Science of the Total Environment 747 (2020) 141006

Contents lists available at ScienceDirect

Science of the Total Environment

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

Ensembles of ecosystem service models can improve accuracy and indicate uncertainty

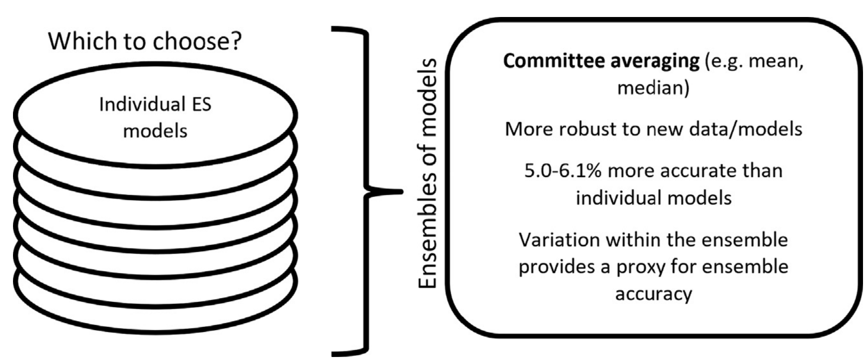
Simon Willcock [a,b,*,1], Danny A.P. Hooftman [c,d,1], Ryan Blanchard [e], Terence P. Dawson [f], Thomas Hickler [g,h], Mats Lindeskog [i], Javier Martinez-Lopez [j,k], Belinda Reyers [l,m], Sophie M. Watts [b], Felix Eigenbrod [b,n], James M. Bullock [d]

[a] School of Natural Sciences, Bangor University, United Kingdom
[b] Biological Sciences, University of Southampton, United Kingdom
[c] Lactuca: Environmental Data Analyses and Modelling, the Netherlands
[d] UK Centre for Ecology and Hydrology, Wallingford OX10 8BB, United Kingdom
[e] Council for Scientific and Industrial Research, South Africa
[f] Department of Geography, King's College London, United Kingdom
[g] Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Germany
[h] Department of Physical Geography, Goethe University, Frankfurt, Germany
[i] Department of Physical Geography and Ecosystem Science, Lund University, Sweden
[j] Soil Erosion and Conservation Research Group, CEBAS-CSIC, Spanish Research Council, Campus de Espinardo, Murcia E-30100, PO Box 164, Spain
[k] BC3 – Basque Centre for Climate Change, 48940 Leioa, Spain
[l] Future Africa, University of Pretoria, Private bag X20, Hatfield 0028, South Africa
[m] Stockholm Resilience Centre, Stockholm University, Stockholm SE-10691, Sweden
[n] Geography and Environment, University of Southampton, United Kingdom

HIGHLIGHTS

- Most ecosystem service (ES) models are uncertain.
- Still, most ES studies use only a single modelling framework.
- Ensembles of ES models are more robust to new data/models.
- Ensembles of ES are 5.0–6.1% more accurate than individual models.
- Variation within the ensemble provides a proxy for ensemble accuracy.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 April 2020
Received in revised form 14 July 2020
Accepted 14 July 2020
Available online 25 July 2020

Editor: Paulo Pereira

ABSTRACT

Many ecosystem services (ES) models exist to support sustainable development decisions. However, most ES studies use only a single modelling framework and, because of a lack of validation data, rarely assess model accuracy for the study area. In line with other research themes which have high model uncertainty, such as climate change, ensembles of ES models may better serve decision-makers by providing more robust and accurate estimates, as well as provide indications of uncertainty when validation data are not available. To illustrate the benefits of an ensemble approach, we highlight the variation between alternative models, demonstrating that there

* Corresponding author at: School of Natural Sciences, Bangor University, United Kingdom.
E-mail addresses: s.willcock@bangor.ac.uk (S. Willcock), danny.hooftman@lactuca.nl (D.A.P. Hooftman), RBlanchard@csir.co.za (R. Blanchard), terry.dawson@kcl.ac.uk (T.P. Dawson), thomas.hickler@senckenberg.de (T. Hickler), mats.lindeskog@nateko.lu.se (M. Lindeskog), jmartinez@cebas.csic.es (J. Martinez-Lopez), belinda.reyers@su.se (B. Reyers), F.Eigenbrod@soton.ac.uk (F. Eigenbrod), jmbul@ceh.ac.uk (J.M. Bullock).
[1] Joint first authors (contributed equally).

https://doi.org/10.1016/j.scitotenv.2020.141006

Keywords:
Africa
Carbon
Charcoal
Firewood
Grazing
Model validation
Natural capital
Poverty alleviation
Sustainable development
Water

are large geographic regions where decisions based on individual models are not robust. We test if ensembles are more accurate by comparing the ensemble accuracy of multiple models for six ES against validation data across sub-Saharan Africa with the accuracy of individual models. We find that ensembles are better predictors of ES, being 5.0–6.1% more accurate than individual models. We also find that the uncertainty (i.e. variation among constituent models) of the model ensemble is negatively correlated with accuracy and so can be used as a proxy for accuracy when validation is not possible (e.g. in data-deficient areas or when developing scenarios). Since ensembles are more robust, accurate and convey uncertainty, we recommend that ensemble modelling should be more widely implemented within ES science to better support policy choices and implementation.

1. Introduction

Planning and implementing sustainable development approaches requires knowledge on the ecosystem services (ES; nature's contributions to people (Pascual et al., 2017)) provided in a region and how they might respond to management choices or other drivers of change (Guerry et al., 2015). Models can provide credible information where empirical data on ES are sparse, which is especially the case in many developing countries (IPBES, 2016; Suich et al., 2015). Although claims of superiority are sometimes made for specific models, independent evaluations of models have often been unable to demonstrate the pre-eminence of any individual model in terms of accuracy or other aspects of their utility (Box 1; Table SI-1-1) (Araújo and New, 2007; Willcock et al., 2019). When models are in disagreement, it is difficult for researchers or practitioners to know which model should be used to support their decision (Willcock et al., 2016). In fact, projections by alternative models can be so variable as to compromise even the simplest assessment; these results challenge the common practice of relying on one single method (Araújo and New, 2007). Put simply, decisions based on a single ES modelling framework are unlikely to be robust (Box 1) (Refsgaard et al., 2007; Walker et al., 2003).

Despite this lack of robustness, most ES modelling applications rely on a single model for each ES (Bryant et al., 2018). For example, the latest state-of-the-art ES models produced via the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) rely on single model outputs with little/no validation (Chaplin-Kramer et al., 2019). Although, few studies have explicitly validated ES models against independent datasets, there are notable exceptions (Bruijnzeel et al., 2011; Mulligan and Burke, 2005; Redhead et al., 2018, 2016; Sharps et al., 2017; Willcock et al., 2019). Willcock et al. (2019) validated multiple models for several ES, testing their accuracy against empirical data across sub-Saharan Africa. Whilst they found that more complex models (i.e. those representing more processes) were sometimes more accurate (Box 1), their results suggested it would be difficult to select a priori the most accurate of a set of models for an ES in any particular context (Willcock et al., 2019).

Box 1
Key definitions

Whilst relatively rare in the ES literature, frameworks for understanding model uncertainty can be found elsewhere in the literature (e.g. see Araújo and New (2007), Refsgaard et al. (2007), and Walker et al. (2003)). Key concepts are defined below:

- Uncertainty – Any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system (Walker et al., 2003).
- Inaccuracy – The deviation from the 'true' value (i.e. how close a modelled value is to the measured value, the latter considered 'true' (Walker et al., 2003).
- Robustness – The level of confidence in the overall patterns/conclusions derived from the model (which may be high even though quantified estimates in individual pixels are inaccurate) (Refsgaard et al., 2007).
- Model Ensemble – A collection of modelled outputs produced by running simulations for more than one set of models, initial conditions, model classes, model parameters and/or boundary conditions (Araújo and New, 2007).
- Committee averaging – A method combining models, giving each an equal weight (e.g. calculating the mean) (Araújo and New, 2007).

One solution to inter-model variation is to utilise ensembles and apply appropriate techniques to explore the resulting range of projections. Ensembles are produced by running simulations for more than one set of models, initial conditions, model classes, model parameters and/or boundary conditions (Araújo and New, 2007). For example, since the current state and processes of the system are often uncertain, small differences in initial conditions or model parameters could result in large differences in model projections (van Soesbergen and Mulligan, 2018). Similarly, different model classes (e.g. statistical models vs process-based models) might be considered competing but equally valid representations of a system, and hence worth exploring (Araújo and New, 2007). If only one model is used, conclusions are dependent on the specific assumptions of that model. If an ensemble is used, conclusions are not dependent on that one set of assumptions and parameters, hence one can consider the variation (or uncertainty) in model outcomes and might obtain a better idea of what the reality might be. Single model forecasts have been criticised due to their potential to result in a decision that imposes rigidity, which might have serious negative consequences if there is large uncertainty and inaccuracies (Araújo and New, 2007).

Whilst running ensembles of models is not the norm in ES studies (Bryant et al., 2018), this practice is commonplace in other disciplines, most famously for climate and weather modelling (Gneiting et al., 2005; Refsgaard et al., 2014). For example, in contrast to IPBES, Intergovernmental Panel on Climate Change (IPCC) publications regularly use ensembles (Collins et al., 2013). These climate change ensembles generate a consensus prediction by measuring the central tendency (e.g. the mean or median) for the ensemble of forecasts (Araújo and New, 2007). Climate change ensemble forecasts might show enhanced performance over some individual models as the averaging results in a smoothing effect, reducing the impact of idiosyncratic responses of any particular model in the area of space and time of interest (Marmion et al., 2009). In short, by averaging multiple models the signal of interest emerges from the noise associated with individual model uncertainties (Araújo and New, 2007; Knutti et al., 2010). Such, so-called, committee averaging gives equal weight to all models. The benefits of these techniques have been observed in multiple disciplines, ranging from agro-ecology (Elias et al., 2017; Refsgaard et al., 2014) and niche modelling (Aguirre-Gutiérrez et al., 2017; Crossman et al., 2012;

Grenouillet et al., 2011) to market forecasting (He et al., 2012) and credit risk analysis (Lai et al., 2006).

The level of variation within an ensemble (i.e. inconsistency among the individual models) may also be informative in itself. Lower variation within an ensemble of models may indicate increased accuracy of the ensemble mean (Puschendorf et al., 2009). Thus, ensembles may also provide an indication of uncertainty when faced with data scarcity, a potential benefit that is perhaps most pronounced in many developing countries, where data collection and model assessment efforts are least advanced (Suich et al., 2015) but reliance on ES for wellbeing is arguably the highest (Daw et al., 2011; Shackleton and Shackleton, 2012; Suich et al., 2015).

In this paper, we demonstrate that decision-making based on single ES models is not robust for large regions within sub-Saharan Africa as high variation between model estimates means that using a different model or incorporating an additional model into the decision-making process is highly likely to result in a different decision. In addition to increased robustness, we show that ensembles of ES models can provide improved accuracy over individual models, as well as an indication of uncertainty. Finally, we discuss how ensemble modelling might become standard practice within the ES community, particularly when supporting high-level policy decisions, such as in IPBES regional, global and thematic assessments used in policy and decision-making.

2. Methods

Recently we validated multiple models for each of six ES in sub-Saharan Africa (stored carbon, available water, water usage, firewood, charcoal, and grazing resources; Table 1) using 1675 data points from 16 independent datasets (Fig. SI1-1; summarised in Table SI1-2, but see Willcock et al. (2019) for further information). In that paper, we used six ES modelling frameworks (InVEST (Kareiva, 2011; McKenzie et al., 2012), Co$ting Nature (Mulligan, 2015; Mulligan et al., 2010), WaterWorld (Mulligan, 2013), benefits transfer based on the Costanza et al. (2014) values, LPJ-GUESS (Smith et al., 2014, 2001), and the Scholes models (comprising two grazing models and a rainfall surplus model) (Scholes, 1998), following Willcock et al. (2019) by using a single set of parameters for each ES per modelling framework, with each framework requiring different inputs (Willcock et al., 2019). We employed two performance metrics to calculate model accuracy in terms of each validation dataset: Spearman's ρ and mean inverse Deviance ($D^{\downarrow}$ the mean absolute distance between normalised model and validation values per data-point, inversed so that a value of 1 represents a perfect fit). Both metrics have real-world relevance, as decision-making can make use of both relative (e.g. rank order of sites or options) and absolute (e.g. the total amount or value of service delivered) values (Willcock et al., 2016), and ρ ranks locations by their relative ES values, whereas $D^{\downarrow}$ reflects the degree to which models consistently reflect absolute values in the validation dataset (Willcock et al., 2019). In the work reported here, we use the model outcomes and calculations, and validation data and methods presented in Willcock et al. (2019) (Fig. 1). This includes our approach of normalising within model variation to fall within a 0–1 scale, following Verhagen et al. (2017), which allows comparability among the different ES studied. The codes we used to do this are deposited here: https://github.com/dhooftman72/ES_Ensembles. All analyses were performed in Matlab (v7.14.0.739), with ArcGIS 10.7 used only for display purposes. $P < 0.05$ was viewed as statistically significant throughout.

2.1. Creating ensembles

To depict among-model variation per service we divided the modelled areas into km^2 gridcells – except water, which is represented in m^3 ha^{-1} per polygon. Since all models do not cover the entire study area, we recorded the number of models with valid values per gridcell. For every gridcell where ≥3 modelled estimates were available, we calculated model ensembles and mapped the standard error of the mean (SEM) among normalised model values.

As described above, ensembles are created by combining individual model outputs, resulting in a smoothing effect whereby the individual model uncertainties are cancelled out and the signal of interest emerges (Araújo and New, 2007; Marmion et al., 2009). However, there are multiple ways by which individual models can be combined into an ensemble. For example, all models could be weighted equally (i.e. committee averaging) or weighted by some measure of reliability or trust. Here, we used committee averaging, but see SI3 for a further exploration of weighting. First, we created committee two ensemble values for each ES by calculating the arithmetic mean and median across the i individual model estimates for each modelled spatial data point (i.e. 1 km^2 grid cell). To evaluate ensemble accuracy, we compared the ensemble estimate (E) to the validation data for that spatial location as described in Willcock et al. (2019).

2.2. Comparing ensembles estimates

To evaluate if the accuracy of the ensemble is an improvement on the accuracy of individual models (Willcock et al., 2019), we performed a comparison between the individual models and each ensemble (i.e. mean and median for each ES) using accuracy statistics Spearman's ρ and Inverse Deviance ($D^{\downarrow}$; Fig. 1). To calculate improvement percentages, Spearman's ρ was normalised using Eq. (1), resulting in a 0–1 scale.

$$\rho_i' = \left(\frac{\rho_i + 1}{2}\right) \tag{1}$$

We analysed the proportional change in accuracy (ρ and $D^{\downarrow}$) for all possible pairs of comparisons between: (i) the individual models, based on the mean accuracy statistics across the group of all possible models (described below), (ii) the different ensembles (mean/median), and (iii) the best performing model according to each validation dataset. We tested whether the accuracy of a first category ("A", e.g., the ensemble mean) was higher – "improved" – or lower than a second category ("B", e.g., the individual models). The accuracy level differed greatly across the 16 validation datasets and the different ES (Willcock et al., 2019). No among ES comparison is possible as 16 validation datasets across six ES provides too low a level of replication per ES, but normalising each ES allows comparisons across the different ES as a whole. Normalising involved dividing the accuracy of A by the accuracy of B for each validation dataset. For simplicity, we refer to the 16 resulting proportions as "improvement values", although they could indicate a loss of accuracy (values <1).

Next, we analysed whether the set of 16 improvement values differ from a normal distribution with mean of 1, using a one-sample Student's t-test (ttest-procedure in Matlab) to determine whether the accuracy of A is significantly higher or lower than B. For ensembles and best-fit models, this analysis involved a direct one-to-one comparison for each possible pair within each validation dataset (i.e. A = the best-fit model vs B = the mean/median ensemble). For individual models as a group, we used an averaging method, where we took per validation set the mean of the one-to-one comparisons between the single value of comparator A, e.g. the best model, and the set of multiple values of models for that validation set as B (Eq. (2)).

$$\left(\left(\sum_i^n \frac{A}{B_i}\right) \times \frac{1}{n}\right), \tag{2}$$

with n total of models for that validation set (i; 4–6 models depending on the service; Table 1).

This was done for each of the 16 validation sets. This averaging method allowed for a fully balanced analysis, with a single

Table 1
Overview of ecosystem service models included in this study, including all ecosystem services covered and their spatial grain (adapted from Willcock et al. (2019)). For more extensive descriptions see Willcock et al. (2019), Bagstad et al. (2013) and Peh et al. (2013).

Model framework	Description[a]	Ecosystem services currently available	Spatial grain	Ecosystem service modelled in this study
WaterWorld	An internally parameterised model of accumulated water run-off. This web-based model incorporates all data required for application.	• Water Supply	1 km^2 gridcells for continental scale calculations	Water supply
Co$ting Nature	A web-based series of interactive maps that defines the contribution of ecosystems to the global reservoir of a particular ES and its realisable value (based on flows to beneficiaries of that service).	• Biodiversity Resources • Carbon Storage & Sequestration • Recreation value • Hazard Mitigation • Water Quality • Water Supply	1 km^2 gridcells for continental scale calculations	Water supply ≈ Clean water run-off Stored Carbon ≈ above and below ground carbon
LPJ-GUESS	The Lund–Potsdam–Jena General Ecosystem Simulator model (Smith et al., 2014, 2001). LPJ-GUESS is a dynamic vegetation/ecosystem model designed for regional to global applications. The model combines process-based representations of terrestrial vegetation dynamics and land–atmosphere carbon and water exchanges in a modular framework.	• Carbon Storage & Sequestration • Nitrogen Storage & Sequestration • Water run-off	0.5 degree≈ 55.6 × 55.6 km gridcells	Water supply Woody species carbon Grazing = C3/C4 carbon
InVEST	A suite of free, open-source software models from the Natural Capital Project, used to map and value the goods and services from nature. InVEST returns results in either biophysical or economic terms.	• Carbon: Terrestrial & Coastal Storage & Sequestration • Crops: Pollination & Production • Scenic Quality, Recreation & Tourism • Fisheries: Marine & Aquaculture Habitat: Quality & Risk • Marine Water Quality • Water Quality: Nutrients and Sediment • Water Supply • Wind & Wave Energy	Any, land-use map input data depending	Water supply Carbon (above ground only)
Benefit transfer	Bespoke adaptations of Costanza et al. (2014) for the study region in $ per hectare. Benefit transfer assumes a constant unit value per hectare of ecosystem type and multiplies that value by the area of each type to arrive at aggregate totals.	• Gas regulation • Climate regulation • Disturbance regulation • Water regulation • Water supply • Erosion control • Soil formation • Nutrient cycling • Waste treatment • Pollination • Biological control • Habitat/Refugia • Food production • Raw materials • Genetic resources • Recreation • Cultural	Any, land-use map input data depending	Water yield ≈ Water supply Carbon ≈ Climate regulation value Charcoal use ≈ Raw materials value Firewood use ≈ Raw materials value
Scholes models	Interpretation of Scholes (1998).	• Grazing • Firewood • Water supply[d]	Any, input data depending	Water surplus[d] ≈ Water supply Grazing use[e] Firewood use[f]
New models[b]	Bespoke calculation of Water use per country, calculated as the sum of all run-off per country[c] divided by the full population per country as calculated from Afripop 2010 (Stevens et al., 2015)	Bespoke models made in this study from Willcock et al. (2019) All models with Water Supply above	Depending on water supply source data	Water use
	Bespoke models for carbon based services grazing, charcoal and firewood using as input the carbon stock output of the existing carbon models and adapted using multiplication factors and spatial masks (see Willcock et al. (2019) for full details).	Co$ting Nature carbon	Depending on carbon source data	Grazing use Charcoal use Firewood use
		InVEST carbon		Grazing use Charcoal use Firewood use
		LPJ-GUESS woody species carbon		Charcoal use Firewood use
		Benefit transfer carbon		Grazing use

[a] All 1 × 1 km in this study, unless otherwise noted. Willcock et al. (2019) investigated the impact of spatial scale on ecosystem service models and found no significant impact (unpublished results). Thus, spatial scales are unlikely to affect results here.
[b] These services were not modelled in these model frameworks when we conducted our model runs (in 2016). We developed new models using carbon stock outputs from existing models as input (see Willcock et al. (2019) for full details). The original models and their developers should not be held responsible for the results from these new models.
[c] Except for accumulated flow from WaterWorld which is the sum over all watersheds within countries of the maximum flow per watershed.
[d] Estimated as number of days that precipitation exceeds evapotranspiration, this service was added by the current study to the available Scholes models (Scholes, 1998).
[e] We have two Scholes grazing models in our study, a generic international model using freely available global data and a locally parameterised South African model (see Willcock et al. (2019) for full details).
[f] Modelled at a 5 × 5 km resolution.

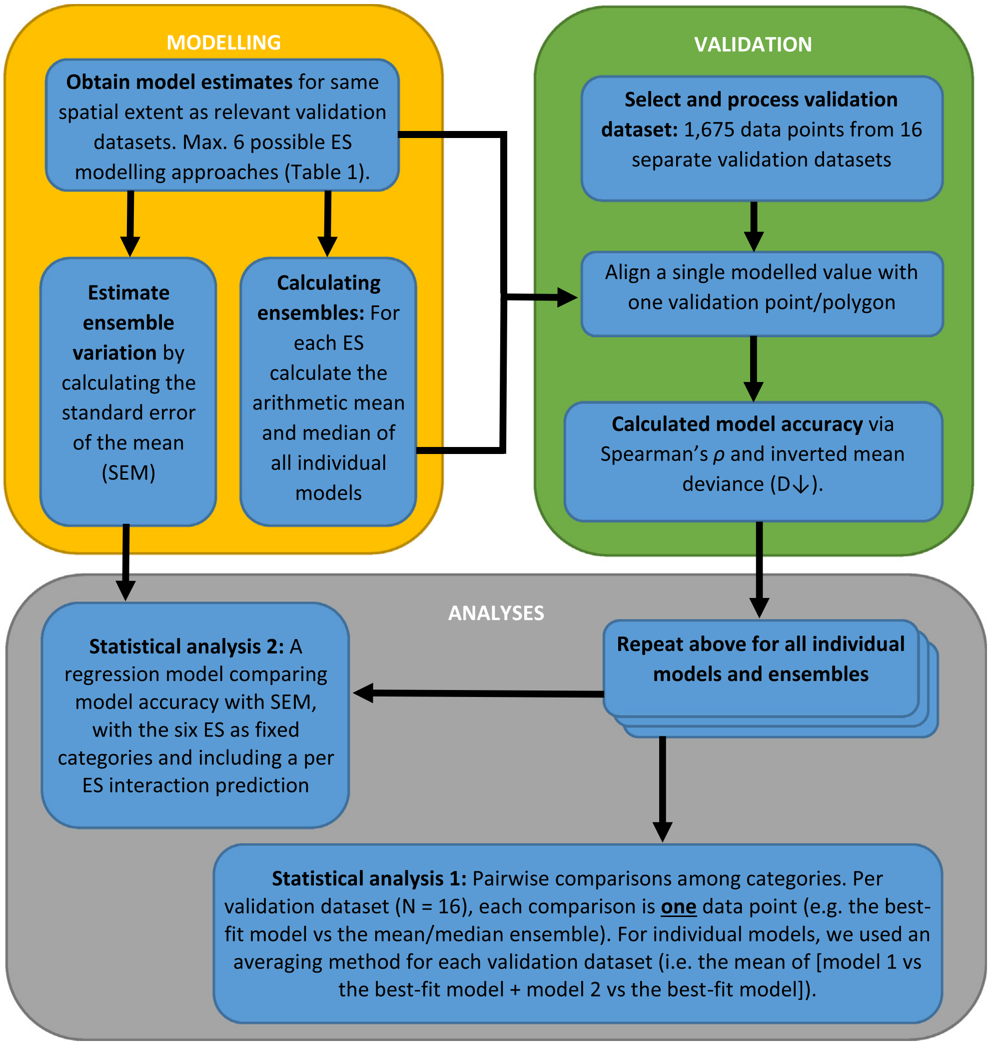


Fig. 1. A summary of the analytical framework, divided into modelling, validation and analysis subsets.

improvement value associated with each of the 16 validation datasets. Alternative analyses in which we included single comparisons for individual models per validation dataset against respective ensemble scores (79 improvement values) showed similar results (Table SI-1-4) as the larger variation was offset by higher degrees of freedom (78 vs 15).

We also tested the correlation between ensemble *uncertainty* and absolute *accuracy* using 1661 of the 1675 individual data-points for validation (anovan-procedure in Matlab). The large sample size meant we were able to differentiate between ES in this analysis. We calculated ensembles from a minimum of three models and so discarded 14 data-points since they only matched ≤2 modelled estimates. For each data-point (X), we calculated the absolute *accuracy* of the mean ensemble ($D^{\downarrow}_{(x)}$) and calculated *uncertainty* as the SEM among-modelled values (Eq. (3)). For statistical comparison, we used an SS type 1 mixed regression model with the six ES as fixed variables and SEM_X as the linear predictor, logit transformed, with correlation coefficient β_1 and constant β_0, and with a per ES interaction prediction with uncertainty (ES_X x SEM'_X). We identified a positive Spatial Autocorrelation (SA) for accuracy with a Moran's I of 0.073 ($P < 0.001$, based on a permutation test), using the Moran's module from https://github.com/dhooftman72/Morans-I. This SA has been corrected for through inclusion of a covariate within the regression model prior to estimating the model parameters of interest, with effect size β_{sa}, describing relatedness between individual samples caused by the spatial structure following Dormann et al. (2007) and Brooks et al. (2016) (Eq. (4)).

$$SEM_X = \left(\frac{\sigma_X}{\sqrt{n_X}}\right), \tag{3}$$

where *X* represents each 1 km² grid-cell, and *n* is the number of models.

$$D^{\downarrow}_{(X)} \sim \beta_{sa}SA_x + ES_X + \beta_1 SEM'_X + \left(ES_X \times SEM'_X\right) + \beta_0 \tag{4}$$

$$\text{With } SEM'_X = \left(\log_{10}\left(\frac{SEM_X}{(1-SEM_X)} + 1\right)\right).$$

3. Results

3.1. Variation among models shows strong spatial patterning

For sub-Saharan Africa, we found large areas for which the variation among models was relatively low (Fig. 2). In these areas all models provide similar normalised predictions and so a decision based on a single model may prove robust. However, there are also notable areas of disagreement, where variation among models was higher. These appear to occur in transition zones between vegetation

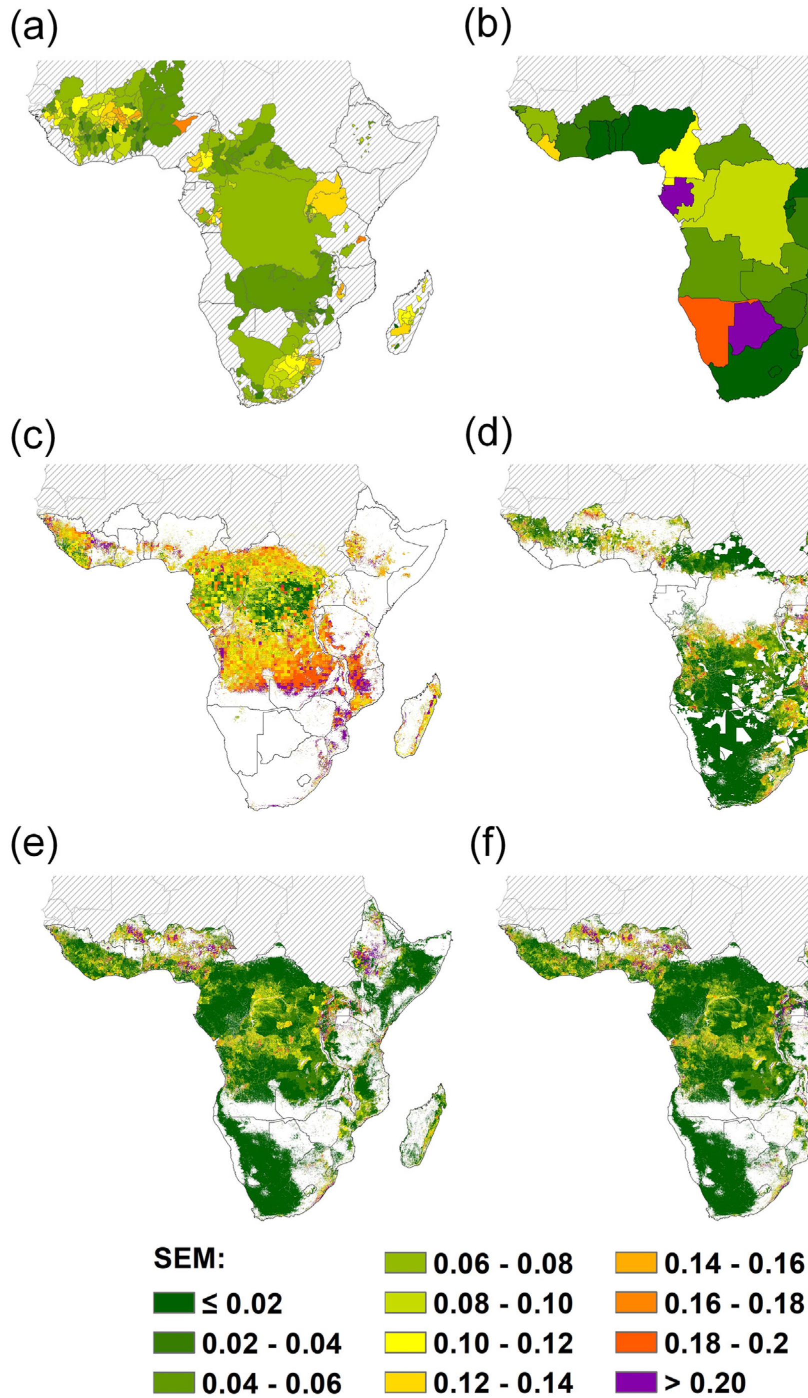


Fig. 2. Among-model variation measured as standard error of the mean (SEM) using normalised model predictions. Non-coloured areas were not modelled (i.e. are outside LCM masks or outside the catchments we analysed). a) Water supply per hectare of the catchment (6 models); b) Water usage (6 models) per hectare of the country; c) Carbon storage in forest vegetation (4 models); d) Grazing use (6 models); e) Firewood usage (5 models); f) Charcoal usage (4 models). Firewood and Charcoal have four models in common that are equal once normalised. However, Firewood contains an additional bespoke Firewood model that generates more variation making (e) and (f) slightly different (see Willcock et al. (2019) for full model details).

types (Fig. 2) and, for aboveground carbon storage models, in less densely forested areas (e.g. miombo woodland; Fig. 2). These maps of variation, as well as the mean and median normalised values, for sub-Saharan Africa at a 1-km-resolution are available through the Environmental Information Data Centre (EIDC; https://eidc.ac.uk/) repository (doi:https://doi.org/10.5285/11689000-f791-4fdb-8e12-08a7d87ad75f). See SI2 and SI3 for further uses of multiple models (i.e. hotspots, weighted ensembles).

3.2. Ensembles perform better than individual models, on average

In general, individual models as a group were inferior to the ensembles created from them: ensembles outperform individual modelling frameworks by 5% to 6% for both ρ and $D^{\downarrow}$ ($P = 0.03$ and 0.008 respectively; Fig. 3; Table SI1-3). Ensembles were outperformed by the best model for each validation set by 13% (mean; $P = 0.04$) and 12% (median; $P = 0.05$) using ρ and 6% ($P = 0.002$) and 7% ($P < 0.001$) using

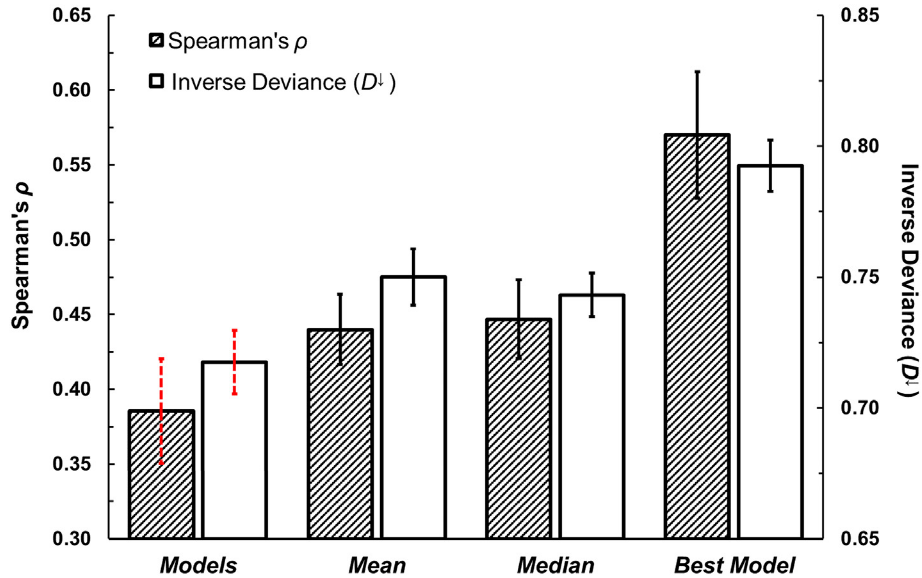


Fig. 3. Mean ρ and $D^{\downarrow}$ of the individual models (as a group), the mean and median ensembles and best-fit individual model. Dark bars = Spearman's ρ; Light bars = Inverse Deviance $D^{\downarrow}$. Black full error bars indicate variation in proportional improvement against the individual models, calculated as $SEM_{imp} = CV_{imp}$ x absolute difference, with CV the coefficient of variation of proportional improvement based on standard error of the mean (SEM). Thus, error bars indicate the variation in improvement against individual models as a group to highlight the range of improvement of ensemble techniques. N = 16 per bar. Red dashed error bars indicate the SEM among all 79 models in this study as indication of overall variation in accuracy.

$D^{\downarrow}$. Unfortunately, which model performs best for each validation dataset was hard to predict as no single model framework is consistently more accurate than others (Table SI1-1, Willcock et al. (2019)). A full matrix of statistical results and means and standard errors of these pairwise comparisons is provided in Table SI1-3.

3.3. Accuracy is correlated to ensemble uncertainty

The accuracy of an ensemble in relation to validation datasets could be in part inferred from the variation among the models within the ensemble (Fig. 4; F-value = 36.2, P < 0.001, df = 1/1637). For example, for every 0.1 increase in the SEM among-modelled values, the inverse deviance decreases by 0.054. We found no significant interaction effects among ES and uncertainty (F-value 1.09, df 5/1637) suggesting results are generalisable among the tested ES in this study.

4. Discussion

We have demonstrated that there is substantial variation between ES models and the difficulty in predicting the best-fit model as no single model was consistently better than others (Table SI1-1) (Willcock et al., 2019). These areas of disagreement highlight regions where decisions

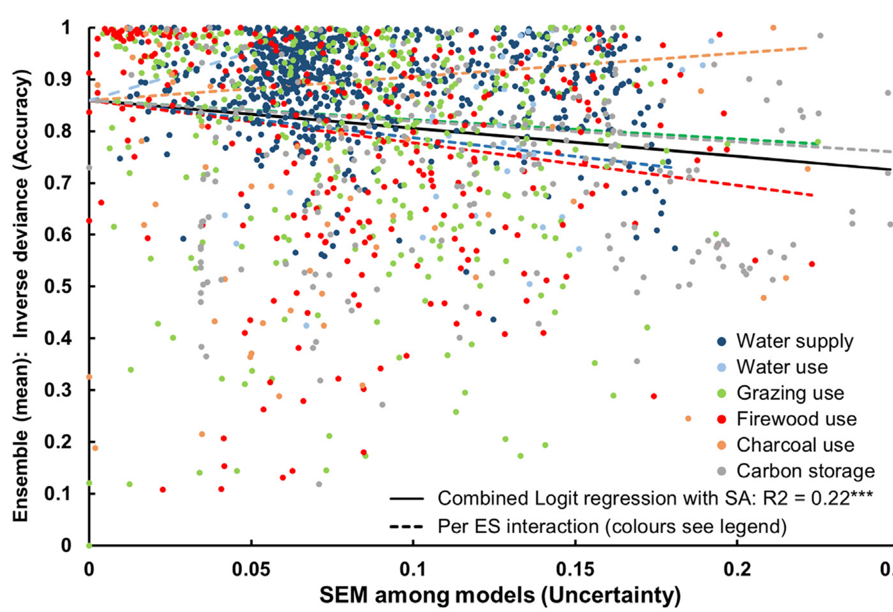


Fig. 4. Relationship between *Uncertainty* among ES models (Standard Error of the Mean of normalised values) and the *Accuracy* of the ensemble (mean) for six ES. ES-specific linear interactions are shown as dashed lines (although the interaction between ES and Uncertainty is not significant) using the same colour palette as the data points– all show a negative correlation against uncertainty, except for water use and charcoal use.

based on individual models are likely not robust (Fig. 2). For example, all ES models agreed less in transition zones between vegetation types. The majority of the models used here (and ES models generally) require input from land cover maps, and transition zones between land cover categories are likely areas of disagreement between maps. Reasons for this might include land cover maps being produced in different years and so locating the forest frontier in different places, maps/models using slightly different definitions of land cover (and so drawing the boundaries between categories in different places), or because land cover categories are more uncertain in transition zones (Dong et al., 2015), partly due to the difficulties of accounting for degradation (Turner et al., 2016). However, even if vegetation transitions are also simulated (here by a Dynamic Global Vegetation Model, LPJ-GUESS), models are more likely to disagree at a transition zone compared to the central area of a vegetation type. Furthermore, vegetation transitions and carbon storage in sub-Saharan Africa are strongly driven by fire, which is difficult to simulate in process-based models (Hantson et al., 2016). The variation between models due to different initial conditions (i.e. land cover maps) is not the focus of this paper, but has been highlighted previously (van Soesbergen and Mulligan, 2018) and can lead to large error propagation in downstream models (Estes et al., 2018). It is likely that such disagreement is also a key factor driving variation between the ES models considered here. Similarly, aboveground carbon storage models also showed disagreement in less densely forested areas (e.g. miombo woodland). Thus, these differences might partly arise due to uncertainties in the carbon data used to parameterise the models. Savanna and miombo ecosystems are understudied, with tree inventory plots showing a bias towards closed canopy forests (Phillips et al., 2002). Added to this, less densely forested areas show higher natural variation in aboveground carbon storage when compared to closed canopy forests as the land cover category definitions typically cover a wider range of canopy cover (e.g. 10–80% vs 80–100%) (Willcock et al., 2014; Willcock et al., 2012). Thus, further collection of primary data is needed, particularly in the areas of disagreement highlighted here, to improve the next generation of ES models.

Despite disagreement between individual models, ensemble modelling has been mostly neglected by the ES community; e.g. a Web of Science search (10 February 2020) for “model ensemble” and “ecosystem service” resulted in no records. This is surprising as: 1) Ensembles are commonly used for model types that simulate output variables closely related to ES, but without emphasising the ES concept in the publication, such as crop models (Rosenzweig et al., 2014), Dynamic Global Vegetation Models simulating carbon uptake (climate mitigation, e.g. Ahlström et al. (2015)) or hydrology models simulating runoff (freshwater supply).; and 2) Other disciplines have found that ensembles can show enhanced robustness and performance over some individual models as the averaging minimises the influence of local idiosyncratic responses of any particular model (Marmion et al., 2009). For example, Inoue and Narihisa (2000) demonstrated that ensemble averaging classification problems resulted in 1–7% improvements in accuracy using computational experiments and similar results are widespread in the literature; e.g. for species distribution models (Grenouillet et al., 2011; Marmion et al., 2009), climate change models (Refsgaard et al., 2014), and economic models (He et al., 2012). These findings from other disciplines mirror ours, that ensembles are around 6% more accurate than individual models (Fig. 2, Table SI1-3). That said, if the desired model output can be validated, then accuracy is increased further by identifying and using the best-fit individual model (gaining a further 12% increase in accuracy). However, using the best-fit model to support a decision does not necessarily increase its robustness as inclusion of new data or models may shift which model is thought to be most accurate (Table SI1-1) (Willcock et al., 2019).

Ensembles will likely have the highest utility when validation using primary data is not possible (IPBES, 2016). In these situations, individual model accuracy is not known, and committee ensemble methods can yield cost-effective solutions decision support tools (Araújo and New, 2007) (see SI3 for a discussion on weighted ensemble techniques). The sustainability agenda desperately requires evidence-based policies and actions for the developing world (Clark et al., 2016). In these regions, ES information is important because the rural and urban poor are often the most dependent on ES (either directly or indirectly (Cumming et al., 2014)), both for their livelihoods (Daw et al., 2011; Suich et al., 2015) and as a coping strategy for buffering shocks (Shackleton and Shackleton, 2012). As such, a single model of unknown certainty could lack credibility, relevance and legitimacy – the major reasons for the ‘implementation gap’ between ES research and its incorporation into policy- and decision-making (Cash et al., 2003; Clark et al., 2016; Wong et al., 2014). Put simply, ensemble models offer a way to reduce as well as acknowledge uncertainty (Bryant et al., 2018) but also potentially offer a future avenue to include other sources of knowledge including local and traditional knowledge in interpreting the outcomes and uncertainty of ensembles to ensure more legitimate and salient knowledge for use in decision making (Díaz et al., 2018; Pascual et al., 2017). Thus, model ensembles may be useful when estimating scenarios of future ES supply and use, but also for contemporary estimates in data deficient areas such as sub-Saharan Africa (Willcock et al., 2016). Furthermore, we suggest that variation among models can provide a first-order estimate of the quality of the prediction when no other information is available (Bryant et al., 2018; Puschendorf et al., 2009). Thus, we believe the benefits of using an ensemble of models in decision-making (increased robustness, increased accuracy over individual models in general, and the ability to estimate uncertainty) substantially outweigh the costs (reduced accuracy when compared to the best-fit model, and additional effort required).

Such ensemble modelling is now possible, as a multitude of ES models have now been developed, with many capable of being run even in data-deficient regions (Willcock et al., 2019). For example, both InVEST (https://naturalcapitalproject.stanford.edu/software/invest) and ARIES (http://aries.integratedmodelling.org/) modelling frameworks are now capable of modelling multiple ES consistently at a global scale (Martínez-López et al., 2019). As a result, for many ES, there are at least three (and often more) independent models for every location across the world. Moreover, the increasing availability of high-speed computing, and a move towards open access code using open source platforms (e.g. InVEST) makes running multiple models increasingly straightforward. Hence, it is now possible for most studies using an ES model to shift to using multiple models. We hope this study encourages ES researchers to do so.

However, whilst using ensembles of ES models is indeed possible, there are several challenges that need to be overcome before it becomes standard practice within ES science. We argue that advances are necessary in two key areas: accessibility and comparability. As more independent models are developed, it might be hypothesised that the ease with which these models can be accessed might increase. Indeed, anecdotal evidence seems to support this as, for example, InVEST historically required access to expensive ArcGIS software and ARIES required extensive computational skills to run. Accompanying the wider shift towards open science (Fecher and Friesike, 2014), InVEST now runs independently of any commercial software, where results can be mapped using open-source GIS (Bagstad et al., 2013; Peh et al., 2013) and ARIES models can be run by non-experts (Martínez-López et al., 2019). Similarly, despite models becoming increasingly complex, the computational capacity required to run some of these models has decreased as many modelling frameworks now make use of cloud-computing resources, putting less stringent requirements on the end-user (Willcock et al., 2019).

Accessing multiple ES models remains a difficult undertaking. For example, whilst the software needed to run InVEST is free, it still requires substantial GIS knowledge and many of the models within this framework are ‘data-hungry’ and therefore require access to data and substantial processing power in order to run (Willcock et al., 2019). By contrast, ARIES and Co$ting Nature store the necessary data and processing power on their servers, but therefore require high-speed internet access (Willcock et al., 2019). Furthermore, to benefit from the full Co$ting Nature model outputs (i.e. disaggregate outputs of individual services) one either needs to enter a partnership with the model owners or pay a subscription of at least 2000 GBP yr^{-1} (http://www.policysupport.org/access-costs). Thus, in order to contrast or combine, for example, carbon models across these frameworks you require access to the internet, adequate data and computational power, as well as the funds to support a model subscription fee and the extra staff time required (i.e. when compared to running a single model). Such resources are likely out of reach of many ES researchers and practitioners and so, for them, ES ensembles are an unfeasible ideal. However, this can be somewhat negated if those with access to these resources make the ensembles they are able to create freely available (e.g. as we have done so through the EIDC repository for our committee averaged ensembles and the SEM [doi:https://doi.org/10.5285/11689000-f791-4fdb-8e12-08a7d87ad75f]).

As well as the issues surrounding the feasibility of running ensembles of models, methodological limitations remain. For example, when validating any model (individual or ensembles) a reference of truth is required (Box 1). Validation data have their own intrinsic inaccuracies and so it may be good practice to validate models against more than one dataset per ES to ensure the accuracy assessment is robust (Willcock et al., 2019). Whilst we use multiple sets of validation data here (Table S-1-2), data deficiency prevented further investigations into the sources of the uncertainty we identified; e.g. running simulations to vary initial conditions (e.g. spatial scale (Hou et al., 2013)), model classes, model parameters and/or boundary conditions (Araújo and New, 2007). This is an exciting avenue for future research, which could also compare using ensembles of models to assess uncertainty with other approaches (e.g. probabilistic models (Bagstad et al., 2014; Willcock et al., 2018)). Whilst both approaches are capable of estimating uncertainty, probabilistic approaches avoid the difficulties associated with running multiple models (above) but provide little insight into model-structural uncertainty, when compared to ensembles of models (Stritih et al., 2019). Thus, future investigations should include more individual models with more varied model-structures and create ensembles using a wider variety of algorithms to deepen our current understanding.

A further outstanding issue for enabling ensemble modelling is that any comparisons or combinations of modelled outputs must involve matching like-for-like variables. This can be problematic, as, at present, a selection of models for a specific ES might, to some extent, be modelling different constructs. For example, Co$ting Nature's stored carbon model includes both below- and above-ground carbon whilst other models predict only above-ground carbon (Willcock et al., 2019). Similar issues arise when linking benefit transfer models (i.e. a valuation output (Costanza et al., 2014)) with both relative and quantitative estimates of available ES resource (i.e. T C ha^{-1}). To reduce these issues and enable like-for-like comparisons, our statistical analyses focused on relative ranking (see Willcock et al. (2019) for further details). Whilst relative rankings allow for some types of questions to be answered and so are useful to support decision-making, biophysical units are required for many sustainable development decisions (Willcock et al., 2019). For example, it is impossible to evaluate if we are operating in the safe and just operating space (Raworth, 2012) without unit estimates predicting if individuals are meeting the threshold supply of a good required to support basic needs, whilst collectively not exceeding planetary thresholds (Rockström et al., 2009). Thus, concerted effort is needed to standardise the outputs of ES models to increase the ease at which they can be compared. Such efforts are perhaps best coordinated by large, multinational organisations, and so the Ecosystems Service Partnership (ESP) or IPBES could play a central role in defining key reporting metrics, akin to the role of the IPCC in providing good practice guidance on the productions of emissions estimates (Knutti et al., 2010). Due to the large quantity and diversity of ES, this is no small challenge. However, the majority of ES modelling and mapping studies focus on relatively few ES (Willcock et al., 2016) and so these could be prioritised. Furthermore, there is potential to use this guidance to converge with other disciplines by aligning on agreed proxies/outputs required to measure and monitor the attainment of the Sustainable Development Goals (SDGs; https://sustainabledevelopment.un.org/) (Xu et al., 2020). At the very least, ES studies must validate model outputs against independent data (Willcock et al., 2019) and transparently convey the identified uncertainty to model users (Bryant et al., 2018; Kleemann et al., 2020). Such practices will increase confidence in ES science and help to reduce the implementation gap between ES models and policy- and decision-making (Cash et al., 2003; Clark et al., 2016; Voinov et al., 2014; Wong et al., 2014).

5. Conclusions

This study highlights that, in most instances, ensemble modelling may provide more robust and better estimates than using single models, as well as an indication of confidence in model predictions when validation data are unavailable. Whilst ES science is not yet ready for ensembles to become standard practice, ensemble modelling should be adopted more widely in ES modelling. In future, studies of high policy relevance (e.g. future assessments of IPBES), as well as efforts to inform decisions and track progress to sustainable development (e.g. the new Global Biodiversity Framework of the CBD and the final decade of the SDGs) would benefit from using ensembles of models.

CRediT authorship contribution statement

Simon Willcock: Conceptualization, Formal Analysis, Writing - original draft. **Danny A.P. Hooftman:** Conceptualization, Formal analysis, Writing - original draft. **Ryan Blanchard:** Writing - review & editing. **Terence P. Dawson:** Writing - review & editing. **Thomas Hickler:** Writing - review & editing. **Mats Lindeskog:** Writing - review & editing. **Javier Martinez-Lopez:** Writing - review & editing. **Belinda Reyers:** Writing - review & editing. **Sophie M. Watts:** Writing - review & editing. **Felix Eigenbrod:** Conceptualization, Writing - original draft. **James M. Bullock:** Conceptualization, Writing - original draft.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work took place under the 'WISER: Which Ecosystem Service Models Best Capture the Needs of the Rural Poor?' project (NE/L001322/1), funded by the UK Ecosystem Services for Poverty Alleviation program (ESPA; www.espa.ac.uk) and 'EnsemblES - Using ensemble techniques to capture the accuracy and sensitivity of ecosystem service models' (NE/T00391X/1). JML acknowledges the support of the Spanish Government through María de Maeztu excellence accreditation 2018-2021 (Ref. MDM-2017-0714). We thank three anonymous reviewers for their insightful comments that improved this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.141006.

References

Aguirre-Gutiérrez, J., Kissling, W.D., Biesmeijer, J.C., WallisDeVries, M.F., Reemer, M., Carvalheiro, L.G., 2017. Historical changes in the importance of climate and land use as determinants of Dutch pollinator distributions. J. Biogeogr. 44, 696–707. https://doi.org/10.1111/jbi.12937.

Ahlström, A., Raupach, M.R., Schurgers, G., Smith, B., Arneth, A., Jung, M., Reichstein, M., Canadell, J.G., Friedlingstein, P., Jain, A.K., Kato, E., Poulter, B., Sitch, S., Stocker, B.D., Viovy, N., Wang, Y.P., Wiltshire, A., Zaehle, S., Zeng, N., 2015. Carbon cycle. The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink. Science 348, 895–899. https://doi.org/10.1126/science.aaa1668.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22, 42–47. https://doi.org/10.1016/J.TREE.2006.09.010.

Bagstad, K.J., Semmens, D.J., Waage, S., Winthrop, R., 2013. A comparative assessment of decision-support tools for ecosystem services quantification and valuation. Ecosyst. Serv. 5, 27–39. https://doi.org/10.1016/j.ecoser.2013.07.004.

Bagstad, K.J., Villa, F., Batker, D., Harrison-Cox, J., Voigt, B., Johnson, G.W., 2014. From theoretical to actual ecosystem services: mapping beneficiaries and spatial flows in ecosystem service assessments. Ecol. Soc. 19, art64. https://doi.org/10.5751/ES-06523-190264.

Brooks, E.G.E., Holland, R.A., Darwall, W.R.T., Eigenbrod, F., 2016. Global evidence of positive impacts of freshwater biodiversity on fishery yields. Glob. Ecol. Biogeogr. 25, 553–562. https://doi.org/10.1111/geb.12435.

Bruijnzeel, L.A., Mulligan, M., Scatena, F.N., 2011. Hydrometeorology of tropical montane cloud forests: emerging patterns. Hydrol. Process. 25, 465–498. https://doi.org/10.1002/hyp.7974.

Bryant, B.P., Borsuk, M.E., Hamel, P., Oleson, K.L.L., Schulp, C.J.E., 2018. Transparent and feasible uncertainty assessment adds value to applied ecosystem services modeling. Ecosyst. Serv. 33, 103–109. https://doi.org/10.1016/J.ECOSER.2018.09.001.

Cash, D.W., Clark, W.C., Alcock, F., Dickson, N.M., Eckley, N., Guston, D.H., Jäger, J., Mitchell, R.B., 2003. Knowledge systems for sustainable development. Proc. Natl. Acad. Sci. U. S. A. 100, 8086–8091. https://doi.org/10.1073/pnas.1231332100.

Chaplin-Kramer, R., Sharp, R.P., Weil, C., Bennett, E.M., Pascual, U., Arkema, K.K., Brauman, K.A., Bryant, B.P., Guerry, A.D., Haddad, N.M., Hamann, M., Hamel, P., Johnson, J.A., Mandle, L., Pereira, H.M., Polasky, S., Ruckelshaus, M., Shaw, M.R., Silver, J.M., Vogl, A.L., Daily, G.C., 2019. Global modeling of nature's contributions to people. Science 366, 255–258. https://doi.org/10.1126/science.aaw3372.

Clark, W.C., Tomich, T.P., van Noordwijk, M., Guston, D., Catacutan, D., Dickson, N.M., McNie, E., 2016. Boundary work for sustainable development: natural resource management at the Consultative Group on International Agricultural Research (CGIAR). Proc. Natl. Acad. Sci. U. S. A. 113, 4615–4622. https://doi.org/10.1073/pnas.0900231108.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W.J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A.J., Wehner, M., 2013. Long-term climate change: projections, commitments and irreversibility. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Costanza, R., de Groot, R., Sutton, P., van der Ploeg, S., Anderson, S.J., Kubiszewski, I., Farber, S., Turner, R.K., 2014. Changes in the global value of ecosystem services. Glob. Environ. Chang. 26, 152–158. https://doi.org/10.1016/j.gloenvcha.2014.04.002.

Crossman, N.D., Bryan, B.A., Summers, D.M., 2012. Identifying priority areas for reducing species vulnerability to climate change. Divers. Distrib. 18, 60–72. https://doi.org/10.1111/j.1472-4642.2011.00851.x.

Cumming, G.S., Buerkert, A., Hoffmann, E.M., Schlecht, E., von Cramon-Taubadel, S., Tscharntke, T., 2014. Implications of agricultural transitions and urbanization for ecosystem services. Nature 515, 50–57.

Daw, T., Brown, K., Rosendo, S., Pomeroy, R., 2011. Applying the ecosystem services concept to poverty alleviation: the need to disaggregate human well-being. Environ. Conserv. 38, 370–379. https://doi.org/10.1017/S0376892911000506.

Díaz, S., Pascual, U., Stenseke, M., Martín-López, B., Watson, R.T., Molnár, Z., Hill, R., Chan, K.M.A., Baste, I.A., Brauman, K.A., Polasky, S., Church, A., Lonsdale, M., Larigauderie, A., Leadley, P.W., van Oudenhoven, A.P.E., van der Plaat, F., Schröter, M., Lavorel, S., Aumeeruddy-Thomas, Y., Bukvareva, E., Davies, K., Demissew, S., Erpul, G., Failler, P., Guerra, C.A., Hewitt, C.L., Keune, H., Lindley, S., Shirayama, Y., 2018. Assessing nature's contributions to people. Science 359, 270–272. https://doi.org/10.1126/science.aap8826.

Dong, M., Bryan, B.A., Connor, J.D., Nolan, M., Gao, L., 2015. Land use mapping error introduces strongly-localised, scale-dependent uncertainty into land use and ecosystem services modelling. Ecosyst. Serv. 15, 63–74. https://doi.org/10.1016/J.ECOSER.2015.07.006.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography (Cop.) 30, 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x.

Elias, M.A.S., Borges, F.J.A., Bergamini, L.L., Franceschinelli, E.V., Sujii, E.R., 2017. Climate change threatens pollination services in tomato crops in Brazil. Agric. Ecosyst. Environ. 239, 257–264. https://doi.org/10.1016/j.agee.2017.01.026.

Estes, L., Chen, P., Debats, S., Evans, T., Ferreira, S., Kuemmerle, T., Ragazzo, G., Sheffield, J., Wolf, A., Wood, E., Caylor, K., 2018. A large-area, spatially continuous assessment of land cover map error and its impact on downstream analyses. Glob. Chang. Biol. 24, 322–337. https://doi.org/10.1111/gcb.13904.

Fecher, B., Friesike, S., 2014. Open science: one term, five schools of thought. Opening Science. Springer International Publishing, Cham, pp. 17–47 https://doi.org/10.1007/978-3-319-00026-8_2.

Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., Gneiting, T., Raftery, A.E.I.I.I., W., A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098–1118. https://doi.org/10.1175/MWR2904.1.

Grenouillet, G., Buisson, L., Casajus, N., Lek, S., 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography (Cop.) 34, 9–17. https://doi.org/10.1111/j.1600-0587.2010.06152.x.

Guerry, A.D., Polasky, S., Lubchenco, J., Chaplin-Kramer, R., Daily, G.C., Griffin, R., Ruckelshaus, M., Bateman, I.J., Duraiappah, A., Elmqvist, T., Feldman, M.W., Folke, C., Hoekstra, J., Kareiva, P.M., Keeler, B.L., Li, S., McKenzie, E., Ouyang, Z., Reyers, B., Ricketts, T.H., Rockström, J., Tallis, H., Vira, B., 2015. Natural capital and ecosystem services informing decisions: from promise to practice. Proc. Natl. Acad. Sci. U. S. A. 112, 7348–7355. https://doi.org/10.1073/pnas.1503751112.

Hantson, S., Arneth, A., Harrison, S.P., Kelley, D.I., Prentice, I.C., Rabin, S.S., Archibald, S., Mouillot, F., Arnold, S.R., Artaxo, P., Bachelet, D., Ciais, P., Forrest, M., Friedlingstein, P., Hickler, T., Kaplan, J.O., Kloster, S., Knorr, W., Lasslop, G., Li, F., Mangeon, S., Melton, J.R., Meyn, A., Sitch, S., Spessa, A., van der Werf, G.R., Voulgarakis, A., Yue, C., 2016. The status and challenge of global fire modelling. Biogeosciences 13, 3359–3375. https://doi.org/10.5194/bg-13-3359-2016.

He, K., Yu, L., Lai, K.K., 2012. Crude oil price analysis and forecasting using wavelet decomposed ensemble model. Energy 46, 564–574. https://doi.org/10.1016/j.energy.2012.07.055.

Hou, Y., Burkhard, B., Müller, F., 2013. Uncertainties in landscape analysis and ecosystem service assessment. J. Environ. Manag. 127 (Suppl), S117–S131. https://doi.org/10.1016/j.jenvman.2012.12.002.

Inoue, H., Narihisa, H., 2000. Improving Generalization Ability of Self-generating Neural Networks Through Ensemble Averaging. Springer, Berlin, Heidelberg, pp. 177–180 https://doi.org/10.1007/3-540-45571-X_22.

IPBES, 2016. The methodological assessment report on scenarios and models of biodiversity and ecosystem services. In: Ferrier, S., Ninan, K.N., Leadley, P., Alkemade, R., Acosta, L.A., Akçakaya, H.R., Brotons, L., Cheung, W.W.L., Christensen, V., Harhash, K.A., Kabubo-Mariara, J., Lundquist, C., Obersteiner, M., Pereira, H.M., Peterson, G., Pichs-Madruga, R., Ravindranath, N., Rondinini, C., Wintle, B.A. (Eds.), Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, p. 348 Bonn, Germany.

Kareiva, P.M., 2011. Natural Capital: Theory & Practice of Mapping Ecosystem Services. Oxford University Press.

Kleemann, J., Schröter, M., Bagstad, K.J., Kuhlicke, C., Kastner, T., Fridman, D., Schulp, C.J.E., Wolff, S., Martínez-López, J., Koellner, T., Arnhold, S., Martín-López, B., Marques, A., Lopez-Hoffman, L., Liu, J., Kissinger, M., Guerra, C.A., Bonn, A., 2020. Quantifying interregional flows of multiple ecosystem services – a case study for Germany. Glob. Environ. Chang. 61, 102051. https://doi.org/10.1016/J.GLOENVCHA.2020.102051.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P.J., Hewitson, B., Mearns, L., 2010. Good practice guidance paper on assessing and combining multi model climate projections. In: Stocker, T., Dahe, Q., Plattner, G.-K., Tignor, M., Midgley, P. (Eds.), Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, p. 13.

Lai, K.K., Yu, L., Wang, S., Zhou, L., 2006. Credit Risk Analysis Using a Reliability-based Neural Network Ensemble Model. Springer, Berlin, Heidelberg, pp. 682–690 https://doi.org/10.1007/11840930_71.

Marmion, M., Parviainen, M., Luoto, M., 2009. Evaluation of consensus methods in predictive species distribution modelling. Divers. Distrib. 15, 59–69.

Martínez-López, J., Bagstad, K.J., Balbi, S., Magrach, A., Voigt, B., Athanasiadis, I., Pascual, M., Willcock, S., Villa, F., 2019. Towards globally customizable ecosystem service models. Sci. Total Environ. 650, 2325–2336. https://doi.org/10.1016/J.SCITOTENV.2018.09.371.

McKenzie, E., Rosenthal, A., Bernhardt, J., Girvetz, E., Kovacs, K., Olwero, N., Tof, J., 2012. Guidance and Case Studies for InVEST Users, Developing Scenarios to Assess Ecosystem Service Tradeoffs. World Wildlife Fund, Washington, USA.

Mulligan, M., 2013. WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally. Hydrol. Res. 44.

Mulligan, M., 2015. Trading off agriculture with nature's other benefits, spatially. In: Zolin, C., Rodrigues, R. de A. (Eds.), Impact of Climate Change on Water Resources in Agriculture. CRC Press.

Mulligan, M., Burke, S., 2005. Global Cloud Forests and Environmental Change in a Hydrological Context. DFID FRP Project ZF0216 Final Technical Report. p. 74.

Mulligan, M., Guerry, A., Arkema, K., Bagstad, K., Villa, F., 2010. Capturing and quantifying the flow of ecosystem services. In: Silvestri, S., Kershaw, F. (Eds.), Framing the Flow: Innovative Approaches to Understand, Protect and Value Ecosystem Services across Linked Habitats. UNEP World Conservation Monitoring Centre, Cambridge, UK, pp. 26–33.

Pascual, U., Balvanera, P., Díaz, S., Pataki, G., Roth, E., Stenseke, M., Watson, R.T., Başak Dessane, E., Islar, M., Kelemen, E., Maris, V., Quaas, M., Subramanian, S.M., Wittmer, H., Adlan, A., Ahn, S., Al-Hafedh, Y.S., Amankwah, E., Asah, S.T., Berry, P., Bilgin, A., Breslow, S.J., Bullock, C., Cáceres, D., Daly-Hassen, H., Figueroa, E., Golden, C.D., Gómez-Baggethun, E., González-Jiménez, D., Houdet, J., Keune, H., Kumar, R., Ma, K., May, P.H., Mead, A., O'Farrell, P., Pandit, R., Pengue, W., Pichis-Madruga, R., Popa, F., Preston, S., Pacheco-Balanza, D., Saarikoski, H., Strassburg, B.B., van den Belt, M., Verma, M., Wickson, F., Yagi, N., 2017. Valuing nature's contributions to people: the IPBES approach. Curr. Opin. Environ. Sustain. 26–27, 7–16. https://doi.org/10.1016/j.cosust.2016.12.006.

Peh, K.S.-H., Balmford, A., Bradbury, R.B., Brown, C., Butchart, S.H.M., Hughes, F.M.R., Stattersfield, A., Thomas, D.H.L., Walpole, M., Bayliss, J., Gowing, D., Jones, J.P.G., Lewis, S.L., Mulligan, M., Pandeya, B., Stratford, C., Thompson, J.R., Turner, K., Vira, B., Willcock, S., Birch, J.C., 2013. TESSA: a toolkit for rapid assessment of ecosystem services at sites of biodiversity conservation importance. Ecosyst. Serv. 5, 51–57. https://doi.org/10.1016/j.ecoser.2013.06.003.

Phillips, O.L., Malhi, Y., Vinceti, B., Baker, T., Lewis, S.L., Higuchi, N., Laurance, W.F., Vargas, P.N., Martinez, R.V., Laurance, S., Ferreira, L.V., Stern, M., Brown, S., Grace, J., Management, R., Vargas, H., York, N., Garden, B., International, W, 2002. Changes in growth of tropical forests: evaluating potential biases. Ecol. Appl. 12, 576–587.

Puschendorf, R., Carnaval, A.C., VanDerWal, J., Zumbado-Ulate, H., Chaves, G., Bolaños, F., Alford, R.A., 2009. Distribution models for the amphibian chytrid *Batrachochytrium dendrobatidis* in Costa Rica: proposing climatic refuges as a conservation tool. Divers. Distrib. 15, 401–408. https://doi.org/10.1111/j.1472-4642.2008.00548.x.

Raworth, K., 2012. A Safe and Just Space for Humanity: Can we Live within the Doughnut? Oxfam Discussion Paper, Oxfam, Oxford, UK

Redhead, J.W., Stratford, C., Sharps, K., Jones, L., Ziv, G., Clarke, D., Oliver, T.H., Bullock, J.M., 2016. Empirical validation of the InVEST water yield ecosystem service model at a national scale. Sci. Total Environ., 1–9 https://doi.org/10.1016/j.scitotenv.2016.06.227.

Redhead, J.W., May, L., Oliver, T.H., Hamel, P., Sharp, R., Bullock, J.M., 2018. National scale evaluation of the InVEST nutrient retention model in the United Kingdom. Sci. Total Environ. 610–611, 666–677. https://doi.org/10.1016/J.SCITOTENV.2017.08.092.

Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the environmental modelling process – a framework and guidance. Environ. Model. Softw. 22, 1543–1556. https://doi.org/10.1016/J.ENVSOFT.2007.02.004.

Refsgaard, J.C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T.A., Drews, M., Hamilton, D.P., Jeppesen, E., Kjellström, E., Olesen, J.E., Sonnenborg, T.O., Trolle, D., Willems, P., Christensen, J.H., 2014. A framework for testing the ability of models to project climate change and its impacts. Clim. Chang. 122, 271–282. https://doi.org/10.1007/s10584-013-0990-2.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, I.I.I.F.S., Lambin, E., Lenton, T.M., Scheffer, M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J., 2009. Planetary boundaries: exploring the safe operating space for humanity. Ecol. Soc. 14.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T.A.M., Schmid, E., Stehfest, E., Yang, H., Jones, J.W., 2014. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proc. Natl. Acad. Sci. U. S. A. 111, 3268–3273. https://doi.org/10.1073/pnas.1222463110.

Scholes, R.J., 1998. The South African 1: 250 000 Maps of Areas of Homogeneous Grazing Potential.

Shackleton, S.E., Shackleton, C.M., 2012. Linking poverty, HIV/AIDS and climate change to human and ecosystem vulnerability in southern Africa: consequences for livelihoods and sustainable ecosystem management. Int. J. Sustain. Dev. World Ecol. 19, 275–286. https://doi.org/10.1080/13504509.2011.641039.

Sharps, K., Masante, D., Thomas, A., Jackson, B., Redhead, J., May, L., Prosser, H., Cosby, B., Emmett, B., Jones, L., 2017. Comparing strengths and weaknesses of three ecosystem services modelling tools in a diverse UK river catchment. Sci. Total Environ. 584, 118–130. https://doi.org/10.1016/j.scitotenv.2016.12.160.

Smith, B., Prentice, I.C., Sykes, M.T., 2001. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. Glob. Ecol. Biogeogr. 10, 621–637. https://doi.org/10.1046/j.1466-822X.2001.t01-1-00256.x.

Smith, B., Wårlind, D., Arneth, A., Hickler, T., Leadley, P., Siltberg, J., Zaehle, S., 2014. Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. Biogeosciences 11, 2027–2054. https://doi.org/10.5194/bg-11-2027-2014.

Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J., Jarvis, A., Hashimoto, H., 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS One 10, e0107042. https://doi.org/10.1371/journal.pone.0107042.

Stritih, A., Bebi, P., Grêt-Regamey, A., 2019. Quantifying uncertainties in earth observation-based ecosystem service assessments. Environ. Model. Softw. 111, 300–310. https://doi.org/10.1016/j.envsoft.2018.09.005.

Suich, H., Howe, C., Mace, G., 2015. Ecosystem services and poverty alleviation: a review of the empirical links. Ecosyst. Serv. 12, 137–147. https://doi.org/10.1016/j.ecoser.2015.02.005.

Turner, K.G., Anderson, S., Gonzales-Chang, M., Costanza, R., Courville, S., Dalgaard, T., Dominati, E., Kubiszewski, I., Ogilvy, S., Porfirio, L., Ratna, N., Sandhu, H., Sutton, P.C., Svenning, J.-C., Turner, G.M., Varennes, Y.-D., Voinov, A., Wratten, S., 2016. A review of methods, data, and models to assess changes in the value of ecosystem services from land degradation and restoration. Ecol. Model. 319, 190–207. https://doi.org/10.1016/j.ecolmodel.2015.07.017.

van Soesbergen, A., Mulligan, M., 2018. Uncertainty in data for hydrological ecosystem services modelling: potential implications for estimating services and beneficiaries for the CAZ Madagascar. Ecosyst. Serv. 33, 175–186. https://doi.org/10.1016/J.ECOSER.2018.08.005.

Verhagen, W., Kukkala, A.S., Moilanen, A., van Teeffelen, A.J.A., Verburg, P.H., 2017. Use of demand for and spatial flow of ecosystem services to identify priority areas. Conserv. Biol. 31, 860–871. https://doi.org/10.1111/cobi.12872.

Voinov, A., Seppelt, R., Reis, S., Nabel, J.E.M.S., Shokravi, S., 2014. Values in socio-environmental modelling: persuasion for action or excuse for inaction. Environ. Model. Softw. 53, 207–212. https://doi.org/10.1016/J.ENVSOFT.2013.12.005.

Walker, W.E., Harremoës, P., Rotmans, J., van der Sluijs, J.P., van Asselt, M.B.A., Janssen, P., Krayer von Krauss, M.P., 2003. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. Integr. Assess. 4, 5–17. https://doi.org/10.1076/iaij.4.1.5.16466.

Willcock, S., Phillips, O.L., Platts, P.J., Balmford, A., Burgess, N.D., Lovett, J.C., Ahrends, A., Bayliss, J., Doggart, N., Doody, K., Fanning, E., Green, J., Hall, J., Howell, K.L., Marchant, R., Marshall, A.R., Mbilinyi, B., Munishi, P.K.T., Owen, N., Swetnam, R.D., Topp-Jorgensen, E.J., Lewis, S.L., 2012. Towards regional, error-bounded landscape carbon storage estimates for data-deficient areas of the world. PLoS One 7, e44795. https://doi.org/10.1371/journal.pone.0044795.

Willcock, S., Phillips, O.L., Platts, P.J., Balmford, A., Burgess, N.D., Lovett, J.C., Ahrends, A., Bayliss, J., Doggart, N., Doody, K., Fanning, E., Green, J.M.H., Hall, J., Howell, K.L., Marchant, R., Marshall, A.R., Mbilinyi, B., Munishi, P.K.T., Owen, N., Swetnam, R.D., Topp-Jorgensen, E.J., Lewis, S.L., 2014. Quantifying and understanding carbon storage and sequestration within the Eastern Arc Mountains of Tanzania, a tropical biodiversity hotspot. Carbon Balance Manag 9.

Willcock, S., Hooftman, D., Sitas, N., O'Farrell, P., Hudson, M.D., Reyers, B., Eigenbrod, F., Bullock, J.M., 2016. Do ecosystem service maps and models meet stakeholders' needs? A preliminary survey across sub-Saharan Africa. Ecosyst. Serv. 18, 110–117. https://doi.org/10.1016/j.ecoser.2016.02.038.

Willcock, S., Martínez-López, J., Hooftman, D.A.P., Bagstad, K.J., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G., Voigt, B., Villa, F., Bullock, J.M., Athanasiadis, I.N., 2018. Machine learning for ecosystem services. Ecosyst. Serv. https://doi.org/10.1016/j.ecoser.2018.04.004.

Willcock, S., Hooftman, D.A.P., Balbi, S., Blanchard, R., Dawson, T.P., O'Farrell, P.J., Hickler, T., Hudson, M.D., Lindeskog, M., Martinez-Lopez, J., Mulligan, M., Reyers, B., Shackleton, C., Sitas, N., Villa, F., Watts, S.M., Eigenbrod, F., Bullock, J.M., 2019. A continental-scale validation of ecosystem service models. Ecosystems 22, 1902–1917. https://doi.org/10.1007/s10021-019-00380-y.

Wong, C.P., Jiang, B., Kinzig, A.P., Lee, K.N., Ouyang, Z., 2014. Linking ecosystem characteristics to final ecosystem services for public policy. Ecol. Lett. 18, 108–118. https://doi.org/10.1111/ele.12389.

Xu, Z., Chau, S.N., Chen, X., Zhang, J., Li, Yingjie, Dietz, T., Wang, J., Winkler, J.A., Fan, F., Huang, B., Li, S., Wu, S., Herzberger, A., Tang, Y., Hong, D., Li, Yunkai, Liu, J., 2020. Assessing progress towards sustainable development over space and time. Nature 577, 74–78. https://doi.org/10.1038/s41586-019-1846-3.