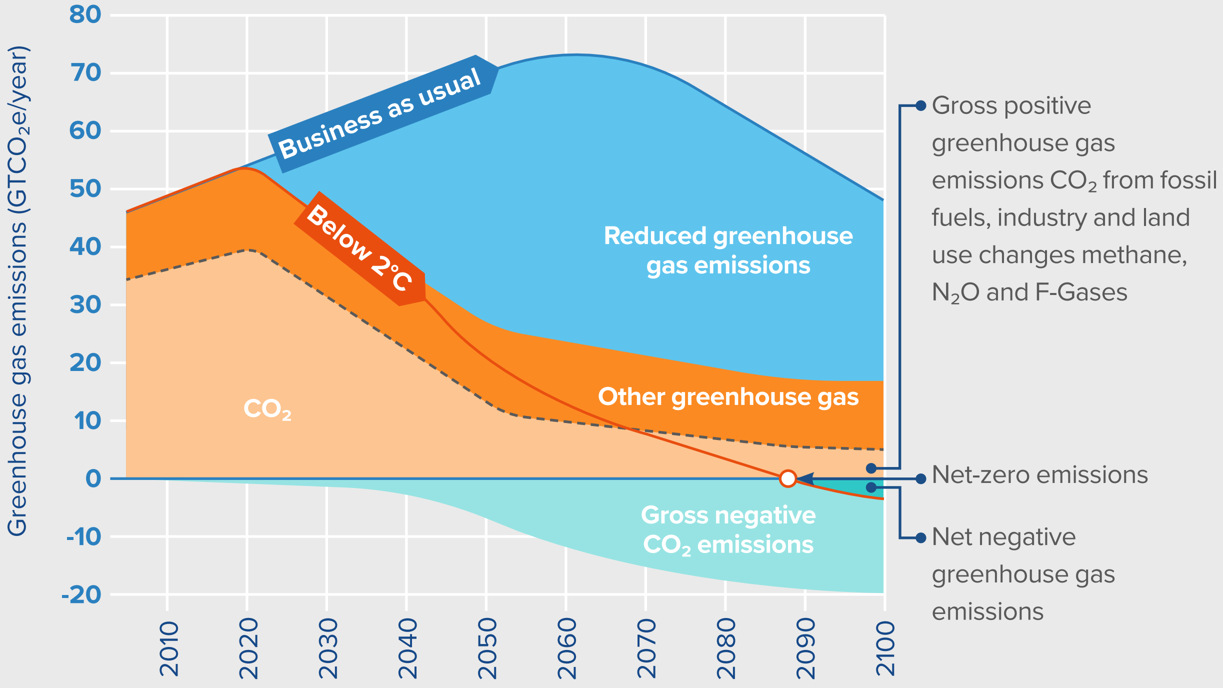
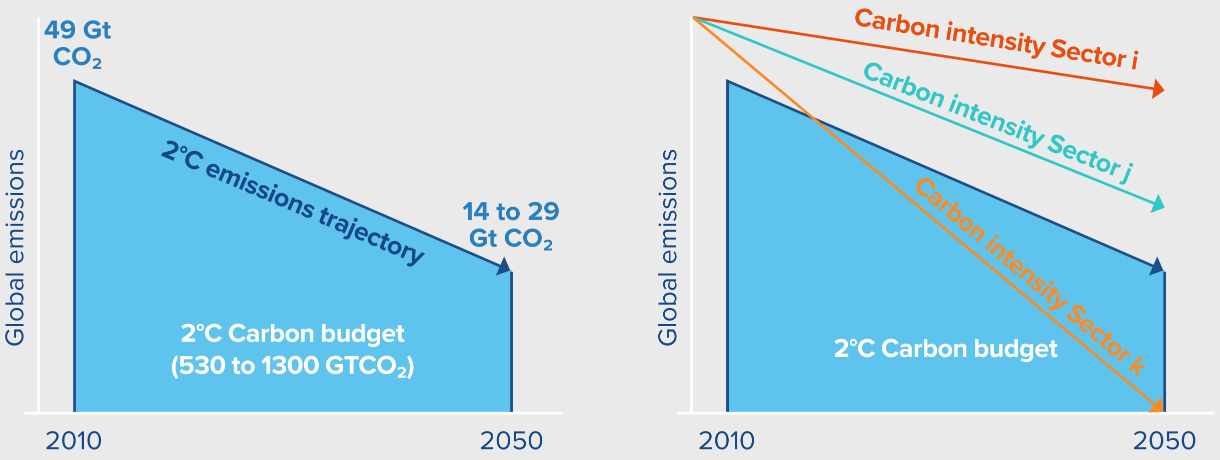


Figure 2:

Absolute reductions vs Sectoral Decarbonisation Approach[17]



There are currently no sector-specific methodologies available for setting and measuring emissions reduction targets in land-based sectors. This is of particular relevance to ISEAL and many of its members as agriculture and forestry pose some unique emissions mitigation challenges and opportunities. The Science Based Targets initiative is undertaking a sector development project, the SBTi Forest, Land and Agriculture project ('SBTi FLAG'), led by WWF, to address this methodology gap. According to the SBTi, "*the effort will focus on the development of methods and guidance to enable the food, agriculture, and forest sectors to set science-based targets (SBTs) that include deforestation, and other land-related impacts*". The project is expected to be completed by Q2 2021.

The important point here goes beyond metrics; it is about defining what 'good' looks like, i.e. what GHG emissions reductions should be made, and by when.

4.3.1.1 Key practical trade-offs

The review of metrics identified some consistent themes relevant to many of the climate metrics.

Communication and benchmarking
A significant limitation of emissions metrics is the challenge in communicating Global Warming Potential (GWP) to non-experts. The measure (carbon dioxide equivalents, CO_2e) is not intuitive and the units (tonnes of emissions) is hard to conceptualise and put into context: is one tonne of emissions a lot or a little? Linked to this, there are also limited opportunities for organisations to compare and benchmark emissions performance due to the inevitable differences in the scope and boundaries of metrics published by different organisations.

These challenges have resulted in many organisations adopting proxy indicators that are easier to communicate and still demonstrate that the issue is being addressed in a substantive way. The clearest example of this is the use of metrics that quantify the proportion of energy from renewable sources (e.g., '*% of energy from renewable sources*', Table 4).

Dependence on models and secondary data
There is a significant dependence on models and secondary data to quantify emissions as it is not practical or economic to measure greenhouse gas emissions and removals directly.

In some areas this is not an issue, as there is a relatively low degree of uncertainty in the systems being modelled, for example, how much carbon dioxide is released from the combustion of a litre of diesel. However, some models have a high degree of model and parameter uncertainty, for example, estimating nitrous oxide emissions from the application of nitrogen fertilizer to soils. In addition, there is a high dependence on secondary sources of emissions factors and 'average data' for developing some GHG inventories. For example, a food manufacturer would commonly use published data on the typical carbon footprint of ingredients to develop an inventory that covered upstream impacts, rather than collecting primary data from their supply chain. The quality and relevance of these data points can be poor and not reflective of reality, particularly in non-conventional production systems. This makes using secondary data problematic for tracking changes in emissions over time in a credible manner, especially when the principles and criteria of VSS demand a change from conventional production (e.g., conventional to organic agriculture). As a result, it is best practice to use as much primary data as possible and ensure any secondary data is representative of the time period, technologies, etc.

In summary, there is a trade-off between the cost and practicality of collecting primary data and the loss of context specificity and responsiveness that comes with heavy reliance on secondary data and models.

Use and reporting of emissions reductions via market-based mechanisms

The use and reporting of emissions reductions via market-based mechanisms – most notably carbon offsets, renewable energy contracts and credit-based zero deforestation certifications – is the most contentious area of emissions metrics. While these have become some of the most popular means of organisations demonstrating that they are tackling climate change, there is on-going debate around the degree to which these market-based systems lead to emissions reductions and how they should be reflected in emissions inventories, if at all. There are also ethical considerations of tying up resources (e.g., land) in developing countries so that companies in developed nations can continue to emit GHGs.

Some emissions reporting initiatives and frameworks allow these reductions to be reported but separately from the main emissions inventory. This is to enable stakeholders to identify that these emissions reductions are not happening within their own operations or value chain. The most developed example of this is the GHG Protocol Scope 2 Guidance which describes how organisations should account for and report emissions from electricity, where the organisation is purchasing renewable energy from a 3rd party (e.g., a 'green tariff' from an energy retailer or a Power Purchase Agreement[18]). It is likely that a greater global consensus on the inclusion of other market-based approaches for the land sector will emerge from new guidance and standards being developed by WWF and the GHG Protocol in 2020/2021:

- GHG Protocol Standards/Guidance on Carbon Removals and Land Use[19]
- SBTi guidance on Net-Zero Targets in the Corporate Sector[20]

Shared value for clients and VSS

A client that wishes to reduce their greenhouse gas emissions is likely to prefer a metric that allows them to see where their largest emissions come from and is able to reflect any changes they make to those sources linked to operational data and efficiency (e.g., '*Greenhouse gas emissions from fertilizers per hectare of production*'). However, this type of metric will be amongst the least useful for many VSS, as it limits claims that can be made about the scheme's contribution to reducing greenhouse gas emissions, or more significantly, reducing GWP to zero. A potential solution to this trade-off would be to limit the metrics demanded of clients to those of most practical use to them, whilst developing specific models, validated with a sample of clients, to estimate the global impact of the VSS on GWP.

Table 4.

Characteristics of GHG metrics

Scale of application
- ⏩ **Supply chain**
- ⚙ **Producer**
- 🏢 **Company**
- 🏞 **Landscape**
- ◔ **Sector**

◉ **High** ◎ **Medium** • **Low**

			Considerations around sustainability definition				Considerations around performance monitoring					Considerations around data use			
Metric type	**Metric description**	**Scale of application**	Metric measures the sustainability issue	Metric is underpinned by evidence	Metric addresses material impacts	Metric is widely used and easily communicable	Practical to collect	Practical to implement	Context-sensitive	Responsive to change	Scalable	Baseline data available to provide context	New tools/software emerging?	Implementation 'user package' exists	**Main issues/comments**
Outcome	Sectoral greenhouse gas emissions and removals	◔	◎	◎	◎	◎	◉	◉	◉	•	◉	✓	✓	✓	Based on sectoral averages so may significanly under- or over-estimate GHG emissions from specific production systems or contexts.
	Direct greenhouse gas emissions ('Scope 1 & 2')	🏢	◉	◉	◉	◉	◉	◉	◉	◉	◉	✓	✗	✓	Outcome metric that is well established and relatively easy to quantify accurately. It covers emissions that reporting organisation has greatest control over. Metric is embedded in many climate reporting and target setting initatives.
	Value chain greenhouse gas emissions ('Scope 1,2 & 3')	🏢	◉	◉	◎	◎	•	◉	◉	◎	◉	✗	✗	✓	Metric that should fully account for the total direct and indirect climate impacts of an oranisation. However challenges in accessing primary data to underpin calculations means that analyes are often based ons econdary data/assumptions that are not sensitive to change.
	Direct greenhouse gas emissions per unit of production ('Scope 1 & 2')	🏢	◉	◉	◉	◉	◉	◉	◉	◉	◉	✓	✗	✓	Outcome metric that is well established and relatively easy to quantify accurately. It covers emissions that reporting organisation has greatest control over. Metric is embedded in many climate reporting and target setting initatives.
	Greenhouse gas emissions from the production and use of fertilizers per tonne of production	⚙	◎	◉	•	◎	◎	◉	•	◉	◉	✗	✗	✓	Example of GHG metric that looks at a subset of emissions that are likely to be highly relevant to some sectors.
	Greenhouse gas emissions from fertilizers per hectare of production	⚙	◎	◉	•	◎	◎	◉	•	◉	◉	✗	✗	✓	Example of GHG metric that looks at a subset of emissions that are likely to be highly relevant to some sectors.
	Value chain greenhouse gas emissions per unit of production ('Scope 1,2 & 3')	🏢	◉	◉	◎	◎	•	◎	◉	◎	◎	✗	✗	✓	Metric that should fully account for the total direct and indirect climate impacts of an oranisation. However challenges in accessing primary data to underpin calculations means that analyes are often based on secondary data/assumptions that are not sensitive to change.
	Carbon dioxide emissions (total or per unit output)	⚙	◉	◉	◎	◎	◉	◉	◉	◉	◉	✓	✗	✓	Basic measure of the relative carbon efficiency of a process or organisation. The exclusion of non-CO2 gases means it will focus mainly on the principle emissions from the combustion of fossil fuels.
	Product carbon footprint	⏩	◉	◉	◎	◎	•	◎	◉	◎	◎	✗	✗	✓	Metric that should fully account for the total direct and indirect climate impacts of an oranisation. However challenges in accessing primary data to underpin calculations means that analyes are often based on secondary data/assumptions that are not sensitive to change.
	Net greenhouse gas emissions	🏢	◎	◉	◎	◎	◉	◉	◉	◉	◉	✗	✗	✗	There is an increasing desire by businesses to use GHG metrics that take account of both emissions and potential removals/avoided emissions. However, there is no consensus yet on how to do this credibly. Global guidance is being developed to address this.
	Emissions per unit area (remote sensing)	🏞	◉	◎	•	◎	•	•	◉	◉	◉	✓	✓	✗	Emerging scientific and research area with potential to reduce the cost and increase quality of jurisdictional assessments of greenhouse gas emissions and track changes over time.
	Carbon dioxide removals	⚙	◎	◉	◉	◉	◎	◎	◎	◉	◉	✓	✓	✗	This metric is likely to become increasingly important in demonstrating business and sectoral contributions to 'net zero' commitments. However the complexity of processes involved and number of methods used means that a consensus metric may be some way off.
Output	% of producers with energy effecency policy or plan	◔	•	•	•	◎	◉	◉	◉	◎	◉	✗	✗	✗	Measures the intent to reduce GHG emissions not the actual reduction, and does not cover all sources of GHG emissions.
	Percentage of producers with credible greenhouse gas reduction targets	◔	◎	◎	•	◎	◉	◉	•	•	◉	✗	✗	✓	Very much a process metric, as companies may not reach their targets or revise them over time.
	Percentage of producers adopting GHG mitigation practices or technologies	◔	◎	•	◉	◉	◎	◉	◉	◉	◉	✗	✗	✓	Metric commonly used by standards and policymakers as a proxy for understanding climate mitigation. Evidence linking practice adoption to outcomes is less strong.
Input	Fossil energy use per unit production (direct)	⚙	◎	◉	•	•	◎	◎	◉	◉	◉	✓	✗	✗	Quantifies key input that drives significant amount of overall climate impact - however ignores some important sources of emissions and is hard to communicate and seldom used outside of life cycle assessment.
	% of raw materials from local sources	⚙	•	•	•	◎	◉	◉	◉	◉	◉	✗	✗	✗	A climate metric with a long history that has fallen out of favour recently due to the fact it frequently only addresses a small proportion of overall impact.
	% of energy from renewable sources	⚙	◎	◎	◉	◉	◉	◉	◉	◉	◉	✓	✗	✓	Popular climate metric that is relatively easy to quantify and communicate. It addresses a key source of impact but some critics challenge the credibility of claims where businesses are using market-based 'green tarriff' products that do not drive new renewables capacity.
	Investment in greenhouse gas mitigation activities	🏢	◎	•	◉	◉	◉	◉	◉	◎	◉	✗	✗	✗	Potentially powerful indicator of the seriousness with which certified companies and VSSs take climate change, but an input indicator not directly related to GHG emissions. Indeed poorly targeted investment could result in an unintended increase in emissions.

Box 1.

Implementation of GHG metrics: The Sustainable Food Lab

Based on an interview with Daniella Malin, Senior Program Director, Agriculture & Climate at Sustainable Food Lab

The Sustainable Food Lab (SFL) has been developing and implementing climate change metrics with a range of organisations in food and agriculture since it was founded in 2004. The non-profit organisation has also played an instrumental role in the development of the Cool Farm Alliance[21] – a pre-competitive collaboration that seeks to enable the quantification of farm-scale greenhouse gas emissions and carbon dioxide removals by farmers and their customers. The Cool Farm Tool has been used by a range of actors seeking to apply greenhouse gas metrics, for example, environmental NGOs, such as WWF (for cotton)[22] and standards setters, such as 4C (for coffee)[23].

The Cool Farm Tool approach includes all significant sources and carbon sinks that occur on a farm and in its value chain. It uses IPCC-derived methods and expresses results in $kgCO_2e$ per unit of production, as well as in absolute terms (e.g. tonnes of carbon dioxide equivalent). Emissions sources and sinks are reported separately to see 'gross' and 'net' emissions results. A summary of methods is available online[24].

The SFL have experience of quantifying emissions using the Cool Farm Tool in a wide variety of sectors through their Cool Farming Options[25] project and work with agricultural businesses, such as Costco (for eggs)[26]. Through this work the SFL has seen how the adoption of a whole-farm GWP metric can be an excellent way of quantifying and communicating the environmental benefits of changes to farming systems and practices. The SFL sees user-friendly web and Excel tools as critical when implementing emissions metrics as they help turn the often complex models into more accessible and understandable calculations that non-experts can engage with. Despite this, they acknowledge the variety and number of input data needed to feed into calculations can be a barrier to adoption. To address this, SFL recommends approaches that seek to automatically link to farm management data systems, where they are used (e.g. data on field-level fertilizer usage). This data integration is likely to be a focus of innovation in the coming years.

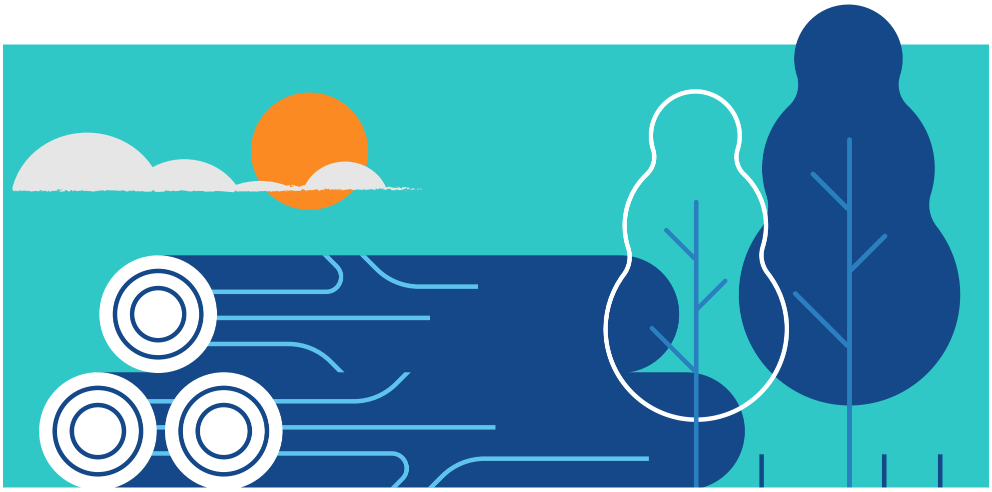
Overall, SFL see a significant amount of consensus on the IPCC-derived methods for quantifying emissions, however there is more diverging opinions on methods for the quantification of carbon dioxide removals from the atmosphere (e.g. into soils and trees). The non-profit sees carbon dioxide removals as a critical part of future climate solutions and so highlights them as an important area for organisations – including voluntary standards - to engage with as they develop climate change metrics. This is particularly the case for organisations in land-based sectors, which are unique in offering the potential to remove carbon dioxide through natural processes.

Finally, as with many sustainability metrics, the SFL has identified it is critical for producers to get value from any metrics process – rather than simply being the subject of data requests from customers, NGOs or standards organisations. Impact is most likely where producers feel engaged in the process, get access to results, and are enabled to explore 'what if' scenarios. This has been most successfully achieved through using the GWP metric as part of a broader programme of support on better farming practices, productivity, etc.

4.4 Summary

Given that climate change is an issue that has been given significant global scientific and policy attention, a large amount of research has been undertaken and published on climate change metrics. The space is dominated by a single measure, Global Warming Potential, upon which numerous metrics are based. It is likely that this will continue to be the default metric for the foreseeable future, given the degree to which it is embedded into international treaties, national regulations and long-established voluntary reporting frameworks. However, there is a growing consensus on the need for organisations to prepare for a transition to a ‘Net Zero’ world through aggressive science-based reductions in emissions, particularly over those which they have direct control.

With the debate on GHG emissions focused on reducing GWP to zero, measurement needs to take this into account (i.e. set the performance threshold for what is sufficient reduction and by when). In that sense, it can be argued that the key issue is that organisations obtain data that allows them to deliver meaningful reductions and, given the wide variety of metrics based on GWP, it is perhaps less important precisely which metric is used.



5 Deforestation

5.1 Background: measuring deforestation

Although deforestation is understood to mean the conversion of forest to other land uses, there are numerous qualifications to this broad definition that are used in the measurement of forest area, and hence in measuring deforestation.

The first set of issues centres around what constitutes a ‘forest’. The FAO – who have the responsibility of reporting on global forest area – define forest as ‘*Land spanning more than 0.5 hectares with trees higher than 5 meters and a canopy cover of more than 10 percent, or trees able to reach these thresholds in situ.*’[27] The term includes forests used for purposes of production, protection, multiple-use or conservation, as well as forest stands on agricultural lands (e.g., windbreaks and shelterbelts of trees), rubberwood plantations and cork oak stands. The definition specifically excludes stands of trees established primarily for agricultural production, for example fruit tree plantations and trees planted in agroforestry systems. Under this definition, deforestation is therefore defined as the permanent reduction of the tree canopy cover below the minimum threshold of 10 percent.

This definition of deforestation is also adopted by the Convention on Biological Diversity and is widely used in deforestation metrics. For example, Global Forest Watch have filters that allow the user to select a 10 percent threshold.[28]

The second set of issues pertain to what forest should be counted when deforestation is being measured. The FAO report a variety of deforestation forest statistics globally, including all forest, production forest, and natural forest only (see Table 5). In this sense, natural forest refers to forest where the trees are native (i.e., occurring within their natural range) and are regenerating without being planted by humans and therefore excludes tree plantations. Other categories of forest that are used to define deforestation metrics include High Conservation Value Forest (HCVF), High Carbon Stock Forest (HCS), protected forest and primary forest. These categories are somewhat overlapping but are essentially designed to identify the forests that are considered to be the most important by some organisations: HCVF for the biological, ecosystem service, social and cultural values they contain, HCS for the high amounts of carbon they store and primary forest because of their (relatively) pristine state. Metrics of deforestation that only include these categories do not report the conversion of other forest categories.

A third qualification used by some deforestation definitions concerns the way in which land is classified rather than the predominant vegetation type. In many countries, some areas of land that are forested may be legally assigned to other land uses (e.g., agriculture, plantation crops) and hence deforestation of those areas is legal. For this reason, some definitions of deforestation distinguish between legal deforestation and illegal deforestation (e.g., in Cargill's proprietary 'Triple S' soy standard[29]).

Finally, deforestation can be measured as a net or a gross metric. Gross deforestation measures the loss of forest that was occurred since a specified reference time. Net deforestation also takes into account the succession of non-forest to forest, and reforestation instigated by human activities, which may replace some or all of the forest converted. However, reforested areas rarely contain the same biodiversity as natural forest, and hence are often not directly comparable. For this reason, definitions of net deforestation often include caveats prohibiting the conversion of primary forest, HCVF and HCS.[30]

A set of common definitions, norms, and guidelines for delivering on ethical supply chain commitments, and in particular deforestation, has recently been launched by the Accountability Framework Initiative (AFi).[31] The AFi was developed by a coalition of conservation and human rights NGOs.

5.2 Overview of the metrics assessed

The metrics included here vary in their applicability and key trade-offs. A detailed description of the characteristics of each metric is given in the accompanying spreadsheet, which is summarised in Table 6.

We assessed thirteen metrics for deforestation (see Table 5). These include eight outcome metrics, three output metrics and two input metrics (Table 6).

Five of the metrics are used by ISEAL members, three are suggested adaptations of data that some VSS already collect and five are from external organisations (noting that there are numerous organisations offering satellite-based deforestation measurement, which are essentially similar and so only one is included as an example).

A notable trend compared with some of the other sustainability issues covered in this report is that a high proportion of the metrics are based on remote sensing; at least nine of the metrics assessed rely wholly or partly on analysis of satellite imagery.

Note that we include metrics of forest area in this assessment, because if the rate of forest loss within certified land management units is lower than in non-certified units, forest area can be used as a metric of avoided deforestation (i.e., with the assumption that forest would otherwise have been lost at the same rate everywhere).

Table 5:

Description of deforestation metrics

Metric	Metric description
Tree cover loss	Tree cover loss from key sourcing countries, or jurisdictions using high quality publicly available Earth Observation data
Natural forest loss	Amount (or proportion) of change in natural forest area. Reported nationally.
Producer-level change in area of High Conservation Value Forest (HCVF)	High Conservation Value areas are defined nationally and several standards (FCS, RSPO, Bonsucro, etc) include criteria that they are identified and maintained. This is verified by audits. The area can be aggregated across all certified entities (by type of HCV if required) to produce a global estimate of the area of HCV maintained and/or deforested. Differs from the previous in that the metric would be calculated as the change in area of individual certified producers otherwise an increase/decrease in the number of certificate holders would cause the figure to increase/decrease.
Producer-level change in area of natural forest	The area of natural forest certified (FSC) could potentially be analysed as the change in area for each certified management unit. This would show the outcome of deforestation, if compared with a baseline such as the FAO change in national area of natural forest (above)
Producer-level change in area under conservation management	The area 'set aside' for conservation (FSC) or actively managed for biodiversity (SAN-RA). Can also be recorded by location and dominant vegetation type. The latter is most appropriate to the issue of deforestation, as not all set-aside is forest. Differs from the previous in that the metric would be calculated as the change in area of individual certified producers otherwise an increase/decrease in the number of certificate holders would cause the figure to increase/decrease.
Land Use Change Analysis (LuCA)	The RSPO requires certified land managers to disclose areas cleared since November 2005 without prior HCV assessment (in 2014). All companies in control of areas with non-compliant land clearance are then required to submit a Land Use Change analysis (by September 2014). Land use/cover in November 2005 is used as a proxy for potential High Conservation Value Forests that may have been lost due to clearance without prior HCV assessment. Different (pre-clearance) land uses are given different weightings, the highest for structurally intact forest, and a zero weighting for perennial monocultures.
Rate of ecosystem destruction or restoration compared to surrounding areas	The metric looks at canopy cover, number of trees, number of tree species and structural complexity based on satellite imagery. Complex statistical analysis is then used to compare certified and non-certified producers within the same area.
Maximum deforestation	Calculates potential deforestation caused by commodity production at the level of municipality based on a combination of official land-use statistics and satellite imagery. Referred to as 'maximum deforestation' because if, for example, if only 10 ha of soy were planted, and 50 ha of deforestation occurred, then the maximum deforestation attributable to soy is 10 ha. Up-to-date analysis on production and trade flows for a small number of countries and commodities, and deforestation estimate only currently available for soy and beef from Brazil.

Table 5 cont:

Description of deforestation metrics

Metric	Metric description
Area of High Conservation Value Forest (HCVF)	High Conservation Value areas are defined nationally and several standards (FCS, RSPO, Bonsucro, etc) include criteria that they are identified and maintained. This is verified by audits. The area can be aggregated across all certified entities (by type of HCV if required) to produce a global estimate of the area of HCV maintained
Area of natural forests in certified operations	Area of managed natural forest in certified operations. Assumes that. Natural forest is not degraded by management.
Area under conservation management ('set aside')	The area 'set aside' for conservation (FSC) or actively managed for biodiversity (SAN-RA). Can also be recorded by location and dominant vegetation type. The latter is most appropriate to the issue of deforestation, as not all set-aside is forest.
Supply chain zero deforestation commitments	The number (or proportion) of certified supply chain companies that have zero deforestation commitments
Deforestation Risk Index (DRI)	The Deforestation Risk Index (DRI) uses earth observations and local economic indicators to predict what areas are more at risk of being deforested in the next year. Could potentially be used to show the difference in deforestation risk around certified farms/farmer groups and non-certified.

5.3 Applicability and trade-offs

5.3.1 Key technical decision points in measuring deforestation

Which definition of deforestation to use?

As discussed above, the definitions of forest and deforestation underpinning deforestation metrics include; all forest, subsets of forest (e.g., natural forest, HCVF), the mode of deforestation (legal or illegal) and whether net or gross deforestation is measured. The definition of deforestation chosen by an organisation will profoundly affect the area of forest loss they report and the value of the baseline they report against. For example, HCVF is only a small fraction of all forests globally, whereas approximately 93% of the world's forests were natural forests in 2015.[32]

The choice of underpinning definition is subject to numerous trade-offs, including:

- Organisational policy, ethos, and stakeholder expectation: For example, an organisation working in sectors in which forests are a marginal concern, may favour an underpinning definition that measures the forests deemed to be most important. Whereas one that promotes itself as a solution to deforestation may prefer to emphasise the contribution it makes to conserving forests of any type. For example, SAN-RA report on the rate of ecosystem degradation/restoration in certified operations compared with nearby non-certified areas.
- Alignment with the organisation's principles and criteria: For example, where a VSS uses HCVF as a forest conservation instrument within its principles and criteria, then a definition of deforestation that focuses on HCVF might be considered appropriate. E.g. the FSC report on the area of HCVF within certified operations.

- Additional reasons for reporting on deforestation: For example, an interest in GHG emissions related to forests, may suggest the use of a metric based on a definition of net deforestation of all forest types, whereas an interest in biodiversity could use a definition based on gross deforestation of protected areas, HCVF and/or natural forest.

Whether to measure outcome, output or input metrics?
As with all metrics, a choice has to be made about whether to measure an outcome, an output or an input. Part of that decision is about what public claim a VSS wants to be able to make around deforestation. For example, to be able to claim that a certification scheme results in less deforestation than uncertified production, as a minimum, an outcome metric is required that also has a baseline with which to compare the performance of certified producers.[33]

Of the main outcome metrics assessed, there is a clear division between those that have an external baseline that allows comparison of the performance of VSS against national or more spatially granular deforestation, and those that do not (Table 6).

The main metrics with external baselines include:
Tree cover loss. The specific example given here is 'Global Forest Watch Pro',[34] however this is just one amongst many examples of deforestation surveillance systems based on the analysis of satellite imagery. There are some differences between these (e.g., in terms of spatial resolution, the analytical algorithms, user interface, cost) but in essence they are broadly similar. Global Forest Watch is used as an example, because the underlying data set used is now the principle data used by academics studying deforestation and has therefore been subject to rigorous assessment. In essence, Global Forest Watch uses automated analysis of satellite data with global coverage and reports on an annual basis. Each 'pixel' (a 30 x 30 m area) is classified as having tree cover or no tree cover (with different thresholds available for defining tree cover, ranging from 10% in line with the FAO forest definition, up to 50%). The change of a 'pixel' from 'tree cover' to 'no tree cover' is then counted as 'tree cover loss'. Like all remote-sensing measurement of deforestation, there are certain limitations: false positives may occur with plantations or agroforestry, and false negatives can potentially occur through large patches of natural disturbance or through clear felling in forestry systems where that is the norm (e.g., most boreal forests). The 'pixel' size also limits its application in certain contexts: it is too coarse grain to define field or farm boundaries, which limits its use as a compliance tool for VSS.

A related metric that is used by SAN-RA in a recent impact report is the **rate of ecosystem destruction or restoration** compared to surrounding areas.[35] This metric measures canopy cover, number of trees, number of tree species and structural complexity based on analysis of satellite imagery. Statistical analysis is then used to compare certified and non-certified producers within the same area, providing a bespoke and context-specific baseline. By including factors in addition to forest area, the metric also indicates the quality of ecosystem and so, unlike the other metrics here, provides a measure of forest degradation as well as deforestation. Degradation of forests is of considerable importance for GHG emissions and biodiversity, and indeed the UNFCCC has been criticised for not taking it into account in its definition of forests.[36] The metric could presumably relatively easily be reported just for forest ecosystems.

*The **change in natural forest area*** (FAO) is a statistic reported by national governments, which is collated and reported publicly every five years in the FAO's Global Forest Resource Assessment.[37] Although the FAO has measurement protocols for national governments, some choose not to report, whilst the data provided by others is at significant odds with independent assessments. The spatial scale of reporting (national) and infrequency with which it is reported mean that it is probably only useful to VSS as the basis of an external baseline against which to gauge progress on natural forest area retention or loss (Table 6).

The maximum deforestation metric used by the TRASE platform[38] is designed for sectoral use, in that it attributes deforestation to other specific land uses (e.g., to specific crops). The metric reports the potential deforestation caused by commodity production at the level of municipality based on a combination of official land-use statistics and satellite imagery. It is referred to as 'maximum deforestation' because if, for example, if only 10 hectares of a crop were planted, and 50 hectares of deforestation had occurred, then the maximum possible deforestation attributable to the crop is 10 ha (i.e., all of the crop was grown on deforested land). Currently, the maximum deforestation estimate is only available for soy and beef from Brazil on the TRASE platform (Table 6).

The above metrics include (or provide) an external baseline against which to compare the deforestation. The remaining outcome metrics rely on internal baselines: data collected repeatedly from a specific cite, or from selected comparison sites in the same area. **The Land Use Change Analysis** (LUCA)[39] has been used by the RSPO Ito understand likely incidence of conversion of HCVF without prior HCV assessment. There is no external baseline for this metric as the potential HCVF considered is all within certified operational areas. All RSPO certified companies in control of areas with non-compliant land clearance were required to submit a Land Use Change analysis in 2014. Land use/cover in November 2005 is used as a proxy for potential High Conservation Value Forests that may have been lost due to clearance without prior HCV assessment. Different (pre-clearance) land uses were given different weightings, the highest for structurally intact forest, and a zero weighting for perennial monocultures. Although this was apparently used as a 'one-off' exercise by the RSPO, in principle it could be repeated by other VSS to measure historical deforestation, with the focus on HCVF, or all forest, as required (Table 6).

Other than those examples cited above, ISEAL members appear to predominantly use output metrics relating to deforestation and its flip side, forest retention and/or reforestation. For example, the **area of natural forests in certified operations** is reported by FSC.[40] As there is no external baseline, and no notion of how the value changes on a producer level, this metric is currently used to describe the scale of FSC certification of natural forest. However, it could potentially be converted to an outcome metric (by FSC or other VSS) of deforestation (or avoided deforestation) by measuring the change in area of natural forest per Land Management Unit,[41] with the FAO natural forest area metric providing a baseline (Table 6).

Similarly, Bonsucro,[42] the RSPO[43] and the FSC report on the **area of High Conservation Value Forest** (HCVF) within certified operations. As described above, HCVF is a concept that defines the most biologically, ecologically, socially and culturally important forests. Hence the area of HCVF indicates the extent to which the VSS's clients are 'looking after' important forest areas, but not their condition, nor whether clients are maintaining the area. Moreover, the amount of HCVF is likely to increase as a scheme expands its number of clients, making it problematic to use as a metric of avoided deforestation. What constitutes HCVF is defined nationally in a way that is consistent with the global definition, but this has not been done in all countries, and even where it has, some parts of the definition do not lend themselves to large-scale mapping. As with the area of natural forest discussed above, a derived metric is suggested here, using the same client data, that measures the **change in area of HCVF per land management unit**. Whilst this metric still lacks an external baseline, it would not necessarily increase as a scheme expands, and would at least partially indicate the effectiveness of the VSS in maintaining High Conservation Values. However, in most cases it would be unusual for the area of HCVF to contract or expand significantly in a single management unit, and hence this metric may not be responsive enough to change to meet an organisation's ends (Table 6).

Table 6.

Characteristics of deforestation metrics

Scale of application: ⏩ Supply chain; ⚙ Producer; ⚑ National; ⛰ Landscape/ Jurisdiction

◎ High ◉ Medium ● Low

Metric type	Metric description	Scale of application	Considerations around sustainability definition				Considerations around performance monitoring					Considerations around data use			
			Metric measures the sustainability issue	Metric is underpinned by evidence	Metric addresses material impacts	Metric is widely used and easily communicable	Practical to collect	Practical to implement	Context-sensitive	Responsive to change	Scalable	Baseline data available to provide context	New tools/software emerging?	Implementation 'user package' exists	Main issues/comments
Outcome	Tree cover loss	⛰	◉	◎	◉	◎	◎	◉	◉	◎	◉	✓	✓	✓	A rapidly developing approach, with new products becoming available regularly. Perhaps the biggest drawback with existing tools (or at least the affordable ones) is the limited ability to define land parcels (e.g., farms, FMUs) which makes its application better suited to larger scale assessment rather than 'producer up'. Data is updated annually, meaning that real time interventions are not possible (though competing platforms will emerge within the next 2-3 years that may resolve this).
	Natural forest loss	⚑	◉	◉	◉	◎	◎	◎	◉	●	●	✓	✗	✗	A well established measure of deforestation, focused on the subset of natural forest that is often the most critical forest for biodiversity. However, the key limitation is that it is only reported nationally, therefore can not reflect producer-level outcomes.
	Producer-level change in area of High Conservation Value Forest (HCVF)	⚙	◉	◉	◉	◉	◉	◉	◎	◉	◎	✗	✓	✓	Has the advantage of being a concept that is already implemented in several VSS, including reporting. However, suffers format least two main issues as a deforestation metric: (a) Is only a subset of forests (albeit some of the most ecologically and socially important); (b) Would only monitor deforestation if changes to the area of each producer were compared year on year. As the majority of certified producers will protect their entire HCV area, change would be minimal.
	Producer-level change in area of natural forest	⚙	◉	◉	◉	●		◉	◎	●	◎	✓	✗	✗	Has the advantage of being a concept that is already implemented in several VSS, including reporting. However, unlikely to be responsive enough to reflect changes over short time periods.
	Producer-level change in area under conservation management	⚙	◉	◎	◎	◉	◎	◉	◎	●	◎	✗	✗	✗	Has the advantage of being a concept that is already implemented in several VSS, including reporting. However, as the majority of certified producers will protect their set aside area, change would be minimal.
	Land Use Change Analysis (LuCA)	⚙	◉	◎	◎	◉	◉	◉	◎	●	◎	✗	✗	✗	Used by the RSPO as a 'one-off' exercise to determine the conservation debt of companies that may have deforested HCVFs without conducting prior assessment. Unlikely to be repeated, and not designed as a monitoring tool.
	Rate of ecosystem destruction or restoration compared to surrounding areas	⚙	◉	◎	◎	◎	●	●	◎	◉	◉	✗	✓	✓	At the current time, this is a rigorous but expensive way of determining the impact of certification on deforestation, possible only as occasional and discrete research exercises. However, the advance in earth observation sciences means that it is likely that this type of analysis can become semi-automated within the next. 2-5 years, potentially opening it up as a more accessible, practical tool for certification monitoring and reporting.
	Maximum deforestation	⛰	◉	◉	◉	◎	◉	◎	◎	◎	◎	✓	✓	✓	TRASE allows specific flows of commodities to be linked to deforestation. However it is currently severely limited as it only covers soy and beef from Brazil.
Output	Area of High Conservation Value Forest (HCVF)	⚙	◉	◉	●	◉	◉	◉	◎	●	◎	✗	✓	✓	Has the advantage of being a concept that is already implemented in several VSS, including reporting. However, a change in area would reflect the increase or decrease in number of producers rather than the retention or deforestation of HCVs.
	Area of natural forests in certified operations	⚙	●	◉	●	◉	◎	◎	◎	●	◎	✗	✗	✗	Metric measures the reach of FSC into natural forest management, but does not measure deforestation and will not be applicable to VSS that are not certifying forest management practices.
	Area under conservation management ('set aside')	⚙	◉	●	◎	◎	◎	◉	◎	●	◎	✗	✗	✗	Has the advantage of being a concept that is already implemented in several VSS, including reporting. However, a change in area would reflect the increase or decrease in number of producers rather than the retention or loss of forest.
Input	Supply chain zero deforestation commitments	⏩	●	●	●	◎	◎	◎	◉	◉	●	✗	✗	✗	Cheap to collect and implement and would potentially be an effective descriptor for the global operations of a VSS. However, the metric suffers from not reflecting the efforts of certified producers, and from being a process indication with an unreliable link to positive change in deforestation.
	Deforestation Risk Index (DRI)	⛰	●	◉	●	◉	●	◉	◎	◎	◎	✗	✓	✓	Could potentially be used to show the difference in deforestation risk around certified farms/ farmer groups and non-certified. However, is likely to be prohibitively expensive, seems to only have been used in Côte d'Ivoire, and ultimately may be more useful as a compliance approach.

A broader metric, **the area under conservation management** ('set aside'), which takes into account forest and non-forest ecosystems that are managed principally for conservation, is used by FSC[44] and SAN-RA.[45] With minor modification to the data collection, this could be reported as forest under conservation. This metric has no external baseline and will tend to increase as the coverage of the scheme increases. Therefore, as for the preceding metrics, a derived metric of the change in area under conservation management per land management unit is proposed as a possible outcome metric (Table 6).

Two further output metrics as assessed here. The first is **Deforestation Risk Index** (DRI) produced by Vivid Economics.[46] This is one example of a number of deforestation risk metrics. The DRI uses a combination of earth observations and local economic indicators to predict what areas are more at risk of being deforested in the next year. Deforestation risk metrics have two potential uses for VSS. The first is in demonstrating the reduced risk of deforestation in certified compared with non-certified operations. However, it is unclear whether the DRI or other similar risk metrics operate at a sufficiently granular scale of information to incorporate individual certification. The second use is to increase the sophistication of their certification system, such as using risk analysis to select a sample of producers to be audited, to identify emerging sustainability issues and in the quality control of their certification bodies and auditors.[48]

Finally, several companies report on the **number (or proportion) of companies in their supply chain that have zero deforestation commitments**. This is a convenient measure for companies with large, complex supply chains, but has the disadvantage that commitments do not necessarily result in successful action. However, it could potentially be used by VSS to report on their supply chain clients, as a way of demonstrating the wider commitment to sustainability of the certification system.

5.3.1.1 Key practical decisions in measuring deforestation

Cost vs robustness

Unlike some of the other sustainability issues covered in this report, the primary data required to measure deforestation generally comes from one or both of two sources: remote sensing (data which is often free) and client data that would typically be collected during audits and reported to the VSS. Both data sources are relatively cheap compared with the on-the-ground surveys that other metrics, e.g. poverty and biodiversity require (the exception being proprietary deforestation risk metrics, which typically contain significant intellectual property and may be intended and priced for large corporations). This may make outcome metrics more accessible to VSS for deforestation than for some other sustainability issues (Table 6).

Value for the VSS and client

Land managers are typically very well aware of deforestation that is happening in their immediate neighbourhood. They may be less aware of deforestation happening in the wider landscape or nationally. There may therefore be some value in placing the efforts of producers to avoid deforestation into a broader context. For VSS, having jurisdictional or national deforestation rates against which to compare the performance of the certification scheme allows reporting of the scheme's impact on deforestation. This suggests that the greatest mutual value might lie in those metrics that include an external baseline (e.g., tree cover loss,

Box 2.

Implementation of deforestation metric: Olam's Forest Loss Risk Index.

Olam are one of the world's largest traders in agricultural commodities, including several that can be associated with deforestation, such as palm oil, coffee and cocoa. Their supply chains are often extremely large and complex, with many thousands of suppliers. As well as some of their own plantations, Olam source from over five million smallholders, sometimes directly, sometimes via traders. In coffee alone, they operate in 20 countries with 154 buying stations, each supplied by multiple producers, and in cocoa they operate in ten countries and use in excess of 18,0000 buying stations.

Ensuring that their supply chains are not contributing towards deforestation therefore represents a significant challenge to Olam. The approach they have taken is to develop a proprietary deforestation-risk assessment tool (the Forest Loss Risk Index), the key requirement of which is that it must allow Olam to prioritise the sourcing areas that they need to focus further attention on.

The key underlying information sources for the deforestation risk tool is the Global Forest Watch (GFW) tree cover and tree cover loss data. The risk index is inferred from the timing and extent of recent deforestation, and remaining forest cover in the area around the sourcing location. The reasons for using GFW data are articulated in a White Paper as "*The World Resources Institute's Global Forest Watch (GFW) is currently the best tool to gather up-to-date and comprehensive data on forests, tree cover loss and land use change. GFW facilitates scanning a large number of sourcing areas and farms to identify deforestation activities and is the platform relied upon for developing this method.*"[49] GFW tree cover loss data is also independent, widely used, and publicly available, which provides an additional layer of credibility for Olam.

Following the risk analysis, implementation proceeds through four further phases: prioritising sourcing locations, investigation, action, and as a last resort, eliminating the supplier from the supply chain. In coffee, Olam were able to prioritise 24 of their 154 sourcing locations for further investigation. Investigation proceeded through compiling detailed information on farmer-supplier locations to establish whether the supply is indeed coming from the areas of deforestation. Where that is the case, Olam makes clear its expectations to farmers, and also supports them to improve their practices through its farmer-support platform.

Buy-in for the approach amongst senior staff at Olam was relatively straightforward, as a commitment toward zero deforestation (formally, excluding deforestation of HCVF, HCS and protected areas) is part of the company's 'Living Landscapes Policy' – a framework that supports a 'net positive' approach to agricultural supply chains and which has the backing of senior management.[50] More challenging was working with agents on the ground to understand that the approach required evidence that the supply is not linked to deforestation, rather than an absence of evidence that it was. Deforestation has not previously been 'on the radar' of agents, whose role is to buy and sell. The Olam CRS team invested significant effort in spreading the 'Living Landscapes' ethos, and more agents are beginning to understand it. Critically, being able to present a deforestation risk analysis to agents, showing where the deforestation is, allows the conversation to move on to action: 'let's work together to show that it isn't our supply chain [contributing to deforestation]'. Furthermore, being able to prioritise a few locations for action makes the task more manageable for agents and CSR staff alike.

Box 2 cont.

Olam report that they have had little push-back on their approach from external stakeholders. This is partly because it is based on credible data and partly because they took the step of publishing the approach in the White Paper. Some comments have been made that GFW data can overestimate deforestation (e.g., in rotationally felled plantations) and that road infrastructure, timber harvesting and other factors also drive deforestation, not just agriculture. However, that is why the on-the-ground investigation phase is critical, and Olam argue that it is far more cost-effective to prioritise and then investigate risk areas, than to develop a deforestation monitoring tool that could distinguish between these factors.

Olam do not consider that the effort required to develop and implement the approach was onerous: the approach was developed with a combination of in-house expertise and consultant resources. The analysis is being repeated each year, is reported publicly in their annual sustainability report, and is being brought fully in-house as a part of one person's job. The next step will be to automate the analysis, to reduce costs further.

And the biggest lessons learned? Firstly, the importance of prioritisation. Secondly, that developing and implementing the tool helped them understand the concept of risk-based approaches: high risk doesn't mean that there is definitely a problem, and there is no substitute for being on the ground and knowing where your product is coming from.

5.3.2 Future horizons

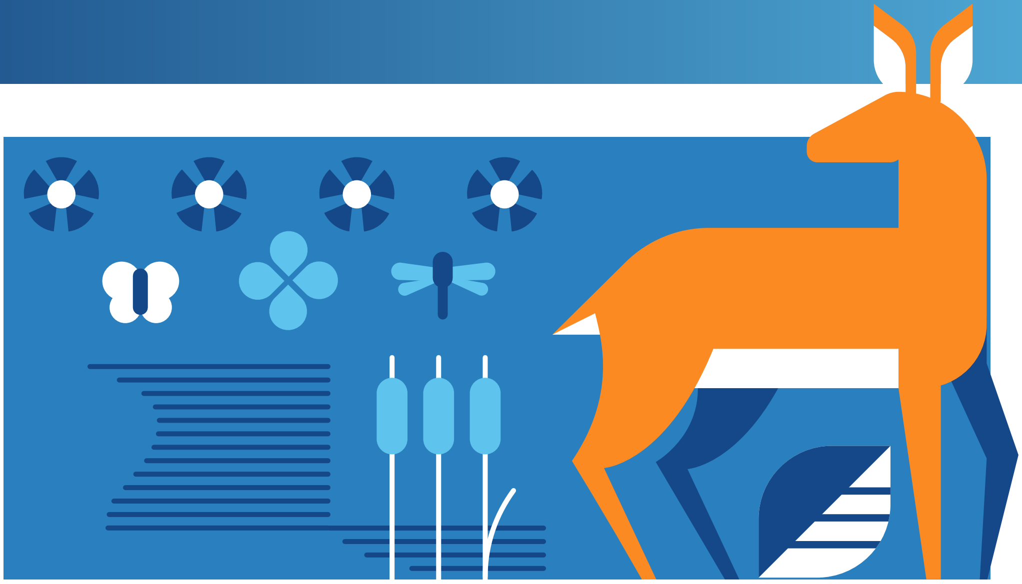
Two main trends have emerged within the last few years in terms of deforestation metrics. The first, represented by the discussion of Global Forest Watch's tree cover loss metric, is the increasing sophistication, credibility and availability of data derived from remote sensing. Analysis of deforestation by remote sensing is likely to become increasingly spatially granular, better able to distinguish between different tree-dominated land use (e.g., natural forest vs agroforestry) and provide near real time updates. These trends will make it increasingly straightforward to monitor compliance as well as report outcomes. Secondly, as it becomes cheaper to overlay multiple geospatial data sets, the type of deforestation risk assessment exemplified by the Deforestation Risk Index (and also in Box 2) are likely to become increasingly accurate in prioritising where deforestation is likely to occur before it has actually happened. This opens up greater opportunities for taking preventative actions. The better the geospatial information that a VSS has on its certified clients, the more meaningful the results of remote sensing metrics will be.

The second trend is a broadening out from the concept of deforestation to encompass the conversion of any natural and semi-natural habitat, whether that is dominated by trees or not. The revision on the Roundtable on Sustainable Soy standard in 2016 was the first multistakeholder sustainability standard to in effect move from 'zero deforestation' to 'zero conversion'.[51] The drive for zero conversion is codified within the Accountability Framework Initiative[52] and has recently been used in a high-profile assessment of corporate environmental performance.[53] Standards organisations may wish to begin incorporating conversion metrics into their effectiveness measurement and reporting systems in advance of likely revisions of their standards.

5.4 Summary

Deforestation has risen rapidly up corporate and policy sustainability agendas in recent years, primarily because of the role of land use change in greenhouse gas emissions. This has been accompanied by a plethora of metrics, largely based on remote sensing. Remote sensing is a highly specialised and technical process, but there are numerous external service providers who collect, analyse and report on deforestation, and high-quality data is available for free. It therefore notable that we could find only two deforestation (or avoided deforestation) metrics based on remote sensing data that is used by ISEAL members: the **rate of ecosystem destruction or restoration** compared to surrounding areas (SAN-RA)[54] and **Land Use Change Analysis** (RSPO).[55] The increasing ease of acquiring data and conducting bespoke analysis means that there will be opportunities for VSS to report (avoided) deforestation outcomes more frequently and globally rather than for specific impact evaluations. As remote sensing of deforestation is evolving rapidly, it is almost certain it will be possible to assess deforestation within individual land management units affordably and in real time within the next two to three years. This opens up increasing possibilities for compliance monitoring.

However, there are multiple definitions of forest and deforestation, and organisations that decide to measure (avoided) deforestation of a very specific subset of forest (e.g., HCVF) may be unable to find suitable global baselines against which to compare the performance of their clients. The current metrics for these definitions are based on the area of forest certified, and hence are likely to increase as the overall area certified increases. As a consequence, they do not relate closely to (avoided) deforestation. Several ‘work around’ metrics are suggested to convert the existing output metrics used by a number of ISEAL members into more outcome-focused metrics. However, they are less responsive to change than remote sensing-based metrics.



6 Biodiversity

6.1 Background: measuring biodiversity

The concept of biodiversity is complex and as a result, reliable measurement can be challenging. Nevertheless, there is increasing pressure and requirements to capture changes in biodiversity and to demonstrate the effectiveness of actions to help halt its loss. Biodiversity underpins ecosystem services which are critical to many functions of business and society. For this reason, in addition to the inherent value of biodiversity, effective conservation is critical.

Biodiversity refers to the variety of living organisms in a system, including animals, birds, insects, plants, fungi and micro-organisms. It is often defined on three levels; i) genetic diversity – the range of genes and characteristics within a species, ii) species diversity – the variety of organisms within a habitat or area, and iii) ecosystem diversity – the diversity of species between different habitats or areas.[56]

Species diversity is one of the most widely understood and commonly measured dimensions of biodiversity, particularly in relation to biological conservation.[57] A basic measure of this is species richness; a count of the number of different species in a system. However, this is seen by some as overly simplistic, and other metrics also incorporate the abundance of the different species, using equations to combine this with richness and calculate a single overall value for biodiversity. This takes into account the evenness of the spread of abundance between species, which has important implications for ecological interactions in the community and for the resilience of biodiversity to external pressures.

In addition to these direct measures, biodiversity is also often captured using related concepts. For example, the presence of certain indicator species is sometimes taken to be a proxy for broader biodiversity. Functional diversity - which describes the various roles and behaviours of organisms within an ecological community e.g. grazing, predation, nutrient cycling – is also sometimes used as a measure of biodiversity and there is some evidence that it correlates well with species richness[58]. Habitat diversity, which encompasses the number of different habitat types and their structural diversity, is another metric used to indicate biodiversity[59].

There are a range of metrics which measure these various dimensions of biodiversity. Many more recent metrics focus on using collated data from the existing body of ecological research. This reduces the need for detailed primary data collection and, given the global scope of the existing research, increases the scale at which biodiversity can be measured and compared. Such metrics are often focused specifically on changes in biodiversity. Accurate measurement and reporting on such dynamics is required by international conventions such as the Aichi Biodiversity Targets[60] which requires actors to "*...take effective* *and urgent action to halt the loss of biodiversity*”. Planning the implementation of suitable actions and monitoring their impacts requires metrics which can measure changes in biodiversity at an appropriate scale.

Table 7:

Description of biodiversity metrics

Metric	Metric description
Biodiversity Impact Metric	Metric which models the biodiversity impact of a company's activities. This is calculated using data about the i) area, ii) intensity and iii) location of the land footprint of the company's activities. The impact of this land-use on biodiversity is then modelled using the Biodiversity Intactness Index (see below), plus consideration of the biodiversity 'quality' in the location in question.
GLOBIO	GLOBIO assesses the impact of environmental drivers on terrestrial biodiversity, measured as Mean Species Abundance (MSA) before and after disturbance. Disturbances include land use change, nitrogen deposition, infrastructure, fragmentation and climate change. MSA is the mean abundance of original species in a system after disturbance relative to their abundance in undisturbed ecosystems; it is an indicator of naturalness or biodiversity intactness, from 0% to 100%. MSA and the impact of disturbance on MSA are calculated based on a meta-analysis of hundreds of peer-reviewed field-based studies. Assessments can be on a national, regional or global scale.
Species count/ species richness	Count and record of the different species present on site.
Global Biodiversity Score (GBS)	Global Biodiversity Score is a tool primarily assessing the biodiversity footprint of companies or investments. Expressed in MSA per km2 where MSA is Mean Species Abundance representing the current intactness of the ecosystem or biodiversity compared to that of a pristine state. MSA values range from 0% to 100%, where 100% represents an undisturbed pristine ecosystem. The MSA per km2 is used to give a spatial footprint of biodiversity impact.
Shannon Index of Diversity	Metric which measures the number of different species and the proportional abundance of each species to give a measure of biodiversity which incorporates i) species richness, ii) population of each species and iii) evenness of species within a community. Calculated by H = - Σpiln(pi) where pi is the proportion of the total number of species in a community made up by species i and ln is the natural logarithm of this proportion.
Biodiversity Intactness Index (BII)	Measures ecosystem or biodiversity intactness based on the current presence and population of species compared to the estimated population that would exist in an area of undisturbed pristine habitat, calculated based on contemporary populations in relatively undisturbed areas of the habitat e.g. protected areas.
Living Planet Index (LPI)	Trends in abundance of species, calculated using over 14,000 population time-series from sources including journals and online databases. Metric is designed for global assessment of biodiversity but can also be applied at regional or national level.

Table 7 cont:

Description of biodiversity metrics

Metric	Metric description
Record of key species	Record of a presence/absence of important species in an area. Species can be defined as 'important' according to different criteria such as local/national/international rarity, conservation status, ecosystem service (e.g. pollination) or as indicator of habitat quality. For example, the LEAF Marque Landscape and Nature Conservation Audit/Enhancement Plan requires listing of key species on farms. UTZ code of conduct requires identification (and protection) of endangered species. FSC HCV metric requires biodiversity to be measured based on indicator species, keystone species or collections ('guilds') of species associated with large intact ecosystems.
Existence of conservation management plans	Record of number or proportion of producers that have an active conservation management plan. Can be aggregated at a company or supply-chain level.
Area of natural or native vegetation cover	Area of vegetation coverage and % of native vegetation species, presented in hectares and/or as % of land area.
Retention of forest or other natural habitat	Area and parameters of forest or other natural habitat on production site mapped and monitored periodically to check for change in area or boundaries. Areas should be maintained. Can be part of baseline mapping as for Rainforest Alliance farm baseline assessment.
Cool Farm Tool Biodiversity	Cool Farm Tool biodiversity metric calculates a score for on-farm actions to enhance biodiversity, such as field margin management, provision of hedgerows etc. Each potential action receives a score to reflect its contribution to enhancing biodiversity, weighted based on expert assessment of published evidence on the likelihood of the effectiveness of each action to enhance biodiversity. Currently only for temperate systems.

6.2 Overview of the metrics assessed

We assessed a total of twelve biodiversity metrics (Table 7). These include four input metrics, six outcome metrics and two impact metrics (Table 8).

Four broad categories of biodiversity metrics are included in the assessment:

- direct measures of biodiversity,
- proxy or ‘surrogate’ indicators of biodiversity,
- metrics that model biodiversity changes using inputs including collated data from existing ecological research, and
- metrics which capture the presence of management measures designed to enhance biodiversity.

6.3 Applicability and trade-offs

Table 8 provides a summary of the characteristics of biodiversity metrics, assessed against the considerations outlined in Section 3.2. The following section draws out some of the main issues identified from this analysis.

6.3.1 Key technical decision points in measuring biodiversity

Which aspect of biodiversity to measure?

One of the key technical choices regarding which biodiversity metric(s) to use concerns whether to measure biodiversity directly, use indicators of biodiversity, modelling approaches based on secondary data or existence of particular management approaches.

The main direct measures of species richness and biodiversity are well established in the literature, but both rely on on-site surveys and significant ecological expertise:

Species count (species richness). This metric captures the number of different species observed on a site and is a key dimension of biodiversity, as described above. Although the data is uncomplicated (simply a list of species encountered at a site) it does require expertise in ecological surveying and species identification. As a metric, it is a relatively basic indicator of biodiversity and only measures one dimension of biodiversity – the number of species. This crucially does not capture anything about the abundance of species, which is an important factor determining the health and resilience of ecological communities. It is therefore not sensitive to changes and there is evidence that significant changes in biodiversity are not captured by measuring species richness[61] (Table 8); most of the population of a species except one or two individuals could be lost without any change to the species count.

A more sophisticated measure of biodiversity commonly used in ecological research is the **Shannon Index of Diversity**. This incorporates both species richness and abundance. These values are combined using an equation to calculate a single value which represents the biodiversity at that particular site at that time. This is a more representative measure of biodiversity than simple species richness and therefore more robust and resilient to scrutiny compared to a species count. The biodiversity index value can be used to compare biodiversity between sites or, if updated data is collected, to measure changes in biodiversity for the same site over time. However, it entails more intensive primary data collection compared to a species count as it additionally requires measurement of the population of each species (Table 8).

The main proxy biodiversity measures at a site level include:

Enumeration of **key or indicator species**. Instead of requiring comprehensive data collection capturing most or all species in a site in order to calculate biodiversity, this instead involves recording the presence of certain 'key' species in an area. The definition of 'key species' could be based on a number of criteria, for example, factors that make them a species of particular interest or conservation priority such as endemism (whether a species is unique to the site or surrounding region) or global conservation status (e.g. according to the IUCN Red List of Threatened Species[62]). Other criteria may be that they are 'indicator species', only occurring in areas where habitat quality is high and therefore more likely to support higher biodiversity, or 'keystone species' which have a large influence on wider biodiversity relative to their population size. The species list could cover high functional diversity, species with a range of habitat requirements, and balanced coverage of all major taxonomic groups, in order to increase the breadth of biodiversity that is represented. Another option, used in the FSC HCV metric, is to look for collections or 'guilds' of species which are indicative of large, intact or high-quality habitats.

An advantage of measuring a subset of species – particularly those that are threatened and endangered – is that it focuses measurement on those elements of biodiversity that are of the greatest concern. However, caution needs to be applied to using indicator species, as there is ample evidence that their presence is often a weak indicator of other elements of biodiversity. In addition, rare species are almost by definition difficult to detect (they are rare!) – witness the number of apparently extinct species that have been 'rediscovered'.[63] Further, it limits the ability to scale up findings, as indicators will be specific to habitats, and therefore will vary within a country (depending on vegetation type, climate, altitude, etc) and even more so between countries (Table 8).

An alternative is to look at the coverage and quality of areas of habitat retained or conserved on a site. Two options included here are the **area of natural or native vegetation cover and the retention of forest or other natural habitat.** The first looks at the area of vegetation cover and prevalence of native vegetation species, reported as a percentage of the land area or in a measure of area such as hectares. This is relatively cost-effective as it can be reported by clients and verified by auditors as part of the normal certification process. However, habitat area by itself gives no indication of the quality of that habitat or the biodiversity it contains. The second metric is similar but brings in an aspect of measuring change over time compared to earlier habitat conditions, for example as in Rainforest Alliance farm baseline assessment. This relies on periodic repetition of data collection. The area of habitats can be estimated using satellite imaging data, if coverage exists at a suitable resolution for the site in question. Ground-truthing using on-site surveys will provide more accurate information and methods for estimating habitat quality exist that are rapid and do not require specialist skills.[64] Reference to resources such as the IUCN Green List of Protected and Conserved Areas or identified Key Biodiversity Areas[65] will also show whether the site in question is within, or close to, critical habitats which may inform management decisions[66].

The third group of metrics use meta-analysis and modelling based on existing data. These metrics are generally designed to measure changes in biodiversity and specifically the state of biodiversity after disturbance and are based on the collated findings from the huge body of existing biodiversity studies. This is the approach behind the **GLOBIO** metric. This measures biodiversity using Mean Species Abundance, which is the mean abundance of original species in a system after disturbance compared to their abundance in an undisturbed system. It is presented as percentage where 100% means the abundance of species is unchanged and 0% means none of the original species remain. This is modelled based on a meta-analysis of around 200 peer-reviewed field-based studies which measure biodiversity before and after disturbances including land use change, fragmentation and climate change[67]. Assessments can be made at a national, regional or global scale.

Another metric, the **Global Biodiversity Score**, combines MSA with inputs of area data to calculate a spatial biodiversity ‘footprint’ of company activities or scenarios for the biodiversity impact of policy or management measures. This metric is specifically designed for setting and measuring progress against biodiversity commitments of corporate and financial organisations and is one of the biodiversity metrics used by the Convention on Biological Diversity[68].

The **Biodiversity Intactness Index** (BII) uses a similar methodology to GLOBIO in that it measures the abundance of species after disturbance compared to a modelled ‘pristine’ state. For this metric however, baseline conditions are based on records of contemporary biodiversity in existing intact areas of similar habitat. The **Biodiversity Impact Metric** uses the BII as a measure of the ‘quality’ of biodiversity and combines it with inputs on the location, area and intensity of a company’s activities to provide a ‘footprint’ of biodiversity impact.

These metrics are largely comparable each other. They are each underpinned by databases and meta-analyses of existing studies and can be used to model the biodiversity impacts of certain disturbances. They have the advantage that they do not require collection of primary on-site data. The only input data required from users is information about the location, area, and intensity or nature of land use associated with their operations (Table 8).

However, the disadvantage is that the outputs are modelled and are therefore based on various assumptions and simplifications which reduce the accuracy of the biodiversity measures. The results are also not specific to the users' sites and will be less sensitive to changes in biodiversity that result from a specific change in management. The reliability of the outputs is dependent on the coverage of the research in the underlying database and on the quantity and replication of studies of relevant studies. For example, the usefulness of outputs for a user wanting to model biodiversity for a perennial crop farm in lowland Brazil will depend how many studies in the underlying database represent this particular set of criteria.

Table 8.

Characteristics of biodiversity metrics

Scale of application
🏢 Company
⚙ Producer
🌐 Global
🏞 Landscape/ Jurisdiction

◉ High
◎ Medium
• Low

Metric type	Metric description	Scale of application	Considerations around sustainability definition				Considerations around performance monitoring					Considerations around data use			
			Metric measures the sustainability issue	Metric is underpinned by evidence	Metric addresses material impacts	Metric is widely used and easily communicable	Practical to collect	Practical to implement	Context-sensitive	Responsive to change	Scalable	Baseline data available to provide context	New tools/software emerging?	Implementation 'user package' exists	Main issues/comments
Impact	Biodiversity Impact Metric	🏢	◎	◉	◉	◉	◉	◉	◉	◉	◉	✗	✗	✓	Input data required from user is straightforward (land use area and location etc) although biodiversity impact is calculated based on modelled data. Availability of data for different areas will affect accuracy of outputs.
	GLOBIO	🌐	◎	◉	◉	◉	◎	◎	◎	◎	◎	✓	✓	✗	Is a modelling framework, largely designed for policy support. Mostly on a global scale, although it has also been used at a national level and to measure a company's biodiversity footprint. Requires availability of detailed data on historical and current land use.
Outcome	Species count/species richness	⚙	◎	◎	◎	◎	◎	•	◉	◎	◎	✗	✗	✗	A direct measure of species diverity on site, but not a comprehensive measure of biodiversity which also encompasses species abundance. Requires collection of primary ecological data.
	Global Biodiversity Score (GBS)	🏢	◎	◎	◎	◉	◉	◎	◎	◎	◎	✗	✗	✗	Most applicable at a global scale. Based on evidence-backed models. Not a direct measure of biodiversity.
	Shannon Index of Diversity	⚙	◎	◉	◎	◉	•	•	◉	◎	•	✓	✓	✗	Very scientific, 'purist' metric of biodiversity requiring collection of primary data on species richness and abundance. Considered a reliable measure of biodiversity.
	Biodiversity Intactness Index (BII)	🏞	◎	◎	◉	◉	◉	◎	◎	◎	◉	✗	✗	✗	Debate about accuracy of the index, particularly original version which was based on expert opinion. Updated version is based on statistical models and linked to the studies in the PREDICTS database, so is based on evidence at some level.
	Living Planet Index (LPI)	🌐	◎	◎	◉	◉	◎	◎	◉	◉	◎	✗	✗	✗	Mainly a global-scale measure of species population trends, although potential to adapt to regional and national context and to model changes in response to 'threats'
	Record of key species	⚙	◎	◎	◎	◉	◎	◉	◉	◎	◎	✗	✗	✗	Difficult to find specific examples of implementation.
Output	Existence of conservation management plans	🏢	•	•	◉	◎	◉	◉	◉	•	◎	✗	✗	✓	Existence of a management plan does not provide a direct measure of biodiversity. Benefits will rely on the plan being based on evidence about effective interventions as well as effective implementation.
	Area of natural or native vegetation cover	⚙	◎	◎	◎	◎	◎	◉	◎	◎	◉	✗	✗	✓	Measure of habitat cover likely to be a good indicator of biodiversity, but is not a direct measure.
	Retention of forest or other natural habitat	⚙	◎	◎	◎	◎	◉	◉	◉	◉	◉	✗	✗	✗	Coverage of forest or other natural habitat is a good indication that biodiversity associated with this habitat will also be retained. However, quality of habitat and connectivity with other natural/semi-natural habitat must be considered.
	CFT Biodiversity	⚙	◎	◎	◎	◎	◉	◎	◎	•	◎	✗	✗	✗	Inputs - of what management measures are in place on land - are easy for producers to record, and the metric's scores to reflect biodiversity impacts are supported by evidence. Only available for temperate systems so far, and, to date, relatively few examples of use are publicly available.

It should be noted that relative species abundance or intactness is only one aspect of biodiversity and does not communicate anything about the status of specific species. Designers of these metrics therefore propose that, in order to be robust, they should be used alongside complementary indicators.[69,79] The models themselves will also require particular expertise, or at least detailed guidance, in order to use them. However, they do have the advantage that they can be calculated at a range of scales, from landscape level to a global scale.

The **Living Planet Index (LPI)** included here is specifically designed as a global-scale metric, although regional or national assessments are also possible[71]. Similar to the other metrics in this category, it uses collated data from existing studies. In this case, the data is approximately 20,000 time series showing trends in population for a range of different species. The LPI is globally influential and used as a biodiversity metric by the Convention on Biological Diversity. It can provide useful evidence-based context on trends in biodiversity or the population of specific species on a global or national scale which may offer a baseline to the biodiversity impacts of a metric user's specific operations.

The final group of metrics, management measures, involve measuring the presence of management measures intended to enhance biodiversity as a proxy indicator of biodiversity. One option included here is recording the **existence of conservation management plans**. This information should be relatively straightforward to collect from the records or reporting of producers or producer groups and can be aggregated along the supply chain. However, the actual impacts for biodiversity will depend on whether and to what to degree the management plans are implemented, and whether the measures within it are actually beneficial for biodiversity (Table 8).

The **Cool Farm Tool biodiversity metric** is designed to address some of these challenges. It is also based on the presence of management measures for enhancing biodiversity, but links this to underlying evidence for the effectiveness of the different measures. In the tool, users select the management options that they are using from a list of possible measures. Each measure gets a score, gaining more points if collated evidence from field-based peer-reviewed studies shows the intervention is beneficial for biodiversity. This means collection of primary biodiversity data is not required and the only information needed from the user is about the management options they are implementing. This should be relatively straightforward to collect based on surveys of producers and/or their sites or from existing producer reports about site management. The tool is only currently available for temperate farming systems but is planned to be extended to cover Mediterranean and Tropical farming systems. Because it does not use primary data, it is relatively insensitive to changes in biodiversity as the user input only captures changes in management (Table 8).

6.3.2 Key practical trade-offs

Cost vs responsiveness

All of the metrics that are based on enumerating species, whether direct measures of biodiversity (e.g., species richness) or indicators (e.g., the presence of rare, threatened or endangered species) require on-the-ground surveys and significant ecological expertise. This is likely to make the cost of them prohibitive for anything other than larger companies or research projects (Table 8). It is perhaps for this reason that some VSS have taken the decision to only require measurements of species when it is critically important to do so (e.g., the FSC's approach to measuring species guilds in HCVF).

By contrast, those metrics that rely on secondary data and modelling (e.g., GLOBIO, the Biodiversity Intactness Index or the Living Planet Index) or that rely on indicators that are several steps away from direct measures of biodiversity (e.g., the area of natural habitat) will rarely be sufficiently responsive to demonstrate the impact of changes in management practice on biodiversity (Table 8). There is very little middle ground in biodiversity metrics.

Shared value for clients and VSS

Biodiversity is, by its essence, place based. The number of species varies hugely across the globe, even in relatively intact habitats. For example, there are around 60 species of tree native to the UK. The same number, or more, can be found in a single hectare of Amazon rainforest. Similarly, the identity of species and habitats varies from one place to the next, sometimes over small spatial scales. This represents a challenge for communicating the global impacts of a VSS on biodiversity: one is never comparing like with like.

From a client's perspective, it is perhaps more useful to understand what they can usefully do to increase biodiversity, rather than the 'amount' of biodiversity per se. Metrics that are either generated by management practices (e.g., the Cool Farm Tool biodiversity metric –unfortunately limited in its application to temperate regions at present) or that combine the area and quality of natural habitat (e.g., the RA-SAN approach to measuring retention of natural habitat) are likely to give clearer indications of what management practices are required to enhance biodiversity. However, these types of metric allow only limited claims on biodiversity impact at a global level, and in some cases are difficult to aggregate to a global scale.

6.3.3 On the horizon

Newer approaches to measuring biodiversity include investigations into the potential to sample genetic material to measure biodiversity. For example, one experiment in Sweden uses long-standing sampling filters to capture DNA via cell fragments circulating in the air in order to measure the relative presence of all types of organisms (plants, animals, fungi, bacteria and viruses) from all types of environments (soil, water, land and air)[72]. Other researchers are also investigating the potential of genetic sampling tools for measuring biodiversity[73]. These techniques are thus far in the early stages of development and require specialist equipment and sampling expertise. They also rely on the existence of, and access to, location-specific reference databases on background genetic diversity for the area in question. As these techniques and databases are developed and become increasingly accurate and accessible, they could become an option for biodiversity monitoring in the future.

Box 3.

Implementation of biodiversity metric: Anglian Water

Anglian Water (AW), a water company managing water supply infrastructure and catchment areas in Eastern England, is adopting a biodiversity metric to inform their efforts towards 'net biodiversity gain', in which more biodiversity is reinstated than is lost following development on a site or a change in site management.

Selecting the metric. This is the first time Anglian Water has used a biodiversity metric. Their starting point was the Biodiversity Impact Metric (BIM, described above) which, according to Chris Gerrard, Catchment and Biodiversity Manager at Anglian Water, was appealing for its "neatness" with its focus on three straight-forward dimensions of biodiversity and land use change: the amount of land, the value of the land for biodiversity and the impact of the proposed land use change for biodiversity.

However, two main factors ultimately led AW to seek an alternative metric to suit their needs. Firstly, as AW's operations have a relatively localised geographical scope, they can easily access their sites and collect higher resolution primary data on biodiversity compared to the BIM data which draws from a global database of studies. Secondly, there has been growing interest in the UK in the concept of 'net biodiversity gain' and an expectation that the government will mandate a particular approach to calculating this.

Anglian Water therefore instead chose to adopt The Biodiversity Metric 2.0 developed by Defra, a department of the UK government.[74] The approach in this metric is comparable to the BIM, measuring; i) area of habitat, ii) quantity (the condition of the habitat; poor, moderate or good) and, iii) importance (distinctiveness or rarity of the habitat), which can be accurately assessed using local primary data and is aligned with the broader UK policy requirements for biodiversity measurement.

Data requirements and measuring biodiversity. Quantity and quality of habitat is used as the measure of biodiversity in this metric. In preparation for using the metric, Anglian Water gathered baseline data for their sites using GIS data to map habitat type and quality. This is translated into preliminary 'biodiversity units' for the habitats in a site. Other components of the metric, such as a measure of the strategic significance of sites in terms of their importance for biodiversity, currently use default scores, but will be updated when on-site surveys are undertaken.

The metric will be used whenever a development or change in management is planned on a site. Independent ecological consultants employed by AW visit the site to 'ground truth' the baseline GIS data on habitat area and assess the habitat condition ('quantity') and distinctiveness ('importance') using established assessment methodologies. They verify whether the number of biodiversity units has been accurately assigned and then review plans for the site development to verify the estimated associated loss of biodiversity units. They then advise on the necessary compensation measures that will be needed to ensure 'net biodiversity gain'.

Benefits. The key benefit for AW is that the metric will allow them to measure their impact on biodiversity in order to "play our part in restoring biodiversity in our region". It will provide tangible evidence to reinforce AW's commitment to delivering more biodiversity alongside its other objectives which, Chris Gerrard sees, will drive innovation. There is also expected to be a reputational benefit for AW and its partners of measuring and reporting their biodiversity impacts.

Box 3 cont.

The benefits of using the Defra Biodiversity Metric 2.0 specifically are that; i) it uses an established approach to measuring biodiversity (also used in the global-scale BIM), ii) it directly meets existing and anticipated UK policy requirements for biodiversity measurement and reporting, iii) it was developed by experts outside of AW so is likely to be seen as more robust and objective than a metric developed internally, iv) any criticisms of the metric will be distributed across its creators and other users and there will be a broad effort to improve it, v) it does not involve any new or complex data collection requirements for ecological consultants, and vi) using it contributes towards the universal adoption of an existing metric rather than the creation of another independent metric with unique requirements for data collection and reporting.

Challenges and lessons. AW is in the early stages of adopting the metric. So far, a major challenge has been ensuring sufficient competence in using it. In particular, having the metric be readily understood by the AW planning teams when they are designing site developments so that they effectively incorporate measures for net biodiversity gain. Measures will also be needed to ensure that the ecological consultants conducting site assessments are trained in how to collect and report the metric data and do so with equal competence and diligence, and following a standardised methodology. The assessment also entails more time for surveying and reporting – estimated at an additional 1-2 days spent by the ecological consultants per site – and therefore implies additional costs for developers. Nevertheless, all of AW's partners support adoption of the metric.

A key lesson for AW has been the importance of fitting the Biodiversity Metric into the already well-established processes, governance and tools they have in place for designing, building and operating assets. They have recognised the need for heavy involvement of end-users in co-designing how the metric is incorporated, otherwise "*the metric is on the back foot from the start until the end users are fully comfortable with how to use it and have had the chance to tweak it so it works from their perspective*".

6.4 Summary

If collecting on-site data is possible, by far the most accurate and robust measure of biodiversity would be a biodiversity index like the Shannon Index. Alternatively, enumeration of specific indicator species or groups of species (such as rare, threatened and endangered species) could be used in specific contexts, as long as care is taken to ensure that the species list is representative of broader biodiversity.

More recently, a number of newer metrics have been developed which take advantage of the significant body of existing ecological studies in order to model biodiversity at various scales. These focus on changes in mean species abundance or overall biodiversity intactness in response to disturbance and are often designed to calculate a biodiversity 'footprint' for company operations. However, they do not capture anything about the status of specific species and designers of these metrics recommend they be combined with other indicators.

Some metrics approach measuring biodiversity from the angle of recording the presence of measures intended to enhance biodiversity. These metrics require modest effort to collect data, but the actual impacts on biodiversity are unknown and will depend on how well the measures are implemented.

7 Water use

7.1 Background: water resource sustainability

The terrestrial water cycle is significantly impacted by human management and use, for example through the use of irrigation water in agriculture[75]. Addressing water scarcity and quality is a critical societal challenge for this century as it underpins broader sustainability challenges such as food and energy security, poverty, conflict, climate change adaptation and biodiversity loss[76]. Improving the sustainability of water resources is summarised by the Alliance for Water Stewardship as seeking to achieve the following outcomes:

- **Good water governance** – This ensures responsible sharing of water resources in the interests of users and the natural environment in line with the principles of water stewardship. It encompasses all aspects of how water is managed by governments, regulators, suppliers, and users.
- **Sustainable water balance** – The condition whereby ongoing water use in the catchment has no long-term negative impact on the natural environment and legitimate water users. This concept is often discussed alongside measures of 'water scarcity', 'water risk' and 'minimum environmental flows'.
- **Good water quality status** – When a water body meets the requirements of native flora and fauna, and for human needs. This includes measures of physical, chemical and biological parameters
- **Important water-related areas (IWRAs) are protected** – These are specific areas of a catchment that, if impaired or lost, would adversely impact the environmental, social, cultural or economic benefits derived from the catchment in a significant or disproportionate manner.
- **Safe water, sanitation and hygiene for all '(WASH')** - Used in the international development sector to refer to the combined area of effort to address basic human water needs and rights related to access to safe and sufficient water for drinking, food preparation and washing.

Impact assessments examining water sustainability aim to understand and quantify the issues above – for instance the ways in which business activities may affect issues such as community access to water, human health, or the in-stream flows required for healthy ecosystems[77]. In this research, the focus is on metrics to quantify '*water use*'. This has been interpreted as measuring the sustainable use of water (i.e. water balance). This is the focus of the remainder of this section, which first describes some of the key considerations affecting the development of water metrics.

7.1.1 Water use vs. consumption

A key aspect of water metric terminology is the different meanings of water use (or withdrawal) and water consumption. The term 'use' refers to the volume of water that is withdrawn from a watershed but then returned for reuse (e.g. water extracted from a river for washing activities that is then returned to the watershed via wastewater processes). 'Consumption', on the other hand, quantifies the volume of withdrawn water that is non-recoverable. For example, evapotranspiration of water from plants is released into the atmosphere, and water included in products that are then exported to other regions and countries[78]. Consumption is sometimes referred to as 'net water use'[79]. These metrics are all volumetric, expressed in litres, cubic metres, etc.

There are benefits to adopting both water 'use' and 'consumption' types of metrics. Water use is a measure of the level of competition and dependency on water resources, while consumption affects downstream water availability; understanding this is essential to evaluating water scarcity and the impacts on ecosystems at the watershed level.

7.1.2 Water balance – the importance of location, scale, and timing

Water is typically a local or regional resource, and so the local context is a critical element when considering the overall sustainability.[80] The timing of water availability and demand is of similar importance. In fact, while the global supply of available freshwater is adequate to meet all current and foreseeable water demands, its spatial and temporal distributions means that there are local areas where demand outstrips supply[81]. So, although water is a 'renewable' resource, its availability is finite in terms of the amount available per unit of time in any one region.

Water supply and demand is most usefully assessed at a 'catchment', 'basin' or 'watershed' level[82]. This is because the 'water balance' is the aggregate outcome of many users within this hydrological unit.[83] The concept of 'water balance' is key to many definitions of water sustainability. A sustainable water balance is defined as "*the condition whereby ongoing water use in a catchment has no long-term negative impact on the natural environment and legitimate water users. It is typically assessed on an annual timescale. For a sustainable balance, total net water abstractions do not exceed natural replenishment of water bodies, while also ensuring water bodies maintain viable flows and water levels to sustain themselves, and the species that depend on them, in a healthy condition. A condition where outflows are consistently larger than inflows is a non-sustainable water balance*"[84]. An alternative term for sustainable water balance is "environmental flow". According to the Brisbane Declaration, the term "environmental flow" is the quantity, timing and quality of water flows required to sustain freshwater and estuarine ecosystems and the human livelihoods and well-being that depend on these ecosystems[85].

Water metrics should therefore also seek to be as specific to locations and seasons as possible, in order to be sensitive to the watershed context.

7.1.3 Water efficiency – Jevons' paradox

At a site-level, it might seem obvious to pursue a strategy focused on increasing water use efficiency and so pursue metrics that measure litres of water used per hectare or tonne of production. However, there is a significant body of evidence which shows that efficiency gains at site level can result in no overall improvement to environmental flows at a watershed level[86]. This is an example of a Jevons' Paradox, where efficiency gains 'rebound' or even 'backfire' in pursuing this goal, causing higher production and consumption.[87] In the case of water, improvements in water efficiency Sears et al. (2018) conclude: "*In arid regions,* *irrigation enhances the productivity of rain-fed cropland, since it allows the production of higher value crops on previously marginal land. More efficient irrigation technology also lowers the effective cost of irrigation by limiting non-consumptive use of applied water, lowering the relative cost of more water-intensive crops. Thus, by making more efficient irrigation technology cheaper to adopt, an incentive-based policy can induce the planting of more water-intensive crops on already irrigated land, as well as a shift away from dry-land crops to irrigated crops, both of which may lead to an increase rather than a decrease in water consumption*".

The implication of this is that water metrics **should not focus on water efficiency without taking into account broader watershed contexts** (e.g. before increasing irrigation efficiency an effective water allocation system that prevents an associated increase in overall water consumption needs to be in place).

7.1.4 Water scarcity, stress and risk

When there is a non-sustainable water balance, a watershed can be considered to suffer from 'water scarcity'. Meeting the needs of society and the environment in the future will be heavily constrained by the scarcity of freshwater[88]. The United Nations describes water scarcity as "*the point at which the aggregate impact of all users impinges upon the supply or quality of water under prevailing institutional arrangements to the extent that the demand by all sectors, including the environment, cannot be satisfied fully.*"[89]

Water scarcity is driven by two main factors: firstly, the climate, which controls the availability of freshwater resources and seasonality in supply, and secondly, water demand, which is a function of economic activities. So, although water scarcity happens in areas with low rainfall, human activities add to the problem, for example where there is intensive agriculture and water demanding industries.[90] Measures of watershed water scarcity are included in metrics that seek to quantify the relative impact of water use in one region compared to another. It benefits from being a physical, objective metric that can be measured consistently across regions and over time.[91]

Other indicators can be added to water scarcity metric to describe 'water stress' (see Figure 3 below). This is a more subjective measure and so not recommended as a basis for sound quantitative comparison at the moment.[91]

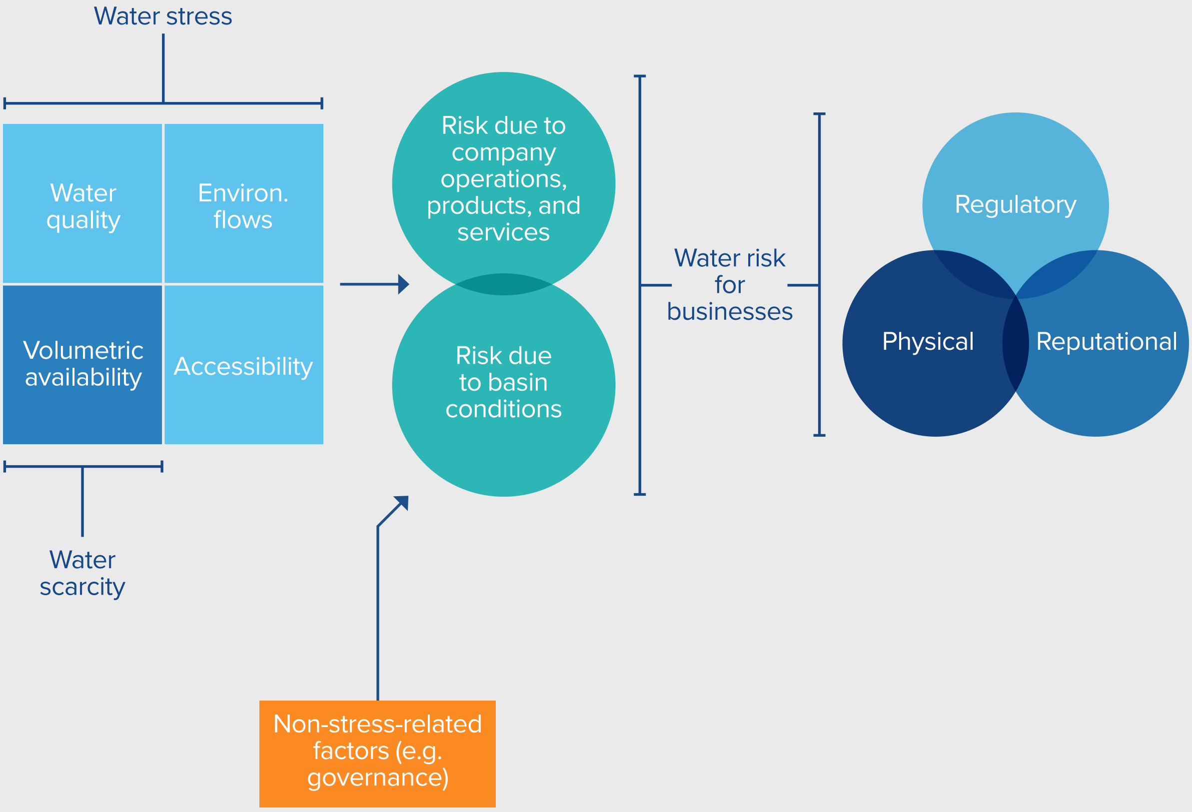
Businesses operating or sourcing from water stressed regions face 'water risk'. Water risk is commonly defined as covering three aspects:

- Physical risk – Too little or too much water or water that is unfit or inaccessible
- Regulatory risk – Changing, ineffective, or poorly-implemented public water policy
- Reputational risk – Stakeholder perceptions of irresponsible water management

Including some measure of local water scarcity in a water metric is therefore important for understanding the sustainability of the use of this resource.

Figure 3:

Water scarcity, stress, and risk[91]



7.1.5 The important role of governance

Finally, given water is a shared resource with multiple competing users, including local ecosystems, effective governance is critical. According to the Water Governance Facility[92], water governance refers to the political, social, economic and administrative systems in place that influence water's use and management. Water governance determines the equity and efficiency in water resource and services allocation and distribution, and balances water use between socio-economic activities and ecosystems. Governing water includes the formulation, establishment and implementation of water policies, legislation and institutions, and clarification of the roles and responsibilities of government, civil society and the private sector in relation to water resources and services. **Credible water use metrics should therefore include some indicators that assess the degree to which local watersheds are sustainably managed – and whether the reporting organisation is participating in these multi-stakeholder processes.**

7.2 Overview of the metrics assessed

We assessed a selection of 13 water metrics (see Table 9). Two of these are outcome metrics, nine measure outputs, and one measures inputs (Table 10).

Note that water use is the only sustainability issue considered in this report for which a full ISEAL member, the Alliance for Water Stewardship, has developed a standard focused on improving it.

Table 9:

Description of assessed water use metrics

Metric	Metric description
Average water stress or 'risk' of catchment or basin where producers operate	Catchment or basin water risk and water scarcity indicators are developed by researchers using a variety of environmental, social and economic data. They are applied by organisations to understand the relative potential exposure to water risk faced by producers and facilities in different locations. For example, WWF's Water Risk Filter's risk assessment is based on a company's geographic location, which informs a site's basin-related risks, as well as characteristics of its operating nature (e.g., its reliance upon water, its water use performance given the nature of the business/site), which informs a site's operational related risks. (The Water Risk assessment method description is available here: https://waterriskfilter.panda.org/en/Explore/DataAndMethod)
Water consumption based on available water remaining (AWARE)	AWARE, is based on the quantification of the relative available water remaining per area once the demand of humans and aquatic ecosystems has been met, answering the question "What is the potential to deprive another user (human or ecosystem) when consuming water in this area?" The resulting characterization factor (CF) ranges between 0.1 and 100 and can be used to calculate water scarcity footprints as defined in the ISO standard.
The total quantity of water used	The total quantity of water used by an organisation. Under normal definitions this includes water from 'mains' supplies as well as those extracted from surface and ground sources.
Number of producers with applied water conservation or water use reduction practices	Amount of money invested in projects, operational changes, new assets, etc. that deliver improvements in water use across single or multiple years.
Proportion of water reused/recycled	Quantity of recycled water divided by total quantity of water used on site.
The area of land irrigated	Area of crop land that has irrigation applied to it.
Quantity of irrigation water used per unit crop produced	Total quantity of irrigation water used by an organisation (from any source) divided by total crop output.
% of water used that comes from storm water or rainwater harvesting	Quantity of rain/storm water use divided by total quantity of water used on site.
Number of sites in a catchment engaged in water stewardship activities	Tally of the number (or %) of sites in a water catchment that are engaged in water stewardship activities.

Table 9 cont:

Description of assessed water use metrics

Metric	Metric description
Water footprint of product or business	The water footprint has three components: green, blue and grey water. Together, these components provide a picture of water use by delineating the source of water consumed, either as rainfall/soil moisture or surface/groundwater, and the volume of fresh water required for assimilation of pollutants.
Cool Farm Tool Water Assessment	The Cool Farm Tool Water Metrics assessment (CFTw) enables farmers and their supply chains to assess their water demand, water consumption and irrigation efficiency with standard crop data, using localised meteorological information. From this, the CFTw produces a Water Footprint Network compliant blue and green water footprint and a crop/soil water balance. At this stage, it does not produce a grey water footprint.
Water exploitation index plus	The water exploitation index (WEI) is the mean annual total abstraction of freshwater as percentage of the mean annual total renewable freshwater resource. Available at basin level in EU
Investment in stewardship activities on and/or off site	Amount of money invested in projects, operational changes, new assets, etc. that deliver improvements in water efficiency on site or improved water stewardship of catchment across single or multiple years.

The assessed metrics can be broadly categorised into four types:

1. Volumetric water withdrawal and consumption metrics.
2. Scarcity-adjusted metrics that consider local water balances.
3. Indicators that measure the adoption of water management practices.
4. Indicators of commitment and investment in sustainable water use.

7.3 Applicability and trade-offs

7.3.1 Key technical decision points in measuring water use

Which aspect of water use to measure: volumetric, scarcity adjusted, practice adoption or investment?

The volumetric water metrics are widely used: approximately one third of the water use metrics used by ISEAL members fall into this category. Examples include the total volume of water extracted (Better Cotton Initiative) and the quantity of water used per unit weight of produce (Utz). Water stewardship activities to date have been largely focused on operational water use efficiency, despite the concerns of rebound effects highlighted above. In many cases, there is limited accounting for the surrounding basin context.[93] Some key examples include:

- Water Footprint. The Water Footprint Network's (WFN) corporate water footprint (WF) calculation itself does not attempt to account for the context of a watershed (e.g., water availability, allocation among users, etc.) or quantify or otherwise assess a company's water-related impacts[94] (Table 10).
- The total volume of water used (cubic metres). Requires primary data on water use

to be collected from each user. Does not take into account the context of water use (Table 10).

- Volume of irrigation water used per unit of crop area or production. Primary data on quantity of irrigation water used at a site, plus data on total crop production (tonnes) is required. Does not take into account the context of water use (Table 10).

The second category of water metrics sees volumetric measures adjusted for local water scarcity. This ranges from simpler metrics such as the **total water withdrawal from all areas with water stress**, a metric recommended under the Global Reporting Initiative (GRI).[95] More advanced models exist, such as **Water Consumption based on Available Water Remaining** (AWARE)[96] and the Cool Farm Tool's Water assessment (Table 10).[97] The **Cool Farm Tool Water assessment** (CFTw) enables farmers and their supply chains to assess their water demand, water consumption and irrigation efficiency with standard crop data, using localised meteorological information. From this the CFTw produces a Water Footprint Network (WFN) compliant blue and green water footprint and a crop/soil water balance. The tool follows the same modus operandi as the CFT greenhouse gas tool (see Section 4), requiring a few additional inputs (sowing date, harvesting date, soil data and irrigation management information).

The water exploitation index plus (WEI+), as an indicator of water scarcity, aims to illustrate the percentage of total renewable freshwater resources used in a defined territory (basin, sub-basin, etc.) in a given period (e.g. seasonally, annually). Values above 20% indicate that water resources are under stress, and above 40 % indicate severe stress and a clearly unsustainable use of freshwater resources.[98]

This type of metric is less common amongst ISEAL members, with only five examples being identified, including the reduction of water scarcity risk in catchments (Alliance for Water Stewardship).

The third category of indicators covers the adoption of specific practices that are likely to reduce water use or improve watershed sustainability, for example

- % of water used that comes from storm water or rainwater harvesting
- Proportion of water reused/recycled

This type of metric is the most commonly used amongst ISEAL members (37 of the 62 water use metrics used by ISEAL that we identified). Examples include water from recycled/non-potable sources (%) and % of total water used that is recycled (Global Infrastructure Basel).

The main advantages of this group of metrics are the relative ease of data collection and implementation, context sensitivity, that they are scalable and relatively widely used. However, they do not have an existing external baseline against which to compare performance, may not reflect the most material issues, and are not especially context sensitive (Table 10).

The final category enumerates commitments to, or investments in, more sustainable water use practices (Table 10), including:

- Investment in stewardship activities on and/or off site.
- Number of sites in a catchment engaged in water stewardship activities

This type of metric is used by some ISEAL members, but less frequently that volumetric or practice adoption measures. Examples include the number of sector-wide water stewardship policies the number of sites in a catchment engaged in water stewardship activities (Alliance for Water Stewardship), The main advantages of these metrics lie in the ease of data collection and implementation (Table 10).

Table 10.

Characteristics of water use metrics

Scale of application
- ⏩ **Supply chain**
- ⚙ **Producer**
- ⚑ **National**
- ⛰ **Landscape/ Jurisdiction**

◎ **High**　◉ **Medium**　● **Low**

			Considerations around sustainability definition				Considerations around performance monitoring					Considerations around data use			
Metric type	**Metric description**	**Scale of application**	Metric measures the sustainability issue	Metric is underpinned by evidence	Metric addresses material impacts	Metric is widely used and easily communicable	Practical to collect	Practical to implement	Context-sensitive	Responsive to change	Scalable	Baseline data available to provide context	New tools/software emerging?	Implementation 'user package' exists	**Main issues/comments**
Outcome	Average water stress or 'risk' of catchment or basin where producers operate	Landscape/ Jurisdiction	Medium	High	Medium	Medium	High	High	High	Low	High	✓	✓	✓	Potentially powerful metric of water stress but would probably need to be used alongside other metrics to reflect the potential water use impacts of certified entities.
	Water consumption based on available water remaining (AWARE)	Supply chain	Medium	High	High	Medium	Low	Low	High	Medium	High	✗	✗	✓	A rigorous metric which allows quantification of water use within the context of water availability. Challenges may include the cost of data collection, management and analysis.
	The total quantity of water used	🏢	Low	Medium	Medium	High	High	High	Low	High	High	✓	✗	✗	Water use metric that is easy to measure and frequently used, but that is not sensitive to local context.
Output	Number of producers with applied water conservation or water use reductopn practices	🏢	Medium	Low	Medium	Low	High	High	High	Medium	High	✗	✗	✗	Potentially powerful indicator of the seriousness with which certified companies and VSSs take water stewardship and efficiency, but an input indicator not directly related to actual impacts on water catchments. Indeed poorly targeted investment could result in an unintended increase in water impacts.
	Proportion of water reused/ recycled	🏢	Medium	Medium	Medium	High	High	High	Low	High	High	✗	✗	✗	Simple water efficiency metric but does not necessarily capture the relative impact of water use by a producer or company.
	The area of land irrigated	Producer	Low	Low	Low	High	High	High	Low	High	High	✓	✗	✗	An indirect measure of water abstraction or use that is not sensitive to local contexts.
	Quantity of irrigation water used per unit crop produced	Producer	Low	Low	Low	High	High	High	Low	High	High	✗	✗	✗	Commonly used water efficiency measure that has the potential to deliver unintended negative outcomes at a basin level.
	% of water used that comes from storm-water or rainwater harvesting	Producer	Medium	Medium	Medium	High	High	High	Low	High	High	✗	✗	✗	Indicates the uptake of sustainable water use practices but is not sensitive to context.
	Number of sites in a catchment engaged in water stewardship activities	Landscape/ Jurisdiction	Low	Low	Low	Medium	High	High	High	Medium	Low	✗	✗	✗	A useful metric for water governance, but in itself does not reflect sustainable water use or the on-the-ground activities of certified entities.
	Water footprint of product or business	Supply chain	Medium	High	Low	High	Medium	Medium	Medium	Medium	Medium	✓	✗	✓	Reflects sources and scale of water use (e.g. rain, abstraction), but not the context or sustainability of that water use.
	Cool Farm Tool Water Assessment	Producer	Low	High	High	Low	Medium	High	High	High	Medium	✗	✓	✓	Provides detailed information to farmers on how they might better mannage water, but less applicable as a metric for global (scheme wide) reporting.
	Water exploitation index plus	Landscape/ Jurisdiction	Medium	High	High	High	High	Medium	High	Medium	High	✓	✗	✓	Available at the EU level only at present, and may not be sufficiently responsive to the actions of certified entities.
Input	Investment in stewardship activities on and/or off site	🏢	Medium	Low	Low	High	High	High	High	Medium	High	✗	✗	✗	Potentially powerful indicator of the seriousness with which certified companies and VSSs take climate change, but an input indicator not directly related to GHG emissions. Indeed poorly targeted investment could result in an unintended increase in emissions.

7.3.2 Key practical trade-offs

Cost and robustness

It is widely accepted that volumetric measures of water use alone are not an adequate indicator of a company's water-related business risks or social and environmental impacts, as they do not consider the local water context. As a result, impact assessments should have two main components: 1) an assessment of the local water resource context, and 2) a quantification of an organisation's share of water use/discharge within that local context. Both approaches pose significant technical challenges to organisations[99].

One of the main challenges when establishing context-based water targets is the availability of basin data e.g. trends in water availability. Unfortunately, many of the open source tools are still too coarse to offer meaningful contextual data at the local scale[100]. Some are producing high resolution ('downscaled') maps to enable more relevant water strategies and metrics to be developed, in places such as the United Kingdom[101] and South Africa[102]. Gathering data on local water balances and environmental flows to understand this, is more technically challenging than simple on-site efficiency measures. This is the key practical trade-off when considering water metrics.

In addition to the lack of data, the other major trade-off between simple volumetric measures and watershed metrics, is the latter requires allocation of staff resource for engagement with local stakeholders (e.g. as part of local water shed management initiatives).

Box 4.

Implementing water metrics: Conor Linstead, Freshwater Specialist at WWF-UK

WWF has had a strategic focus on freshwater for many years as it is often abstracted from at-risk ecosystems and is one of the greatest risks to the global economy if not properly managed. The environmental NGO has developed and applied water metrics through its on-the-ground projects with producers, local NGOs and corporate partners in many regions of the world, from citrus production in Spain[103] to textile production in India[104]. WWF was also one of the founding board members of the Alliance for Water Stewardship, notable as being the only full ISEAL member that has developed a standard focused on improving one of the sustainability issues reviewed within this project.

WWF has championed the use of basin/catchment 'water risk' indicators – to identify locations that are likely to be most at risk from over-abstraction or pollution – alongside 'water stewardship', a continuous improvement framework that identifies practical steps organisations can take to better manage freshwater resources and become good water stewards[105]. There is a relatively strong consensus between many stakeholders on these concepts when it comes to corporate action on water sustainability, for instance they are promoted by global coalitions such as the AWS, CEO Water Mandate[106] and researchers such as the Pacific Institute[107].

From Conor's perspective one of the main challenges in implementing water metrics that are relevant at a catchment or basin scale is poor catchment/basin data availability. Globally, there are significant gaps in data on flows and in many cases, even where these data are available environmental flows[108] haven't been defined or expressed in terms of sustainable water allocations for users. This lack of available data on catchment context makes it extremely challenging for individual water users to understand whether the amount of water they are using is within the limits of sustainability for their catchment. In addition, there are limited platforms for sharing

Box 4 cont.

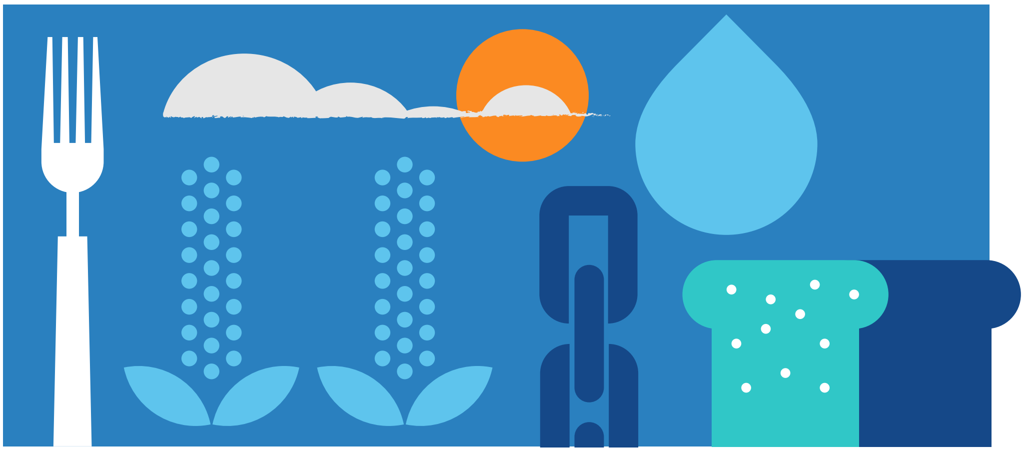
and aggregating water use data and some of the emerging technologies for gathering water data (e.g. remote sensing) are not easily accessible to non-experts.

Given these challenges, Conor encourages organisations engaging in freshwater sustainability to develop a basket of water metrics on water quantity that includes organisation-level metrics, but places these in the context of catchment and basin level metrics. Organisation level metrics should include both absolute volumetric measures on water use and metrics on an appropriate per-unit basis (e.g. tonnes of production, Ha of area). Catchment/basin scale metrics could include whether environmental flows have been defined, legislated, or are being implemented/achieved in the wider catchment (e.g. is Good Ecological Status being achieved if in the jurisdiction of the EU). In the absence of such data or environmental flow processes, water risk evaluation tools (such as the Water Risk Filter) can be used to assess wider catchment/basin condition. In addition, qualitative metrics can be used to track engagement in water management and governance processes that indicate good water stewardship practice. E.g. tracking adoption of relevant water management best practices, and/or measuring wider understanding and engagement with basin-level governance and contexts (such as knowledge and participation in water management plans, understanding of water sources, etc). These measures speak both to business profitability (e.g. lower water use costs) but also helps ensure producers are more open to collective action – something that is ultimately needed to ensure catchments are sustainable. These recommendations mirror the findings of a review undertaken by WWF Germany on the adoption of Water Stewardship measures in sustainability standards in 2017[109]. The conclusions of this work still hold true in 2020: "*Water stewardship integration begins with a deeper understanding of your context and agricultural water risks, be sure you are considering collective actions and engagement in water governance, ensure efficiency requirements are supplemented with cumulative basin impact considerations, and collaborate as much as possible*".

7.4 Summary

Site water targets informed by catchment context are considered best practice by leading NGOs. Water targets should respond to priority water challenges within a watershed[110] and because an organisation's water risks depend heavily on external factors, companies are likely to focus on a combination of quantitative accounting of internal operations (e.g. water use efficiency measures) plus indicators that measure the water balance of local watersheds.[111] However, there is less alignment on precise metrics for measuring water use impacts.

As has emerged in climate impacts metrics space, best practice is to use context-based water targets (CBWTs) that measure the degree to which water use by a company's sites (and those of its suppliers), is sustainable relative to local water resource constraints[112]. For example, a purely volumetric measure of water efficiency (e.g. m^3/tonne) is not supported by some water specialists, as reduced water use does not necessarily translate to reduced environmental impact in catchments and can result in unintended consequences.[113] Instead, the focus should be on encouraging the sustainable management of water resources by local users and policymakers.



8 Poverty

8.1 Background: measuring poverty

8.1.1 Definitions of poverty

Poverty is a contested and evolving concept, with numerous definitions and, consequently, a wide array of metrics associated with its measurement.

Fundamentally, poverty is the state of people not being able to meet basic material needs, such as food, clothing, clean drinking water and shelter. Many of the earliest widespread poverty metrics implicitly understand this to be due to a lack of income (i.e., income poverty). Metrics of income poverty – such as the number or proportion of people living below the 'poverty line' – have been widely used by governments for decades, first coming into use in the 20th Century.

However, in the last 30-40 years, theoreticians and practitioners in poverty reduction have recognised that poverty is multi-faceted in both its causes and as it is experienced. Income is only one of those facets. For example, Amartya Sen defined human development in terms of freedoms, and poverty as the lack of one or more freedom: political freedom; access to economic facilities (e.g., to labour markets); social opportunities (e.g., access to healthcare); security (e.g., social safety nets); and transparency guarantees (e.g., mechanisms for seeking justice).[114] Other multi-faceted definitions of poverty have arisen, and most development organisations (e.g., international NGOs, the UN) now use these more complex notions to define poverty and guide how they measure it. The scope of measurement is frequently the household rather than the individual.

From a measurement point of view, more comprehensive notions of poverty have several consequences. Firstly, they make poverty metrics more complicated to measure, analyse and communicate, as multiple dimensions have to be measured, analysed and reported. Secondly, they can provide information on the likely causes of poverty, as well as its occurrence (e.g., where lack of access to improved sanitation is both a facet of poverty and a cause of it).

The above summary refers to *absolute* poverty. There is also a group of metrics that are commonly used by governments, academics and others to measure *relative* poverty: when a person cannot meet a minimum level of living standards relative to others in the same time and place.

8.1.2 Data and benchmarks

Poverty can only be measured directly with reference to a benchmark: the primary data required to measure poverty typically does not capture information on poverty itself. For example, a survey might estimate the income of a community, finding that the first household has an annual income of US$6,500, the next US$7,400, then US$8,800 … and so on, up to US$25,000. Until that income distribution is assessed in relation to a reference point - the poverty line, a living income benchmark, etc., - it is not a metric for poverty, it is simply a distribution of incomes. In the example cited above, if the living income benchmark was US$8,500, then two households would be below a living income. We include a combination of metrics that for which a number of data collection techniques can be used to compare with a stated baseline (e.g., proportion of households receiving a living income), and metrics that can be used to generate an internal poverty benchmark (e.g., as part of the methodology of the Household Economy Approach, communities self-define poverty groups).

8.2 Overview of the metrics assessed

Table 12 provides a summary of the characteristics of poverty metrics, assessed against the considerations outlined in Section 3.2. Eighteen poverty metrics were included in the assessment (see Table 11). These include nine metrics of absolute income poverty, three of relative income poverty; and six of (absolute) multifaceted poverty. Ten of the selected metrics consider the outcome of poverty, with eight being output indicators. Nine of the metrics originate from external organisations, with a further nine also being used by ISEAL members as part of their monitoring or impact evaluation frameworks. The following section draws out some of the main issues identified from this analysis.

8.2.1 Characteristics of the metrics

A notable trend, when compared to some of the other sustainability issues considered in this report, is that all of the poverty metrics require the collection of primary data. Collecting and analysing poverty data – even if confined to income poverty – can often be time consuming and costly as it usually requires surveys of a sample of individuals or households. For this reason, many organisations choose to use metrics that measure outputs (e.g., such as changes in the wages paid to workers) rather than the outcome (poverty).

A number of the metrics are sufficiently widely used that an external baseline exists, against which the sample population can be compared over space and time (Table 12).

Table 11:

Description of poverty metrics

Metric	Metric description
Proportion of households below the poverty line	The World Bank sets a global 'poverty line' that currently stands at US$ 1.90 a day (purchasing-power parity). This is the absolute minimum level of income deemed adequate to avoid absolute poverty. The poverty line is a reference against which income surveys can be compared.
Proportion of people in relative poverty	The UNDP (and others) measure relative poverty - i.e., the number or percentage of people below a set proportion (usually 60%) of the median income of people in that country.
Proportion of households attaining a Living Income	A living income is defined as the net annual income required for all members of a household to afford a decent standard of living. Elements of a decent standard of living include food, water, housing, education, healthcare, transport, clothing, and other essential needs. The Living Income baseline is significantly higher than the Poverty Line. The allied concept, Living Wage, is used where the poverty concern is with waged labourers, and is used by ISEAL members including ASC. Actual income/wages are acquired through survey.
Human Development Index (HDI)	The HDI was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. The Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.
Multidimensional Poverty Index (MPI)	The MPI assesses poverty at the individual level. If someone is deprived in a third or more of ten (weighted) indicators, the global index identifies them as ‘MPI poor’, and the extent – or intensity – of their poverty is measured by the percentage of deprivations they are experiencing. As with the HDI, the poverty outcomes are focused on measures of standard of living, health and education.
Household Economy Approach (HEA)	Rural livelihood zones are demarcated based on land use, climate, and economic information. In each zone, locations are selected to represent the range of variation across the zone. At each location, interviews establish the characteristics of wealth groups that are recognised there. Using this information, further interviews are conducted with a focus group selected from each wealth group to establish the incomes and expenditures of a ‘typical’ household in that wealth group in a ‘typical’ reference year.
Sustainable Livelihoods Framework	The sustainable livelihoods approach provides a framework that promotes systematic analysis of the underlying causes of poverty. At the level of individual or household, it focuses on five types of 'capital': financial, human, social, natural and physical. This is put into a context of underlying vulnerabilities and the potential forces for transforming poverty. Unlike most other poverty metrics, it is explicitly concerned with sustainability - environmental, social, institutional and economic. The results can be compared against a number of reference points to establish the incidence of poverty.

Table 11 cont:

Description of poverty metrics

Metric	Metric description
Household consumption	Researchers have found household consumption to be a more reliable metric than income, particularly in rural contexts. Measurement of consumption requires detailed surveys in which respondents are asked to detail the food and non-food items they have consumed (over appropriate time periods). The household's consumption is then calculated on a per-day, per-person basis. This is often combined with household wealth indicators, ownership of assets, and access to services.
Poverty Probability Index (PPI)	The Poverty Probability Index estimates the probability that a household is living below the poverty line based on a series of 10 questions about household characteristics and ownership of assets. The answer to each question generates a score, which are then summed and compared to a statistical table to estimate the probability that the household is living beneath the poverty line. Statistical tables are available for 60 countries.
Percentage of farmers/workers experiencing food insecurity	Food insecurity is usually defined as when people do not have physical and economic access to sufficient safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life at all times. The presence of food insecurity therefore indicates several underlying causes of poverty, not only economic poverty.
Ratio of the wages paid by certified entities compared to non-certified entities in the same region	Ratio of the wages paid by certified entities compared to non-certified entities in the same region. This can be reported as either the number or proportion of certified entities that pay a higher wage. The ASC base their version of the metric on three measures: % of overtime compensation paid at premium rate; % of workers receiving living wages and lowest weekly wage for different types of workers
Ratio of lowest entry level wage including benefits to minimum wage and benefits required by law	This compares the number (or %) of workers receiving the actual lowest wages (and benefits) in each producer company to the legal minimum wage as a ratio: a ratio greater than one indicates that the lowest paid workers are receiving more than the statutory minimum, and a ration lower than one indicates that they are receiving less. Fairtrade use a similar metric, but applying to all 'general workers'. Assumes that the legal minimum wage is sufficiently high so that workers who receive it are not in poverty.
Percentage of producer organisations for which the lowest wage paid to general workers increased	Percentage of producer organisations for which the lowest 'real' wage - i.e., adjusted for inflation - paid to general workers increased in the last calendar year, by type of contract and gender.
Percentage of households which have made investments	This metric records the percentage of households which have made investments in the last 3 years, and type of investments made, by gender of member/worker. Note that the version used by SAN-RA includes savings as well as investments. In theory, households that are in a position to make investments in their future are more likely to be improving their poverty status, and so this metric tracks the 'journey' out of poverty rather than the incidence of poverty.

Table 11 cont:

Description of poverty metrics

Metric	Metric description
Percentage of households which have struggled to repay debts	This metric records the percentage of households which have struggled to repay debts in the last calendar year, disaggregated by gender. In theory, households that are increasingly able to repay debts are more likely to be improving their poverty status, and so this metric tracks the 'journey' out of poverty rather than the incidence of poverty. Note that Good Weave use a similar metric; 'Average end of season debt balance by worker' that is perhaps less subjective.
Percentage of households who perceive that the economic situation of their household has improved	This metric records the percentage of producer organisation members and workers who perceive that the economic situation of their household has improved in the last 3 years. It is a subjective metric that does not measure poverty *per se*, but a perceived improvement in economic circumstances.
Household financial income	As used by Fairtrade, this metric records the household income of small-scale producer organisations but could be extended to other types of producer. The similar metric used by Good Weave, 'Average monthly salary for monthly workers', is focused on income from certified entities and does not attempt to estimate all sources of income. Where income from certified entities makes a major contribution to household income, this will be significantly easier to collect than total household income.
Perceived change of profitability of the farm	This metric measures the perception of farmers that their farms are becoming more (or less) profitable. It is not a direct measure of poverty in that farmers can make a profit whilst still living in poverty - but can provide an indication of the 'journey' out of poverty. Note that UTZ also include a metric of the actual change in profitability of the farm amongst their metrics. This will inevitably be more time consuming to collect than a perception but allows more insightful qualitative analysis.

8.3 Applicability and trade-offs

8.3.1 Key technical decision points in measuring poverty

Which definition of poverty to use: income poverty or multi-faceted definitions?
There is no right or wrong in this decision. However, it should be noted that metrics of multi-faceted definitions of poverty have an additional advantage in that they allow greater understanding and engagement with the range of root causes of poverty. This may be important for some standards-setting organisations during standards revision processes and if they are designing on-the-ground projects to reduce poverty.

The main metrics of income poverty are:

Proportion of people below the poverty line (or poverty threshold). This is probably the single most widely used measure of poverty as it is nationally reported in most countries, with varying frequencies. These national censuses, which are also reported at sub-national levels, provide an external baseline with which data collected by VSS could be compared over space and time. The long history of widespread use makes the poverty line a readily understood metric, and numerous guidelines for its measurement and use exist.[115] However, collecting the required data from clients – net household income – requires surveys, which are expensive, and a global poverty line (currently set at US $ 1.90 per day, purchasing power equivalent)[116] is considered by many insufficient to ensure a decent standard of living, whilst also not always reflecting local or sector-specific realities (Table 12).

Proportion of households attaining a Living Income. Comparing net household income against a Living Income (or Living Wage where the individuals of interest are waged workers) is in effect an attempt to correct some of the perceived shortcomings of the poverty line as a reference against which to measure income poverty. The living income sets a reference point for poverty of the cost of a decent standard of life rather than basic survival and is set locally to ensure that it reflects local and sectoral circumstances. The actual (net) household income can be measured in a variety of ways (e.g., the HEA, below), and the gap between actual and living incomes (or proportion attaining or not attaining a Living Income) calculated. However, in places where there is no existing reference point for a Living Income, one has to be set, implying significant investment. Protocols and communities of practice are emerging to support measurement, implementation and influencing on Living Income, for example, the Living Income Community of Practice[117] and the Global Living Wage Coalition[118] (Table 12). A number of VSS use this metric, including Fairtrade and Utz.

ISEAL members use a number of bespoke measures of income poverty, including the proportion of producer organisations for which the lowest wage paid to general workers increased (Fairtrade); the percentage of households that have made investments (SAN-RA, Fairtrade); and the proportion of households that have struggled to repay debts (Good Weave, Fairtrade). Note that all of these are, in poverty terms, output metrics, and are considered further below.

The main metrics that respond to multi-faceted definitions of poverty are:

The **Human Development Index** (HDI) and the **Multidimensional Poverty Index** (MPI) are both used for global poverty reporting, which means that national baselines are available against which to compare a population over time and space (Table 12).[119] Both metrics emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic performance alone. The Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The MPI also combines measures of standard of living, health and education attainment but with a broader set of sub-indicators. The MPI is considered by its designers to be appropriate to use at small scale, and there are some resources available for its implementation.[120] However, there would be significant trade-offs to its use by VSS in that data collection and analysis would be likely to incur significant effort and cost (Table 12).

The **Household Economy Approach** (HEA) uses surveys within selected populations to define wealth groups and establish the incomes and expenditures of typical households within each wealth group. The data gathered typically includes a wide range of information, and it is most used for tracking short-term changes in levels of poverty (e.g., tracking food insecurity needs) rather than structural shifts in poverty status,[121] and data collection is intensive and expensive. However, there are multiple humanitarian and development organisations and individuals that have significant experience in collecting, analysing and reporting the HEA at multiple scales, including sub-national jurisdictions. Several of these have produced materials to support collection and implementation of the HEA (Table 12), including Save the Children, Evidence for Development and FEWS Net.[122] Evidence for Development are developing new software to support the HEA.[123]

Unlike most other poverty metrics, the **Sustainable Livelihoods Framework**[124] is explicitly concerned with sustainability – environmental, social, institutional and economic – and allows exploration of both the causes and outcomes of poverty. The framework focuses on five types of 'capital': financial, human, social, natural and physical. This is put into a context of underlying vulnerabilities and the potential forces for transforming poverty. It is designed to be adapted to specific institutional and poverty situations, and this flexibility could be of interest to VSS. Its limitations include that there is no set measurement or reporting protocol, nor widely applicable baselines, which would mean that VSS would have to create their own. Although we could not identify a formal community of practice, the approach is used by several major international development NGOs and has been championed by the UK Government's Department for International Development, meaning that expertise and user resources are available. Nonetheless, data collection and analysis are intensive and expensive, and the results do not lend themselves to straightforward communication (Table 12).

Measurement of **Household Consumption** was developed in response to the inaccuracies in estimating household income that many researchers have found over the years. Consumption can be recalled more reliably and integrates income with other factors that can affect what and how much is consumed (e.g., access). It is largely used in poverty research, and also by Oxfam in their 'state of the art' Effectiveness Reviews,[125] and a measurement protocol has been established by the World Bank.[126] There is no external baseline and data collection requires surveys of households (or focus group discussions) which makes the data costly to collect, analyse and manage (Table 12).

The **Poverty Probability Index (PPI)** estimates the probability that a household is living below the poverty line based on a series of 10 questions about the household's characteristics and ownership of assets. The answer to each question generates a predetermined score, which are then summed and compared to a predetermined statistical table to estimate the probability that the household is living beneath the poverty line. Statistical tables are available for 60 countries.[127] The PPI is used by a range of NGOs and companies, but has been criticised due to the fact that the statistical relationship between household characteristics and poverty is not constant over space and time, and because interventions can in effect 'game' the metric to record false improvements in poverty.[128] However, the simple survey and pre-prepared scoring and analysis significantly reduces the cost of the PPI compared with most other poverty metrics.

Of the other metrics assessed, the **percentage of people experiencing food insecurity** (used by Fairtrade US) differs from most of the poverty metrics considered here, in that food security is an outcome of various underlying causes of poverty. As such, it is a potentially powerful metric, and baselines (updated according to likelihood of humanitarian crises) exist in some developing countries. Protocols for measuring food insecurity include the Food Insecurity Experience Scale and the Household Food Insecurity Access Scale (HFIAS), both of which rely on short questionnaire-based surveys asking respondents to recall the occurrence and frequency of occurrence of food insecurity. They can be used to both measure the incidence of food insecurity within a population and its changes over time, but are heavily based on perception and recall. Food insecurity is a complex notion and is likely to prove expensive to collect, but is a useful option for impact evaluations, especially in contexts in which food insecurity is likely to occur.

Whether to measure absolute poverty or relative poverty

In general, absolute poverty is likely to be preferred, as metrics of relative poverty arguably record inequality rather than poverty *per se*: it is possible for someone's relative poverty to be low whilst they are still unable to secure their basic needs.

Table 12.

Characteristics of poverty metrics

Scale of application
⚑ National
⚙ Producer
🏢 Company

◎ High
◉ Medium
● Low

Metric type	Metric description	Scale of application	Considerations around sustainability definition				Considerations around performance monitoring					Considerations around data use			Main issues/comments
			Metric measures the sustainability issue	Metric is underpinned by evidence	Metric addresses material impacts	Metric is widely used and easily communicable	Practical to collect	Practical to implement	Context-sensitive	Responsive to change	Scalable	Baseline data available to provide context	New tools/software emerging?	Implementation 'user package' exists	
Outcome	Proportion of people living below the poverty line	⚙	◉	◎	◉	◎	●	◉	◉	◉	◎	✓	✗	✓	Pros include that there are long-established baselines, and that it is a readily communicable metric. Limitations are that it represents a limited conceptualisation of poverty, and would necessitate significant cost to collect and implement for some VSS.
	Relative poverty	⚙	●	◎	◉	◉	●	◉	◉	◉	◎	✓	✗	✓	Pros include that there are long-established baselines. Limitations are that it represents a limited conceptualisation of poverty (i.e., focuses on income only), and would necessitate significant cost to collect and implement for some VSS.
	Proportion of households attaining a Living Income	⚙	◉	◉	◉	◎	●	◉	◎	◎	◎	Partly	✗	✓	Pros include that the metric is very context specific and is set higher than the very basic poverty line. Limitations include the significant resources required to set new baselines and that it represents a limited conceptualisation of poverty (i.e., focuses on income only).
	Human Development Index	⚑	◎	◉	◎	◎	●	●	◉	◉	◉	✓	✗	✓	Pros include that the HDI is focused on the outcomes of poverty reduction. Challenges would include that the metric is essentially intended for the national-level report, and collection and implementation at lower levels may prove costly for VSS.
	Multidimensional Poverty Index (MPI)	⚙	◎	◎	◎	◉	●	◉	◉	◉	◎	✓	✗	✓	Pros include that the MPI is focused on the outcomes of poverty reduction. It is also appropriate to use at smaller scales (although the best known uses are at national and global levels). However, collection and implementation at lower levels may prove costly for VSS.
	Household Economy Approach (HEA)	⚙	◉	◎	◎	◎	●	◉	◎	◎	◉	Partly	✓	✓	Pros include that it includes a wide range of information, various aspects of which can be reported. However, it is most often used to tracking short-term changes in levels of poverty rather than structural shifts, and data collection is intensive and expensive.
	Sustainable Livelihoods Framework	⚙	◎	◎	◎	◉	●	◉	◎	◎	◉	✗	✗	✓	The Sustainable Livelihoods Framework is unique, in that it explicitly places livelihoods in a context of sustainability. This allows for greater exploration of both the outcomes and causes of poverty. However, there is no set measurement protocol, so VSS would have to create their own.
	Household consumption	⚙	◎	◎	◎	◎	●	◉	◎	◎	◉	✗	✗	✗	There is robust evidence that household consumption is a more reliable metric of wellbeing than income. However, it requires significant effort to collect, and is probably best used in evaluations rather than as a monitoring metric.
	Poverty Probability Index (PPI)	⚙	◉	◉	●	◎	◉	◎	◉	◉	◎	✓	✗	✓	The main advantage is the significant cost reductions created by short pre-determined household surveys, pre-determined scoring and statistical look up tables. However, the metric is a likelihood rather than an occurrence of poverty, meaning information produced is less actionable.
	% farmers/workers experiencing food insecurity	⚙	◎	◎	◎	◎	●	◉	◉	◎	◎	✗	✓	✓	Differs from other poverty metrics, in that food security is an outcome of various underlying causes of poverty. It is a potentially powerful metric, and baselines exist for many developing countries. However, it is complex topic and potentially expensive to collect data for.
Output	Ratio of the wages paid by certified entities compared to non-certified entities in the same region	⚙	●	◉	◉	◎	●	◎	◎	◉	◎	✗	✗	✗	Relatively easy to collect and is made context specific by having local comparators. Potentially useful in systems where waged workers are not the main poverty concern. However, it needs to be linked to an absolute measure of poverty and focuses on income only.
	Ratio of lowest entry level wage including benefits to minimum wage and benefits required by law	🏢	◉	◉	◉	◎	◉	◎	◎	◉	◎	✓	✗	✗	Relatively easy to collect, and has a clear independent baseline that is context sensitive (the national legal minimum wage). Readily communicable. Ideally would be reported gender-disaggregated. Only applies to waged labourers, which may not be applicable for all VSS.
	Percentage of producer organisations for which the lowest wage paid to general workers increased	🏢	●	●	◉	◎	◉	◎	◎	◎	◎	✗	✗	✗	Relatively easy to collect. Ideally, reporting is gender-disaggregated. However, with no external baseline, it does not allow any straightforward interpretation about the poverty status of waged workers. Best suited to be used amongst a suite of other poverty metrics (as Fairtrade do).
	Percentage of households which have made investments	⚙	◉	◉	◉	◎	●	◉	◉	◎	◎	✗	✗	✗	Potentially valuable for monitoring the 'journey' of cohorts out of poverty. However, with no external baseline, the metric is difficult to interpret. E.g., a natural disaster might lead to reduced investment, even if members of certified schemes fared better than non-certified producers.
	Percentage of households which have struggled to pay debts	⚙	◉	◉	◉	◎	●	◉	◉	◎	◎	✗	✗	✗	Potentially valuable for monitoring the 'journey' of cohorts out of poverty. However, with no external baseline, the metric is difficult to interpret. E.g., a natural disaster might lead to reduced investment, even if members of certified schemes fared better than non-certified producers.
	Percentage of households who perceive that the economic situation of their household has improved	⚙	●	◉	◉	◎	●	◉	◉	●	◎	✗	✗	✗	A relative and subjective metric that does not directly measure poverty status. However, because it is based on perceptions, it could act as a useful way of engaging producers in the issues that affect them.
	Household financial income	⚙	◉	◉	◉	◎	●	◉	◉	◎	◎	✗	✗	✗	On its own, records changes in income not poverty. Used with a baseline, becomes an input to other metrics (e.g., poverty line, Living Income).
	Perceived change of profitability of the farm	⚙	●	◉	◉	◎	●	◉	◉	●	◎	✗	✗	✗	A relative and subjective metric, not a direct measure of poverty status. However, because it is based on perceptions, it can be a useful way to engage producers on issues that affect them. E.g., Utz used this metric with non-certified comparison groups in programme evaluations.

8.3.2 Key practical trade-offs

Cost and robustness

As articulated in the previous section, a major practical consideration with outcome metrics of poverty, is that they almost inevitably require surveys of a sample of client or worker households. This makes them expensive to collect, and in some cases the analysis and data management is also likely to prove challenging for VSS (e.g., the Sustainable Livelihoods Framework, MPI). Even the output metrics require primary data collection. This could be done via certification bodies, which is not by any means a straightforward or cheap process, but will be significantly cheaper than surveys.

The trade-off between cost and robustness means that poverty outcome metrics may be better suited to evaluations rather than scheme monitoring. However, there are some factors that will make them likely to be more practical to collect and implement: where there is an existing external baseline; where data analysis and management is relatively straightforward; where communication of the results is clear (see Table 12).

Value for VSS and for clients

The ideal metric would be useful to both clients (allowing them to monitor progress and change their approach if necessary) and to VSS for reporting effectiveness and impact. With poverty metrics, the use to clients is likely to be limited (if you are poor, knowing that you are poor does not help make you any less poor), other than in a few specific instances:

- When the poverty is focused on waged workers. Depending on the VSS and/or the specific sector in which a client works, poverty may be an issue for workers. In these circumstances, output measures (such as the proportion of clients who have increased the wages of general workers) can act as a useful prompt for to clients on the specific actions they need to take. Outcome metrics, such as the poverty line or living income (or living wage) may provide higher joint value still, in that they demonstrate to the client the extent of action needed, whilst also providing the VSS with an absolute measure of their effectiveness.
- When the VSS (and or partner organisation) is intending to develop an on-the-ground poverty reduction programme to complement certification. In this case, some of the most useful metrics will be those that lend themselves to participatory approaches and which include consideration of both the incidence of poverty and its underlying causes. Examples of this include the Sustainable Livelihoods Framework, or the Household Economy Approach. These metrics are often used by international development NGOs for project planning, securing baselines and evaluations.

Box 5.

Implementing the Poverty Probability Index – based on an interview with Harveen Kour, MEL Manager, Fairtrade International

Fairtrade have invested significant resources and thought into developing poverty metrics, but this case study will focus on just one of those developments: longitudinal measurement of poverty.

Work to systematise the longitudinal measurement of poverty (i.e., measuring changes over time) started in earnest in 2015. Fairtrade has an extensive Monitoring and Evaluation system for poverty but relied on external researchers for longitudinal data. This was expensive, the data was not always comparable, and so Fairtrade decided to bring the measurement and knowledge of longitudinal poverty in house.

Box 5 cont.

The first step was to assess existing metrics, specifically the MPI and the Poverty Probability Index (PPI, see Table 14). They decided to use the PPI for a number of reasons. Firstly, the PPI is customised for each country, Secondly, the methods are fully explained and publicly available. Thirdly, "*the support available for using the PPI* [from JPAL] *is superb*".

The PPI is based on ten questions that have a statistical relationship with extreme poverty. Some of the questions appear non-intuitive (e.g., number of bath towels each member of the household owns) but they do work in the local context. The answers to each question are given a predetermined score, and the overall scores are converted into a likely incidence of poverty via reference tables. Fairtrade combine the PPI approach with additional survey questions on the acquisition of new assets, perception of socio-economic chain, and repayment of debts. Together this forms a rich data set on poverty.

The idea was to do household surveys of a sample of producer organisations and repeat the surveys with the same producer organisations every 3-4 years. They have now surveyed over 1,200 households, and after three years of implementation, the surveys are working well.

Implementation challenges. The first major challenge was that there was some resistance to measuring multi-dimensional poverty, perhaps because of Fairtrade's traditional focus on the very precise measurements of income poverty required to calculate the Fairtrade Premium. The next challenge was supporting people to understand the PPI questions: Fairtrade developed training programmes and provided real-time support to staff in the field. Cleaning and analysing the data took more effort than anticipated, particularly as staff changes meant that continuity was lost. The legal issues of who owns the data also had to be accommodated. Finally, there is pressure to release the analysis, which meant that decisions had to be made about whether to publish it all, or just those parts where they were confident that the data quality was high.

Positives. Anything involving household surveys is always costly, but the PPI requires fewer resources than traditional poverty metrics. The main uses of it will be to show changes over time, and also to pinpoint where to focus efforts. If producers are falling back in one place, having that data means that you can try to understand why, and even if the cause is outside of the control of Fairtrade, they can still provide tailored interventions to support and empower producers (e.g., training).

Looking forward. The idea has always been that local representatives could go to the field and conduct the surveys, as part of their normal schedule of visits to producer organisations. This is getting closer. Measurement started off with "paper and excel spreadsheets" and is now moving to an app. Developing the app has not been a smooth process, but it will shortly be fully operational.

Main Learnings: measure what's important: prioritise what you want to collect. Share the results as soon as possible, otherwise future investments may be reduced. Consolidate what you measure – it's difficult when so many issues are important, but Fairtrade are now consolidating their measurement approach rather than measuring different things in different places.

8.3.3 Future horizons

Although not without its flaws, the PPI – designed to significantly reduce the cost of estimating the likelihood of poverty via statistical relationships between a small number of household characteristics and poverty – hints at one potential direction for poverty metrics in the future. It is likely that more agile statistical approaches between household characteristics and poverty will be developed in the coming years, perhaps based on 'big data' analytics.

A second horizon that has already shown promise in reducing the cost of poverty data collection is the use of rapid mobile telephone surveys, which have been used in initial assessments at the beginning of humanitarian crises.[129] This type of technique is rapidly becoming more applicable as mobile phone use amongst people living in poverty becomes more widespread, as witnessed in the last decade in Asia and Africa, and opens up the possibility of more rapid and cheaper surveys of poverty.

8.4 Summary

There is no shortage of poverty metrics available, with some significant demarcations in theory and practice between them. Nonetheless, all of the metrics considered require some form of primary data collection which, on the one hand makes them responsive to context and change over time, but on the other, entails often significant effort in data collection, analysis, management and communication. This is particularly the case as the causes of poverty vary significantly between groups; survey questionnaires typically have to be adapted repeatedly to capture the poverty status of populations in different places and working in different sectors.

The key theoretical demarcations include whether income poverty is the desired focus of measurement, or if a more multi-faceted view of poverty is desired. If the former, then outcome metrics such as Living Income (or Living Wage) and poverty line are amongst the most widely applicable. If the latter, then metrics such as household consumption, the proportion of people experiencing food insecurity, and the HEA may be more appropriate. If the purpose of the metric is to also to understand the environmental, social, and economic sustainability of poverty reduction, then a metric such as the Sustainable Livelihoods Framework may be particularly useful.

Secondly, there are significant practical differences between the outcome and output poverty metrics. Outcome metrics of poverty will reveal the poverty status of people and changes in that, but without exception, require surveys of households and/or individuals. In practice, this makes the data expensive to collect, in some cases the analysis complex (e.g., the MPI) and the results do not lend themselves to clear communication (e.g., the Sustainable Livelihoods Framework). Implementation costs will be reduced where there are existing national and sub-national baselines (e.g., poverty line, MPI), which saves having to create one. Nonetheless, in the context of VSS, this perhaps lends them more to impact evaluation rather than regular progress monitoring.

The output metrics assessed here also require primary data collection, but typically from clients and via certification bodies, hence fitting in with existing systems (although not without cost to collect and analyse). Examples of the output metrics currently used by ISEAL members include: the proportion of producer organisations for which the lowest wage paid to general workers increased in the last calendar year; the proportion of households which have made investments in the last three years; the proportion of households which have struggled to repay debts in the last calendar year; the proportion of households who perceive that the economic situation of their household has improved in the last 3 years; and the perceived change of profitability of farm. However, as output metrics, none of them are able to provide information on the poverty status of the people under consideration.

9 Forced Labour

9.1 Background: Measuring forced labour

There is a widely accepted definition of Forced Labour provided by the International Labour Organisation's Forced Labour Convention[130]: "*all work or service which is exacted from any person under the threat of a penalty and for which the person has not offered himself or herself voluntarily*". The ILO further defines a set of eleven indicators for measuring cases of forced labour:

1. Abuse of vulnerability
2. Deception
3. Restriction of movement
4. Isolation
5. Physical and sexual violence
6. Intimidation and threats
7. Retention of identity documents
8. Withholding of wages
9. Debt bondage
10. Abusive working and living conditions
11. Excessive overtime

When measuring forced labour, any single indicator may indicate its existence or a combination of indicators may be needed to identify a possible case.[131]

The Forced Labour Convention has been ratified by 178 countries and the definition and indicators are frequently used to underpin labour policies and codes of conduct.

The elimination of forced labour in supply chains is a requirement of the law in many countries. For example, it is enshrined in legislation including the Modern Slavery Act (UK) and Section 1589 of the US Code[132] which '*criminalises and punishes both forced labour and those who benefit from forced labour*'. In 2016, a loophole in this Code, which allowed imports of goods into the US despite their production by forced labour if domestic production fell short of demand, was closed by the introduction into law of the Trade Facilitation and Trade Enforcement Act of 2015.[133,134]

Being able to measure forced labour is therefore critical. However, despite the wide acceptance of ILO definitions, actual measurement of forced labour is very challenging and it can often go undetected by audits for a number of reasons. Firstly, forced labour is often associated with activities that are illegal or forbidden and therefore hidden, and secondly, workers and other informants may be unwilling to discuss such sensitive issues, especially if they fear they might be punished or penalised for doing so.[135] Detection of forced labour can be particularly challenging on marine fishing vessels, where it is an acute problem, due to the difficulty of accessing ships at sea and because of loopholes in vessel registration laws which make it difficult to accurately track ships and their activities[136].

These challenges in measuring forced labour mean that many of the metrics used rely on; i) detecting proxy indicators, ii) evaluating the risk that forced labour may be occurring, or iii) measuring the presence of measures intended to address forced labour.

9.2 Overview of the metrics assessed

Table 14 provides a summary of the characteristics of forced labour metrics included here, assessed against the considerations outlined in Section 3.2. The following section draws out some of the main issues identified from this analysis.

In total, we assess twelve metrics for forced labour (Table 13). These include one outcome metric, five output metrics and six input metrics (Table 14).

9.2.1 Characteristics of the metrics

Perhaps in response to legislation that requires due diligence or outright elimination of forced labour in supply chains,[137] the majority of metrics are supply chain related. However, this should not present an insurmountable obstacle for VSS who are interested in assessing clients rather than suppliers, and indeed many of the commonly used supply chain metrics have analogues used by VSS, such as GoodWeave and Fairtrade (Table 14).

Table 11:

Description of poverty metrics

Metric	Metric description
Number of forced labour cases addressed	Good Weave measures a number of comparable metrics; number of forced laborers remediated, receiving appropriate support for their needs, or who have found appropriate alternative livelihoods to escape forced labour.
ILO Indicators of Forced Labour	Eleven indicators (Abuse of vulnerability, Deception, Restriction of movement, Isolation, Physical and sexual violence, Intimidation and threats, Retention of identity documents, Withholding of wages, Debt bondage, Abusive working and living conditions, Excessive overtime) can be used to assess whether individuals are involved in forced labour, and to calculate statistics on prevalence of forced labour within a country or a supply chain. A similar metric is used by GoodWeave which monitors the number of bonded/forced labour cases identified.
SEDEX Forced Labour Indicator Reports	A report gives a scored assessment of a business based on the degree to which 11 indicators are present in its supply chain. The indicators are based on the ILO indicators of forced labour (see above) and are scored at a site level according to the degree of risk that they are present; i) Definite, ii) Strong or iii) Possible. Assessments are based on auditor assessments and/or self-assessment questionnaires.

Table 11:

Description of poverty metrics

Metric	Metric description
Incidence of non-conformity on workers' rights and wages	Analysis of the number of reported non-conformities on workers' rights and wages, with particular focus on situations that involve forced labour (which can be defined using ILO definition/indicators). A comparable metric is already used by Fairtrade; 'number and percentage of certified organizations found to be non-compliant and/or suspended for forced labour'
Identification of areas, operations or suppliers with higher risk of forced labour	Reporting organisation provides details about operations or suppliers in their supply chain which may have a higher risk for forced labour due to the type of operation and/or the country or region e.g. as done by Rainforest Alliance and by Good Weave which maps areas of high and low risk for forced labour.
Public reporting on actions to address forced labour	Ongoing reporting of actions to address forced labour.
Number of Community Structures in place to prevent forced labour	Assessment of the existence of community structures which could help prevent forced labour, as done by Good Weave. No definition of accepted 'community structures' found in Good Weave's literature.
Public commitment to eliminating forced labour	Record of existence of a public commitment to elimination of forced labour from company's operations, such as a public Code of Conduct e.g. like that of UTZ which requires "No forced, bonded, trafficked or other involuntary labour is used at any stage of production and processing by the group or group members". Could relate to own operations, or also to those of suppliers and others in their supply chain. Could involve public reporting on progress (see below).
Existence of an accessible grievance procedure	Reporting entity proves the existence of a grievance procedure for their operations and demonstrates how this is accessible to the workforce including the process to address and resolve grievances and how decisions about reported grievances are made. A similar metric is used by Good Weave which assess the existence of Grievance Committees and whether they are meeting regularly.
Number of actors with a policy to protect workers against forced labour	Number of actors in supply chain with a policy to protect adult workers against forced labour. A comparable metric is used by Fairtrade which records the number of Hired Labour Organisations with a policy to prevent forced labour.
Reporting of management approach for forced labour	Reporting organisation provides detailed explanation of its management approach for dealing with forced labour including any policies, commitments, goals/targets, responsibilities, resources, grievance mechanisms or other specific actions. The Fairtrade certification has a comparable metric; 'Number/ Percentage of organizations with a policy and/or Integrated Management System to protect vulnerable adults for prevention of forced, bonded, or involuntary labour'
Proportion of suppliers assessed for the risk of forced labour	Records the proportion of suppliers that a company interacts or does business with that have been assessed for the risk of forced labour. Risk may be assessed using a methodology similar to the SEDEX approach, above, and a proportion of suppliers that have been assessed calculated.

9.3 Applicability and trade-offs

9.3.1 Key technical decisions in measuring forced labour

Whether to measure forced labour, the risk of forced labour, approaches to reduce its likelihood or remedial actions?

We identify four main types of forced labour metrics, those that: i) directly measure the incidence of forced labour or its indicators as defined by the ILO; ii) measure the risk that forced labour may be occurring, iii) measure the existence of policies and commitments to reduce or eliminate forced labour, and iv) measure the structures put in place to prevent or respond to forced labour.

Metrics that directly measure the incidence of forced labour or its indicators include:

- Direct **recording of ILO forced labour indicators** by auditors. This can be used to assess whether there are individuals involved in forced labour and subsequently to calculate statistics on the prevalence of forced labour within a country or a supply chain. A comparable metric is used by GoodWeave which monitors the number of bonded or forced labour cases identified.
- **Incidence of non-conformity on workers' rights and wages**. This involves analysis of the number of reported non-conformities on workers' rights and wages, with particular focus on situations that involve forced labour (which can be defined using ILO definition/indicators). A comparable metric is already used by Fairtrade, which monitors the number and percentage of certified organizations found to be non-compliant and/or suspended for forced labour.
- Direct **reporting of forced labour** is perhaps the most robust metric for forced labour, but relies on primary collection of complex qualitative data (discussed further below) (Table 14).

The advantages of this first category of metrics, is that they tend to align with the agreed definition of forced labour provided by the ILO. They also address the material impacts of forced labour, are widely used and may have an external baseline against which to assess performance. However, due to the complexities of capturing the data (discussed further below) they can be costly to implement (Table 14).

The second category of metric estimates the risk of forced labour. This is one step removed from measuring the actual occurrence of forced labour or its indicators and instead evaluates the likelihood that they may exist in a site or area of operation. This may make them slightly easier to measure, but the presence of risk does not necessarily indicate the actual presence of forced labour. Examples include:

- **SEDEX's Forced Labour Indicator Reports** (Table 14). SEDEX generate a report which is a scored assessment of the degree of risk that the ILO indicators are present in a supply chain. The indicators are assessed at a site level based on audits or self-assessment questionnaires.
- **Identification of areas, operations or suppliers with higher risk of forced labour** (Table 14). This involves the reporting organisation providing details about operations or suppliers in their supply chain which may have a higher risk for forced labour due to the type of operation and/or the country or region. This has been done by Rainforest Alliance and a similar metric is implemented by GoodWeave, mapping areas of high and low risk for forced labour.

The third type of metric measures the existence of measures and structures designed to address forced labour, including policies, codes of conduct and commitments to eliminate forced labour. This group of metrics has advantages in terms of ease of data collection and implementation, but they have the disadvantage that they do not offer any measurement of the actual occurrence of forced labour. There is relatively little evidence that measures such as code of conducts effectively address the sustainability issues they are intended to. These metrics are unable to measure any change in rates of forced labour, although they do indicate levels of commitment and action to addressing forced labour (Table 14). Examples include:

- Counting the number or proportion of **public commitments to eliminating forced labour** amongst operators in a supply chain. This involves recording the existence of a public commitment by a company to eliminating forced labour from its operations, such as a public Code of Conduct e.g. like that of UTZ which requires that "*No forced, bonded, trafficked or other involuntary labour is used at any stage of production and processing by the group or group members*". This could relate to a company's own operations, or also to those of suppliers and others in their supply chain. It could also involve public reporting on progress (Table 14).
- Other similar metrics that measure commitments to reduce or eliminate forced labour include the **number of actors with a policy to protect workers against forced labour**, and reporting of the existence of a **management approach for dealing with forced labour** (Table 14).
- The most straight-forward measure of this is a record of the **number of forced labour cases addressed**. This is, in effect, the flip side to measuring forced labour and emphasises the responsibility of organisations to act to address it. GoodWeave measures a number of comparable metrics: number of forced laborers remediated, receiving appropriate support for their needs, or who have found appropriate alternative livelihoods to escape forced labour.

The final group of metrics looks at the structures that are in place to prevent forced labour from occurring in the first place:

- **Existence of an accessible grievance procedure.** This involves the reporting entity proving the existence of a grievance procedure for their operations and demonstrating how this is accessible to the workforce. This includes the process to address and resolve grievances and how decisions about reported grievances are made (Table 14). A similar metric is used by GoodWeave which assess the existence of Grievance Committees and whether they are meeting regularly.
- The **number of community structures in place to prevent forced labour** which may lessen the risk of forced labour (Table 14). This is a measure described by GoodWeave, but note that there is no accepted definition of which 'community structures' would be included within this metric. For example, it could be the availability of schooling and paid employment opportunities in an area, or existence of committees similar to the Grievance Committees, tasked with identifying instances of forced labour and providing support to those affected.
- **Public reporting on actions to address forced labour**. This entails ongoing reporting of actions to address forced labour. The requirement for them to be made public introduces transparency as a potential lever of change. In addition, there can be opportunities for reputational benefits, particularly when reporting shows positive results (Table 14).

These metrics are likely to be easier and cheaper to measure compared to the other categories of metrics. However, they do not provide any data on the actual occurrence of forced labour, which may mean they are more vulnerable to scrutiny e.g. from consumers.

9.3.2 Key practical decisions in measuring forced labour

Cost and robustness

As with many other sustainability issues, there is a clear trade-off between the cost of using metrics that rely on the collection, management, analysis and reporting of primary data obtained from individual clients, and those that rely more on secondary data or indicators. The latter will often be cheaper to collect and implement but may also be less robust.

Metrics that directly measure the incidence of forced labour or its indicators provide the most robust measure of the actual presence of forced labour, in addition to being based on the ILO's internationally recognised indicators. They therefore stand up to scrutiny and offer the potential reputational benefit to users of demonstrating transparency and commitment to rigorous monitoring and action on the issue of forced labour. They require the collection of primary data, which will require specialised auditors or surveyors plus expertise to interpret and analyse the information – which will often be qualitative and subjective – to determine whether forced labour is present.

Some relevant data may already be collected during existing auditing processes, for example records on wages or labour contracts. This information will be particularly relevant for the metrics measuring the incidence of non-conformity on forced labour requirements or the number of cases addressed. These could therefore fit well into existing auditing and reporting processes.

However, the collection of other accurate data on forced labour can be challenging. Firstly, forced labour is often associated with hidden illegal activities, making identification and access to relevant sites and workers difficult. In addition, the indicators of forced labour can be hard to capture or observe. They will also often be hidden and not evidenced in any record keeping. Data collection therefore relies on observation, investigative and ideally impromptu site visits coupled with interviews with affected people to determine if, for example, intimidation and debt bondage are occurring. It is likely that workers and other informants may be unwilling to discuss such sensitive issues and the data collection will need to be done sensitively to avoid leading to any punishment or repercussions for respondents. Deciding whether or not the indicators of forced labour are present relies on some degree of subjective judgement. For example, defining whether the actions of an employer count as 'intimidation' will require assessors to have specialist skills for recognising forced labour.[138] However, detailed user guidance on how to measure forced labour is available (Table 14).

Table 14.

Characteristics of forced labour metrics

Scale of application

🏢 Company

⚙ Producer

🌐 Global

⛰ Landscape/ Jurisdiction

◉ High ◎ Medium ● Low

			Considerations around sustainability definition				Considerations around performance monitoring					Considerations around data use			
Metric type	**Metric description**	**Scale of application**	Metric measures the sustainability issue	Metric is underpinned by evidence	Metric addresses material impacts	Metric is widely used and easily communicable	Practical to collect	Practical to implement	Context-sensitive	Responsive to change	Scalable	Baseline data available to provide context	New tools/software emerging?	Implementation 'user package' exists	**Main issues/comments**
Outcome	Number of forced labour cases addressed	»	◎	●	◉	◉	◎	◎	◉	◉	◉	✗	✗	✗	A relatively straightforward metric of rates of forced labour which also captures efforts to address it. Can be used in addition/combination to an assessment of risk such as the SEDEX indicators. Little evidence of current use, although reportedly used by GoodWeave. Limitations include that it assumes all efforts to detect forced labour are equally effective.
Output	ILO Indicators of Forced Labour	⚑	◉	◉	◉	◉	●	●	◉	◎	◉	✓	✗	✓	Internationally-accepted standard for indicators of forced labour, but rely on the collection of primary data which is likely to be challenging and resource-intensive.
	SEDEX Forced Labour Indicator Reports	»	◉	◉	◎	◉	◎	◎	◉	◎	◉	✓	✗	✓	Comprehensive but easily communicable indicator based on international standard indicators for forced labour, but relies on collection of primary data which can be challenging due to the subject matter.
	Incidence of non-conformity on workers' rights and wages	»	●	●	◎	◉	◎	◎	◉	◉	◉	✗	✗	✗	Not a direct measure of forced labour but rights and wage violations may indicate existence of conditions of forced labour. Relies on collection/collation of sensitive and qualitative primary data, which presents a challenge.
	Identification of areas, operations or suppliers with higher risk of forced labour	🏢	◎	◎	◎	◉	◎	◎	◉	◎	◉	✓	✓	✓	Assessment to identify possible risk suppliers or operations based on type or location. Requires different use or analysis of data, but can be measured based on data that is already collected. Is an indication that risk of forced labour might exist, rather than a direct measure of forced labour.
	Public reporting on actions to address forced labour	🏢	◎	●	◎	◉	◉	◉	◉	●	◉	✗	✗	✓	Ongoing reporting on actions to address forced labour is more 'active' than a simple commitment to address forced labour and is straightforward to measure, but does not provide any information on actual presence/rates of forced labour.
Input	Number of Community Structures in place to prevent forced labour	»	●	●	◎	◉	●	●	◉	●	◎	Partly	Partly	Partly	Indirect input or process indicator based on collection of qualitative data in communities where operation activities are occurring. May be possible to link with measurement of poverty or other social metrics.
	Public commitment to eliminating forced labour	🏢	●	●	◉	◉	◉	◉	◉	●	◉	✗	✗	✗	Easy to show whether a public commitment to eliminating forced labour exists, but does not measure or communicate anything about actual presence or rates of forced labour.
	Existence of an accessible grievance procedure	»	●	●	◎	◉	◉	◉	◉	●	◉	✗	✗	✗	Easy to show whether an accessible grievance procedure exists, but does not measure or communicate anything about actual presence or rates of forced labour.
	Number of actors with a policy to protect workers against forced labour	»	●	●	◎	◉	◉	◉	◉	●	◉	✗	✗	✓	Straightforward measure of whether actors in a supply chain have a policy against forced labour, but does not provide any information on actual incidence or rate of forced labour.
	Reporting of management approach for forced labour	»	●	●	◎	◉	◉	◎	◉	●	◉	✗	✗	✗	Captures a number of different measures a company may be taking to address forced labour (e.g. code of conduct, different internal performance indicators etc), requires new collation and interpretation of information. Does not provide any measure of actual presence/rate of forced labour.
	Proportion of suppliers assessed for the risk of forced labour	»	●	●	●	◉	◎	◎	◉	●	◉	✗	✗	✓	Fairly high level and straightforward reporting on whether or not a supplier has been assessed for forced labour risk. Data may require time and resources to collate from all suppliers. Does not provide any measure of actual incidence of forced labour.

Metrics used to measure past incidence of forced labour, for example incidences of non-compliances and number of cases addressed, are by comparison, often easier and less expensive metrics to collect and analyse. The data can even be gathered and reported during standard certification audit processes. However, incidences in which forced labour has been found and recorded are likely to be under-reported compared to the actual rate of forced labour occurring. This is not least because it will be in an actor's interest to hide any instances of forced labour in its operations and avoid punishment for breaching laws and international agreements. An organisation may also choose not to record a case that it has addressed, and auditors typically find forced labour hard to detect.

Metrics which estimate the risk of forced labour in regions, operations or suppliers (i.e., one step removed from measuring forced labour or its indicators itself) rely more on existing or secondary data, or a combination of primary and secondary data (e.g., SEDEX). As a consequence, they are likely to be less resource intensive than those metrics that rely solely on the collection of primary data. However, risk indicators are likely to be less responsive to change in the rate of forced labour and, as ever with risk indicators, low risk does not mean no risk. In addition, metrics which map the risk of forced labour, based on operations being located in high-risk regions, are reliant on low-resolution national-level data or assessments. They are therefore not specific to the actual operations within the country and not responsive to any changes that might have been made in management. Metrics such as the SEDEX reports also rely on data collected via self-assessment questionnaires. This risks that the assessment is done by individuals within organisations who are not specifically skilled in recognising forced labour and/or will have an incentive to under-report in order to present their organisation in a positive light. Reducing this risk of self-assessments requires the expertise and resources to collate and interpret risk data. This is not a trivial investment for operations. However, once a system has been established it could be updated relatively easily and risk assessments can be useful to identify priority areas, where subsequently more expensive direct measurement metrics could be implemented.

Recording the existence of commitments to reduce or eliminate forced labour requires less intensive data collection, management and analysis. It will often be possible to capture the necessary information from company reporting and/or during certification auditing processes. However, this type of metric is likely to be more open to scrutiny or criticism as they do not show stakeholders the actual rates of forced labour or even the likelihood or risk of its occurrence within a supply chain or an actor's operations. Being able to demonstrate that action being taken to address forced labour may be seen as positive, but not being able to show that you have records of the actual rate of forced labour associated with your operations could be seen as a lack of transparency. This type of indicator could be strengthened by including some assessment of the quality or effectiveness of the measures to address forced labour, for example; the level detail of the written content, the strength and ambitiousness of any targets, 'bindingness' of the commitment(s), repercussions for not following, etc. Even so, the metric would remain far-removed from measuring forced labour or its indicators directly.

The last type of metric – recording the number of instances of on-the-ground measures to prevent or redress forced labour – again measures positive action against forced labour rather than forced labour or its indicators. Nonetheless, it is likely to be relatively easy to collect through audit processes and is straightforward to analyse.

Shared value for clients and VSS

Clients will often prefer to have their efforts to address forced labour acknowledged by the metrics they use so that they can report positive progress on addressing the issue, for example, through having grievance procedures and remediation mechanisms in place. The more 'investigative' approaches (e.g., measuring the incidence of forced labour indicators) may be less attractive to clients. Showing progress over time on preventing and responding to forced labour is also likely to be an attractive option for VSS, particularly those operating in sectors and geographies where forced labour is endemic and requires long-term solutions (e.g., cocoa in Côte d'Ivoire[139]). However, some stakeholders may want to see proof of elimination of forced labour, in which case, metrics that directly measure the incidence or indicators of forced labour may be desirable instead of measurements of preventative action.

9.3.3 On the horizon

Recent efforts on measuring and tackling forced labour include investigations into using Blockchain software which collates automated, permanent and unchangeable records of transactions occurring along supply chains. Blockchain is designed to create greater transparency and traceability in global sourcing. For forced labour, this could include monitoring contracts and labour agreements. For example, Coca-Cola is investigating the use of 'smart contracts' to create transparent records of agreements between parties in their supply chains[140]. It can also be used to measure accountability and record-keeping quality.

However, as with other metrics discussed above, measuring the presence or absence of fair labour agreements is a proxy metric and does not directly capture the occurrence of forced labour. It is also important to consider who is able to create and validate records within Blockchain and whether workers themselves are able to access the relevant systems and accurately report their working conditions. It requires access to computer technology, an internet connection and expertise to use the software, which may be a challenge for smaller suppliers, especially in regions that are less technologically developed[141]. It also relies on the respondents being free to report accurately, which will be hindered in situations where intimidation and coercion of workers is occurring.

Box 6.

Implementing Forced Labour Metrics – based on an interview with Steph Velez, Sainsbury's Supermarkets Ltd.

Monitoring forced labour is challenging for any Fast Moving Consumer Goods retailer. The challenges include the numerous and complex supply chains; limited understanding of the issues within the company and amongst suppliers; and the difficulty of detecting it (victims don't always identify as victims, and traffickers are increasingly sophisticated, "always one step ahead"). To make matters more complicated, NGOs and investors ask for a wide range of different metrics on forced labour, which requires a company response. The UK's Modern Slavery Act requires many UK companies to publish a Modern Slavery Statement, but few companies are able to report good metrics. The most common one is the number of people trained, but how good is the training and how effective is it at increasing the detection of forced labour and improving the responses to it?

A significant focus of the company has been on getting better at detecting the risk of forced labour. They invested in a proprietary Modern Slavery Risk Assessment Tool, which overlays product location and spend data with country and sectoral forced labour risk, using an array of indicators. This highlighted some high-risk products that hadn't been considered a risk before: canned and packaged food products such as stocks, sauces and stir fries due to the multiple spices used and general merchandise products in the metal, glass and wood production industries as well as confirming

Box 6 cont.

expected risks such as coffee and fish. The company reports on some of these risks in its annual Modern Slavery Statement.

The Modern Slavery Risk Assessment Tool is complemented by a range of other approaches and data. SMETA audits are used where direct site risk is high according to the Sedex risk rating. Although social audits do not typically identify forced labour, the company reports the number of non-conformances found each year under ‘harsh and in-humane treatment’ and ‘employment is not freely chosen’. Cases of child labour have been identified through these audits and are reported on, alongside cases identified directly by its suppliers. However, the really important aspects are training and collaboration. Training, so that people within the company and its suppliers have a better understanding of the forced labour indicators and are better equipped to identify the risk of forced labour. Collaboration because the company needs to develop good tools (e.g. a progress reporting tool developed with Stronger Together is used to identify risk and also used to develop a metric), but critically, the company needs someone on the ground to call if a risk of forced labour is detected. If they detect a potential case in the UK, they can be reasonably confident that that person will be safe, but they need partners to work with in other countries where they have no presence and where the incident could be several tiers removed from them in the supply chain. The company has a wide range of collaborations, including with the Ethical Trading Initiative, Verité, Stronger Together, Issara Institute and FNET.

One of the key enablers behind their current approach was the UK’s Modern Slavery Act 2015. Forced labour was already an issue the business focused on, but the introduction of the Act drove the issue up to Board level. This resulted in more investment (e.g. resources to commission the Modern Slavery Risk Assessment tool) and a greater focus on it.

9.4 Summary

There is a widely accepted definition of forced labour and a number of related indicators but it is challenging to measure in practice. Although it is possible to measure forced labour or its indicators directly, doing so is likely to be costly to implement and relies on collection and analysis of data which requires specialist skills. Evidence can also be difficult to capture due to the hidden nature of forced labour. Metrics which measure the past incidence of forced labour (non-compliances and redressed cases) are an alternative direct measure of forced labour, but are likely to be under-reported compared to the actual rates of forced labour. Risk-based metrics do not measure forced labour directly and suffer from the generic issue that all risk-based metrics possess; that low risk does not equate to no forced labour. They may also be unresponsive to the actions of individual clients if there is a heavy reliance on national-level data in the risk analysis, although they can be used to subsequently target more costly direct measurements. Measures of on-the-ground actions to prevent or address forced labour do not measure its incidence, but have the advantages of being more cost-efficient and focusing on positive actions to reduce forced labour. Metrics that report commitments to reduce or eliminate forced labour if used alone provide no measure of the incidence of forced labour, or that forced labour is being dealt with. They can, however, be a useful starting point in long, complex supply chains where more direct measurement becomes extremely challenging. They are therefore perhaps best thought of as part of a hierarchical system of forced labour metrics, where they can be used to prioritise where more cost-intensive outcome measurements are needed.

10 Summary and conclusions

The research is focused on six critical sustainability issues: deforestation, biodiversity, water use, forced labour, poverty, and GHG emissions. The aim of the research is to document a range of leading metrics and indicators and commensurate data sources, focusing on the best practices, limitations and trade-offs associated with their use.

This work is intended to support standard systems and other sustainability initiatives which use metrics to measure these sustainability issues and want to move towards more data-driven outcome claims, whilst also providing insight into the process of considering and understanding the suitability of certain metrics. As it has drawn on experiences across sectors and a range of sustainability issues, we are able to highlight common challenges and emerging solutions across this spectrum of sustainability issues.

10.1 The growing need for robust sustainability metrics

Many modern supply chains are highly fragmented, taking place through numerous transactions, actors and geographies.[142] The main driver for the increasingly complex and fragmented supply chains is a cost cutting logic,[143] but also results in deficits of information about how something is produced and trust that it will be produced ethically and sustainably. Voluntary Sustainability Standards are one of the few viable mechanisms for reducing the risk of unsustainable practices in many complex supply chains, whilst also contributing to positive change. From this perspective, supply chain actors need to know that risk is actually reduced, and that positive change is occurring.

The same need for evidence of performance applies to the NGOs and donors that support VSS, and many other stakeholders. For VSS themselves, the questions are perhaps subtly different – 'how do we know we're achieving what we want to achieve, and how can we improve?' – but still require performance monitoring as one of a range of inputs in order to answer this.

10.2 Choosing a metric

Choosing an appropriate thing to measure and a suitable way of measuring it – a metric – is one of the key early steps in developing a way of better understanding sustainability performance. Whilst some metrics are objectively better or worse, for example based on the extent to which the metric actually measures the sustainability issue, there are a series of more complex choices.

The first choice concerns the definition of the issue. Only two of the sustainability issues included within this study have relatively uncontested definitions: GHG emissions and forced labour. Within biodiversity, water use, poverty and deforestation, multiple working definitions or theoretical constructs exist, that have spawned a range of organisational targets and commitments and supporting metrics.

The purpose of this research has not been to pass judgement on which underlying definitions of these sustainability issues are better or worse – if such a thing is even possible – instead, we have summarised the main differences and the practical consequences of these. We would suggest that VSSs consider which definition best aligns to their organisational goals, ethos and to the expectations of their stakeholders, individually or collectively under the ISEAL umbrella.

The subsequent decisions involve a series of trade-offs. The most common trade-offs are between cost of data collection, analysis, management and communication and a range of other attributes: whether the metric reflects an outcome, its robustness, the degree to which it reflects different contexts and changes over time, etc. The metrics assessed show one common characteristic not confined to sustainability: High quality information that is responsive to change over time and context and which is able to reflect the outcomes of management practices generally requires significant effort to collect, analyse, manage and report. Metrics that measure outputs and inputs are generally cheaper to collect (e.g., through the audit process) but typically do not allow clear statements about environmental and social change. Although numerous metrics are included that are heavily based on secondary data and modelling, and hence often require less investment to record and implement, these tend – perhaps with the exception of deforestation metrics – to be insufficiently granular to reflect the work of VSS clients which, by definition, is often outlying from the modelled normative practice. Similarly, relatively few metrics were found to have an external baseline against which VSS can demonstrate progress.

These three challenges – the cost of primary outcome data, limitations of secondary data and model-reliant metrics, and the scarcity of baselines – present a particular challenge to VSS. This is especially true for those that operate across radically different contexts with huge numbers of client organisations. This perhaps explains why the majority of metrics used by ISEAL members are input or output measures that are collected through the audit process. As far as we were able to ascertain, where outcome metrics based on primary data are used by ISEAL members, it is for the most part only in occasional impact evaluations. However, there would seem to be opportunities for shared effort to begin to address some of these issues. For example, agreeing a set of outcome measures for impact evaluation might allow consistent reporting and greater shared learning. There would potentially be significant cost saving, too, especially where a baseline needs to be generated.

Finally, a number of the case studies have highlighted the potential of a hierarchical approach to using metrics. In these approaches, the sustainability issue is initially measured at a higher level and broad scale (e.g., risks). This information is then used to prioritise where digging deeper is necessary (see Boxes 2, 5 and 6 for examples). This type of approach would seem particularly relevant for those sustainability issues with few options for measuring outcomes other than on-site surveys, such as biodiversity, poverty and forced labour. An example of current usage of a hierarchical approach within the ISEAL community is GoodWeave's combination of a risk-based metric of forced labour with various more outcome-focused metrics, such as the ILO indicators of forced labour.

10.3 Questions of scale

The metrics assessed across all sustainability issues are predominantly focused at the producer (client) level. As indicated in the summary tables, it is possible to aggregate many of these to make claims at greater scales at least in terms of the difference between the outcomes of certified client's management practices in comparison to conventional practices (e.g., 'deforestation was 10% less amongst scheme members than amongst similar non-certified organisations'). However, difference at scales larger than a producer does not equate to describing an outcome at a larger scale. For example, at a landscape or jurisdictional level, in which there may be only one or a few certified entities, the landscape-level outcome may be increased deforestation, whatever the outcome at the level of the certified producer.

Conversely, those metrics designed to operate at a level greater than the individual producer (landscapes or watersheds) rarely provide sufficiently granular data to reflect the management practices of individual certified entities. However, there may be value in aligning producer-level metrics to those that are produced at a larger spatial scale, largely for the purpose of having external, context-specific baselines against which to monitor performance.

10.4 Glimpses of the future?

The report highlights some of the recent advances, and those that are on the horizon, for each sustainability issue.

Technology is changing the way that information is collected (e.g., rapid mobile phone surveys of poverty status), the scales at which data can be collected (e.g., big data analytics to identify the risk of forced labour within jurisdictions) and the uses to which data can be put (e.g., remote sensing of deforestation moving towards being both a performance measure and a near real-time compliance tool). Technological advances will continue to open up new possibilities for VSS performance monitoring.

VSS can enable the use of many existing and future technology-enabled metrics for most (if not all) of the sustainability issues covered in this report is by ensuring that they have accurate, granular and up-to-date location data on their clients.[144] High quality location data will strategically pre-position VSS to take advantage of many of the metrics that are emerging, as remote sensing and spatial modelling becomes increasingly sophisticated. In the more immediate term, reliable and granular location data could facilitate a more hierarchical approach to measuring sustainability issues, where a 'broad and shallow' metric is used to prioritise where more expensive and labour-intensive outcome measurements are most needed (see Section 10.2). For example, fewer 'false positives' are likely to be generated by a broad-scale or risk-based measure of deforestation or forced labour if client location data is granular and reliable. It would also potentially extend the range of uses of existing metrics – for example using remote sensing measures of deforestation as a compliance tool as well as an outcome metric.

There are also other, non-technological changes that will influence the way that VSS measure and communicate performance. For example, the increasingly obvious climate crisis has shifted the debate away from emissions metrics and towards targets, progress against which can be measured by any number of rigorous metrics.

The effect of these technological and non-technological changes is that the metrics that are used in 5-10 years' time may be very different to those used today. However, the same choices about definitions and trade-offs will still need to be made.

11 Sources

1 IPCC https://www.ipcc.ch/

2 It is important to note that different versions of GWP exist that quantify the impact of emissions over different timescales e.g. 20 years, 100 years or 500 years. GWP100 (i.e. a one hundred year timescale) is generally used

3 Myhre, G., D. Shindell, F.-M. Bréon, W. Collins, J. Fuglestvedt, J. Huang, D. Koch, J.-F. Lamarque, D. Lee, B. Mendoza, T. Nakajima, A. Robock, G. Stephens, T. Takemura and H. Zhang, 2013: Anthropogenic and Natural Radiative Forcing. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change

4 UK Government (2019) Greenhouse gas reporting conversion factors www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2019

5 https://www.ecoinvent.org/

6 https://www.springer.com/journal/11367

7 GHG Protocol establishes comprehensive global standardized frameworks to measure and manage greenhouse gas (GHG) emissions from private and public sector operations, value chains and mitigation actions. https://ghgprotocol.org/

8 Cool Farm Alliance https://coolfarmtool.org/

9 For examples of the equations used to model agricultural GHG emissions see IPCC 2006 Good Practice Guidelines Volume 4 Chapter 10 Emissions from Livestock and Manure Management

10 Allen, M (2015) Short-Lived Promise? The Science and Policy of Cumulative and Short-Lived Climate Pollutants https://www.oxfordmartin.ox.ac.uk/publications/short-lived-promise-the-science-and-policy-of-cumulative-and-short-lived-climate-pollutants/

11 Greenhouse Gas Protocol: https://ghgprotocol.org/

12 Greenhouse Gas Protocol: https://ghgprotocol.org/

13 Sustainable Food Lab (n.d.). Scope 4: How to grow sustainable sourcing to measurable impacts in regions and commodity systems. http://www.sustainablefoodlab.org/wp-content/uploads/2015/12/Journey-to-Scope-4-ver-12comp.pdf

14 CO_2 Human Emissions: https://www.che-project.eu/introduction

15 Science Based Targets Initiative Tool https://sciencebasedtargets.org/step-by-step-guide/

16 Royal Society (2018). Greenhouse gas removals

17 Science Based Targets Initiative

18 GHG Protocol Scope 2 Guidance https://ghgprotocol.org/scope_2_guidance

19 https://ghgprotocol.org/blog/new-greenhouse-gas-protocol-standardsguidance-carbon-removals-and-land-use

20 SBTi https://sciencebasedtargets.org/sbt_events/net-zero-targets-in-the-corporate-sector-option-1/

21 https://coolfarmtool.org/

22 WWF (2013) Greenhouse-gas emissions from cotton farms in Warangal district of Andhra Pradesh http://coolfarmtool.wpengine.com/wp-content/uploads/2017/09/WWF_Cotton_Carbon_Emission.pdf

23 GIZ (2013) Climate Change Adaptation and Mitigation in the Kenyan Coffee Sector

24 CFT methods https://coolfarmtool.org/coolfarmtool/greenhouse-gases/

25 Cool Farming Options case study https://sustainablefoodlab.org/wp-content/uploads/2015/12/Cool-Farm-Tool.pdf

26 Costco eggs case study: http://coolfarmtool.wpengine.com/wp-content/uploads/2016/11/CFA_CaseStudy_Eggs.pdf

27 FAO (2012). Forest Resource Assessment 2015: Terms and Definitions. Food and Agriculture Organisation of the United Nations, Rome.

28 https://www.globalforestwatch.org/

29 https://www.soja3s.com/en/cargill-en/

30 For example, the Consumer Goods Forum: https://www.theconsumergoodsforum.com/initiatives/environmental-sustainability/about/our-commitments+and+achievements

31 The Accountability Framework Initiative: https://accountability-framework.org/

32 FAO (2016). Global Forest Resource Assessment 2015: How the World's Forests are Changing. Food and Agriculture Organization of The United Nations, Rome.

33 There are, of course, other caveats on claiming impact that apply to all metrics, such as being able to demonstrate that certified producers are in all other ways similar to non-certified producers, other than the fact that they are certified.

34 https://pro.globalforestwatch.org/about

35 Rainforest Alliance (2018). Rainforest Alliance Impacts Report 2018: Partnership, Learning, and Change. Available at https://www.rainforest-alliance.org/impact-studies/impacts-report-2018

36 Sasaki, N. and Putz, F.E. (2009), Critical need for new definitions of "forest" and "forest degradation" in global climate change agreements. Conservation Letters, 2: 226-232. doi:10.1111/j.1755-263X.2009.00067.x

37 FAO (2016). Global Forest Resource Assessment 2015: How the World's Forests are Changing. Food and Agriculture Organization Of The United Nations, Rome.. Note that the FAO also reports the extent of all forests, in a similar manner

38 https://trase.earth/

39 RSPO Guidance for Land Use Change Analysis. Revised version March 2017. Available at https://www.rspo.org/resources/search-result?search=land+use+change+analysis&s_lang=&s_type=&s_publication_date=

40 FSC (2012). Global certificates: types and distribution. Available at https://fsc.org/en/facts-figures

41 We use the term 'land management unit' as some producers may have more than once certified farm, forest, etc.

42 Bonsucro (2019). Outcome Report 2018. Available at https://www.bonsucro.com/our-impacts/

43 RSPO (2019). Impact Report 2019. Available at https://www.rspo.org/impact

44 ISEAL Metrics Catalogue

45 Newsom, D. & Milder, J. (2018). Rainforest Alliance Impacts Report 2018: Partnership, Learning, and Change. Available at https://www.rainforest-alliance.org/impact-studies/impacts-report-2018

46 https://www.vivideconomics.com/images/

47 Other examples of deforestation risk tools include Ecometrica, https://ecometrica.com/essential_grid/deforestation-risk-model; Verisk Maplecroft, https://www.maplecroft.com/risk-indices/deforestation-index/

48 ProForest (2017). Quantifying sustainability risks among suppliers and certificate holders Best practices and lessons learned A study commissioned by ISEAL Alliance.

49 Moray McLeish, Marc Fasel, Samuel Vionnet, Chris Brown (Olam International) & Ravi Abeywardana (October 2018). Forest Loss Risk Index: A risk-based approach to prioritize action on supply chain deforestation. Olam International.

50 Olam International (2018). Living Landscapes Policy. Available at https://www.olamgroup.com/sustainability/policies-codes-standards/living-landscapes-policy.html

51 RTRS Standard for Responsible Soy Production Version 3.0 02 of June, 2016. https://www.responsiblesoy.org/nueva-version-del-estandar-rtrs-de-produccion-de-soja-aprobada/?lang=en

52 https://accountability-framework.org/

53 https://palmoilscorecard.panda.org/

54 Rainforest Alliance (2018). Rainforest Alliance Impacts Report 2018: Partnership, Learning, and Change. Available at https://www.rainforest-alliance.org/impact-studies/impacts-report-2018

55 RSPO Guidance for Land Use Change Analysis. Revised version March 2017. Available at https://www.rspo.org/resources/search-result?search=land+use+change+analysis&s_lang=&s_type=&s_publication_date=

56 The International Convention on Biological Diversity, n.d. CBD Article 2: Use of Terms. Online at: https://www.cbd.int/kb/record/article/6872?RecordType=article

57 Englund & Berndes, 2014. How do sustainability standards consider biodiversity? Online at: http://publications.lib.chalmers.se/records/fulltext/196539/local_196539.pdf.

58 Petchey & Gaston, 2002. Functional diversity (FD), species richness and community composition. Online at: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1461-0248.2002.00339.x

59 Magurran, 1991. Ecological diversity and its measurement.

60 Convention on Biological Diversity, n.d. Aichi Biodiversity Targets. Online at: http://www.cbd.int/sp/targets/

61 Hillebrand et al, 2017. Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring. Online at: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/1365-2664.12959

62 IUCN, 2020. IUCN Red List of Threatened Species. Online at: https://www.iucnredlist.org/

63 Ladle, R.J., Jepson, P. Malhado, A.C.M, Jennings, S. and Barua, M. (2011). The causes and biogeographical significance of species' rediscovery. Journal Frontiers of Biogeography, 3(3).

64 For example, see Brown, N., Jennings, S., Wheeler, P., & Nabe-Nielsen, J., (2000). An improved method for the rapid assessment of forest understorey light environments. Journal of Applied Ecology, 37, 1044-1053; and Jennings, S.B., Brown, N.D. & Sheil, D. (1999). Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. Forestry, 72, 59-73.

65 IUCN, 2020. IUCN Green List of Protected and Conserved Areas. Online at: https://www.iucn.org/theme/protected-areas/our-work/iucn-green-list-protected-and-conserved-areas

66 IUCN, 2020. Key Biodiversity Areas. Online at: https://www.iucn.org/regions/mediterranean/our-work/mediterranean-species-programme/biodiversity-status-and-trends/key-biodiversity-areas

67 Globio, 2020. Science behind Globio. Online at: https://www.globio.info/what-is-globio/science-behind-globio

68 Club B4B+, 2019. Global Biodiversity Score: a tool to establish and measure corporate and financial commitments for biodiversity 2018 technical update. Online at: http://www.mission-economie-biodiversite.com/wp-content/uploads/2019/05/N14-TRAVAUX-DU-CLUB-B4B-GBS-UK-WEB.pdf

69 Globio, 2020. Impact on biodiversity. Online at: https://www.globio.info/what-is-globio/how-it-works/impact-on-biodiversity

70 Mace et al, 2018. Aiming higher to bend the curve of biodiversity loss. Nature Sustainability. Online at: http://macroecointern.dk/pdf-reprints/Mace_NS_2018.pdf

71 WWF & ZSL, 2020. Living Planet Index. Online at: https://livingplanetindex.org/projects?main_page_project=AboutTheIndex&home_flag=1

72 https://www.umu.se/en/research/projects/swedish-biodiversity-in-time-and-space/

73 https://onlinelibrary.wiley.com/doi/full/10.1111/eva.12661

74 http://publications.naturalengland.org.uk/publication/5850908674228224

75 Mehran, A. et al. (2017) Compounding Impacts of Human Induced Water Stress and Climate Change on Water Availability. Nature Scientific Reports 7: 6282

76 Orr, S et al. (2009) Understanding water risks a primer on the consequences of water scarcity for government

and business. WWF

77 UNEP (2010) Corporate water accounting: An Analysis of Methods and Tools for Measuring Water Use and Its Impacts

78 Pimentel, D. et al (2004) Water Resources: Agricultural and Environmental Issues. BioScience, Volume 54, Issue 10

79 Alliance for Water Stewardship Standard 2.0

80 Pfister, S. et al. (2011) Environmental Impacts of Water Use in Global Crop Production: Hotspots and Trade-Offs with Land Use. Environ. Sci. Technol. 2011, 45, 5761–5768

81 Cosgrove, W. J., and D. P. Loucks (2015), Water management: Current and future challenges and research directions, Water Resour. Res., 51, 4823–4839

82 The Alliance for Water Stewardship defines catchments as: "The area of land from which all surface runoff and subsurface waters flow through a sequence of streams, rivers, aquifers, and lakes into the sea or another outlet at a single river mouth, estuary, or delta; and the area of water downstream affected by the site's discharge. Catchments, as defined here, include associated groundwater areas and may include portions of water bodies (such as lakes or rivers). In different parts of the world, catchments are also referred to as watersheds or basins (or sub-basins)."

83 Fugle, K. et al (2016) G20 MACS White Paper: Metrics of Sustainable Agricultural Productivity

84 Alliance for Water Stewardship Standard Version 2.0

85 Brisbane Declaration (2007). "The Brisbane Declaration: environmental flows are essential for freshwater ecosystem health and human well-being," in 10th International River Symposium

86 Sears, L. et al (2018) Jevons' Paradox and Efficient Irrigation Technology. Sustainability, 10, 1590

87 Jevons' paradox. Ecological Economics. Volume 54, Issue

88 Orr, S et al. (2009) Understanding water risks a primer on the consequences of water scarcity for government and business. WWF

89 UNESCO (2003) Water for People, Water for Life.

90 European Environment Agency (2019) Water use and environmental pressures. https://www.eea.europa.eu/themes/water/european-waters/water-use-and-environmental-pressures/water-use-and-environmental-pressures

91 CEO Water Mandate (2014) Driving Harmonization of Water Stress, Scarcity, and Risk Terminology

92 Water Governance Facility URL: https://www.watergovernance.org/governance/what-is-water-governance/

93 Pacific Institute (2017) Exploring the case for corporate context-based water targets

94 UNEP (2010) Corporate water accounting: An Analysis of Methods and Tools for Measuring Water Use and Its Impacts

95 GRI 303: Water and Effluents 2018

96 Pfister, S.; Koehler, A.; Hellweg, S. Assessing the environmental impacts of freshwater consumption in LCA. Environ. Sci. Technol. 2009, 43, 4098–4104

97 Cool Farm Tool Water Assessment https://coolfarmtool.org/coolfarmtool/water/

98 European Environment Agency (2019) Water use and environmental pressures. https://www.eea.europa.eu/themes/water/european-waters/water-use-and-environmental-pressures/water-use-and-environmental-pressures

99 UNEP (2010) Corporate water accounting: An Analysis of Methods and Tools for Measuring Water Use and Its Impacts

100 Pacific Institute (2017) Exploring the case for corporate context-based water targets

101 WWF UK http://assets.wwf.org.uk/downloads/water_risk_filter_for_the_uk_report_vfinal.pdf

102 WWF https://www.wwf.org.za/?18121/Water-Risk-Filter

103 AWS (2018) Water Stewardship in Andalusia, Spain. The case of Iberesparragal. https://mk0a4wsorgk6akjsiboq.kinstacdn.com/wp-content/uploads/2018/10/AWS_IBER_Brochure_Final.pdf

104 WWF (2019) Summary of WWF Water Stewardship Projects, focus on Noyyal-Bhavani https://d3bzkjkd62gi12.cloudfront.net/downloads/wwf_india_stewardship_web.pdf

105 WWF water risk and water stewardship summary: https://wwf.panda.org/our_work/water/water_management/

106 https://ceowatermandate.org/

107 https://pacinst.org/

108 According to IUCN, an environmental flow is "*the water provided within a river, wetland or coastal zone to maintain ecosystems and their benefits where there are competing water uses and where flows are regulated. 'Environmental flows' is a tool for allocating water among multiple, competing uses in a watershed or river basin and building consensus on allocation decisions. Adoption and implementation requires that environmental flows are incorporated into water policies and laws. These should include the need for a negotiated consensus on flow allocation among all stakeholders. Environmental flows provide the means for integrated management of river flows to meet the needs of people, agriculture, industry, energy and ecosystems within the limits of available supply and under a changing climate*". Source: https://www.iucn.org/downloads/water_briefing_eflows.pdf

109 WWF (2017) Water risk in agricultural supply chains: How well are sustainability standards covering water stewardship https://wwf.panda.org/?315290/Water-Risk-in-Agricultural-Supply-Chains

110 UN Global Compact CEO Water Mandate, Pacific Institute, CDP, The Nature Conservancy, World Resources Institute, WWF, UNEPDHI Partnership Centre for Water and Environment. 2019. Setting Site Water Targets Informed by Catchment Context: A Guide for Companies

111 PWC (2011) The true value of water: Best practices for managing water risks and opportunities

112 Explanation of context based water targets is available via the CEO Water Mandate initiative https://www.

ceowatermandate.org/files/context-based-targets.pdf

113 Conor Linstead (2018) The Contribution of Improvements in Irrigation Efficiency to Environmental Flows. Front. Environ. Sci.,

114 Sen, A (1999). Development as Freedom. Oxford University Press, Oxford, UK.

115 E.g., https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/bur/2017/October/14_Add.pdf

116 https://data.worldbank.org/topic/poverty

117 https://www.living-income.com/

118 https://www.globallivingwage.org/

119 E.g., UNDP (2019). Human Development Report 2019 Human Development Report 2019 Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century http://hdr.undp.org/en/2019-report

120 The Oxford Poverty and Human Development Initiative https://ophi.org.uk/multidimensional-poverty-index/

121 FEWS Net https://fews.net/

122 FEWS Net https://fews.net/

123 Evidence for Development http://efd.org/methods/the-household-economy-approach-hea/

124 UK Government's Department for International Development (1999). Sustainable Livelihoods Guidance Sheets 2.1. Available at https://www.ennonline.net/attachments/872/section2.pdf

125 https://policy-practice.oxfam.org.uk/our-approach/monitoring-evaluation#ec2ea376-6c14-473f-ad41-02aa6ea5dfb1

126 Deaton, A. & Zaidi, S. (2002). Guidelines for Constructing Consumption Aggregates for Welfare Analysis. The World Bank Washington D.C. Available at http://documents.worldbank.org/curated/en/206561468781153320/Guidelines-for-constructing-consumption-aggregates-for-welfare-analysis

127 Poverty Probability Index https://www.povertyindex.org/about-ppi

128 Pape, U. (2019). Why measuring poverty impacts is more difficult than simply using score cards. World Bank Blogs. Last accessed 02 June 2020 at https://blogs.worldbank.org/developmenttalk/why-measuring-poverty-impacts-more-difficult-simply-using-score-cards

129 Pape, U. (2018). Informing rapid emergency response by phone surveys. World Bank Blogs. Last accesses 02 June 2020 at https://blogs.worldbank.org/developmenttalk/informing-rapid-emergency-response-phone-surveys

130 ILO, 2017. Forced Labour Convention, 1930 (No. 29). Online at: https://www.ilo.org/dyn/normlex/en/f?p=1000:12100:0::NO::P12100_ILO_CODE:C029

131 ILO, n.d. ILO Indicators of Forced Labour. Online at: https://www.ilo.org/wcmsp5/groups/public/---ed_norm/---declaration/documents/publication/wcms_203832.pdf

132 https://accountabilityhub.org/provision/18-usc-%c2%a7-1589/

133 https://www.arnoldporter.com/en/perspectives/publications/2017/04/understanding-the-us-ban-forced-labor-goods

134 https://www.ghy.com/trade-compliance/forced-labour-enforcement-and-supply-chain-due-diligence/

135 For example, O'Rourke, D. (2000). Monitoring the monitors: A critique of PricewaterhouseCoopers labor monitoring. Available at http://www.bollettinoadapt.it/old/files/document/18107ROURKE_2000.pdf, and Clean Clothes Campaign 2005. Looking for a quick fix: How weak social auditing is keeping workers in sweatshops, Amsterdam. Available at https://www.evb.ch/fileadmin/files/documents/Konsum/Quickfix05.pdf

136 ILO, 2013. Caught at Sea: Forced Labour and Trafficking in Fisheries. Online at: https://www.ilo.org/wcmsp5/groups/public/---ed_norm/---declaration/documents/publication/wcms_214472.pdf

137 For example, in the US, forced labour incurs penalties and the TVPA 2000 requires mandatory restitution for victims of forced labour (see https://accountabilityhub.org/country/the-united-states/). For imports, companies are required to provide paperwork to US Customs and Border Protection to show that goods were not produced using forced labour (https://www.cbp.gov/sites/default/files/assets/documents/2017-Jan/170103_Forced%20Labor%20Importer%20Due%20Diligence%20Fact%20Sheet.pdf). Since the 2016 update of the TFTEA (see above), the US Customs and Border Protection has issued a number of orders denying importation of foods into the US e.g. cotton from Turkmenistan (https://www.lexology.com/library/detail.aspx?g=085ff7c6-a334-4af7-853b-48e3282cf102)

138 Verité, n.d. Research on indicators of forced labour: Successes, Challenges and Reflections on Future Engagement. Online at: https://www.verite.org/wp-content/uploads/2016/11/Lessons-Learned-During-Research-on-Indicators-of-Forced-Labor-in-the-Production-of-Goods-v2.pdf

139 Verité, n.d. (2019). Assessment of Forced Labor Risk in the Cocoa Sector of Côte d'Ivoire. Verité, Inc., Massachusetts, USA

140 https://www.investopedia.com/news/cocacola-and-us-state-dept-fight-forced-labor/

141 https://theconversation.com/blockchain-can-help-break-the-chains-of-modern-slavery-but-it-is-not-a-complete-solution-115358

142 Gereffi, G., Humphrey, D. & Sturgeon, T., (2005). The governance of global value chains, Review of International Political Economy 12:1, 78–104

143 Bitran, G.R., Gurumurthi, S. & Sam, S.L. (2006). Emerging Trends in Supply Chain Governance Paper 227. MIT Sloan School of Management.

144 The ISEAL Alliance's 'Certification Atlas' work represents a significant step in this direction. Visit the public page for more info: https://www.isealalliance.org/innovations-standards/innovations-projects/iseal-certification-atlas-putting-certified-sites-map Also for ISEAL members visit: https://community.isealalliance.org/online-community/projects/the-certification-atlas-project-hub

12 Acknowledgements

We are grateful to the following for sharing their insights on implementing sustainability metrics within their organisations:

Chris Gerrard, Chris Gerrard, Anglian Water Anglian Water

Harveen Kour, MEL Manager, Fairtrade International

Conor Linstead, Freshwater Specialist, WWF-UK

Daniella Malin, Senior Program Director, Agriculture & Climate at Sustainable Food Lab

Moray McLeish, Vice President, Corporate Responsibility, Olam International

Stephanie Velez, Sustainable Sourcing Manager, Corporate Responsibility and Sustainability, Sainsbury's Supermarkets Ltd

We also thank Robert Lefebure and Adam Romo of the ISEAL Alliance for their constant support and input throughout this project.

The work was funded by a grant to ISEAL from the German Federal Ministry for Economic Cooperation and Development (BMZ) through GIZ and the Walton Family Foundation.

Supported by

Federal Ministry
for Economic Cooperation
and Development

Implemented by

giz Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

The WALTON FAMILY FOUNDATION