Title: Model Ensembles of Ecosystem Services Fill Global Certainty and Capacity Gaps

Authors: Simon Willcock[1,2±*], Danny A. P. Hooftman[3,4±], Rachel A. Neugarten[5,6,7], Rebecca Chaplin-Kramer[8,9,10], José I. Barredo[11], Thomas Hickler[12,13], Georg Kindermann[14], Amy R. Lewis[2], Mats Lindeskog[15], Javier Martínez-López[16,17] & James M. Bullock[4]

1. Net Zero and Resilient Farming, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, United Kingdom; simon.willcock@rothamsted.ac.uk
2. School of Natural Sciences, Bangor University, Bangor, Gwenydd, LL57 2DG, United Kingdom; amy.lewis@bangor.ac.uk
3. Lactuca: Environmental Data Analyses and Modelling, the Netherlands; danny.hooftman@lactuca.nl
4. UK Centre for Ecology & Hydrology, Wallingford OX10 8BB, United Kingdom; jmbul@ceh.ac.uk
5. Department of Natural Resources & Environment, Cornell University, 226 Mann Drive, Ithaca NY 14853 USA; ran63@cornell.edu
6. Conservation International, 2100 Crystal Drive #600, Arlington, VA 22202, USA
7. Cornell Lab of Ornithology, Cornell University, 159 Sapsucker Woods Rd, Ithaca, NY 14850, USA
8. Global Science, Word Wildlife Fund, 131 Steuart Street, San Francisco, CA 94105, USA becky.chaplin-kramer@wwf.org
9. Institute on the Environment, University of Minnesota, 1954 Buford Ave, St. Paul, MN, USA
10. Natural Capital Project, Stanford University, 327 Campus Drive, Stanford CA, USA
11. European Commission, Joint Research Centre, Ispra, Italy; jose.barredo@ec.europa.eu
12. Senckenberg Biodiversity and Climate Research Centre, Frankfurt, Germany; thomas.hickler@senckenberg.de
13. Institute of Physical Geography, Goethe-University, Altenhöferallee 1, 60438 Frankfurt am Main, Germany.
14. International Institute for Applied Systems Analysis, Laxenburg, Austria; kinder@iiasa.ac.at
15. Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden; mats.lindeskog@nateko.lu.se
16. Department of Ecology, University of Granada, Avda. del Mediterráneo s/n, E-18006 Granada, Spain; javier.martinez@ugr.es
17. Instituto Interuniversitario de Investigación del Sistema Tierra en Andalucía (IISTA), Universidad de Granada, Avda. del Mediterráneo s/n, E-18006 Granada, Spain

[±] Joint 1st author
* Corresponding author

Teaser

Global ensembles of ecosystem service models have increased accuracy and fill data gaps for less wealthy regions

Abstract (max 150 words)

Sustaining ecosystem services (ES) critical to human wellbeing is hindered by many practitioners lacking access to ES models ('the capacity gap') or knowledge of the accuracy of available models ('the certainty gap'), especially in the world's poorer regions. We developed ensembles of multiple models at an unprecedented global scale for five ES of high policy relevance. Ensembles were 2-14% more accurate than individual models. Ensemble accuracy was not correlated with proxies for research capacity – indicating accuracy is distributed equitably across the globe and that countries less able to research ES suffer no accuracy penalty. By making these ES ensembles and associated accuracy estimates freely available, we provide globally consistent ES information that can support policy and decision making in regions with low data availability or low capacity for implementing complex ES models. Thus, we hope to reduce the capacity and certainty gaps impeding local to global-scale movement towards ES sustainability.

Keywords: Accuracy; Ensemble; Implementation gap; Modelling; Nature's contributions to people; Uncertainty.

Introduction

There is a burgeoning number of ecosystem service (ES) maps delineating an ever-growing understanding of the ways in which nature benefits people (e.g. *1*, *2*). However, when ES data are available, they are typically inconsistent between countries, making standardized measurement or reporting difficult (*3*). Global maps (based on satellite and other data integrated in a variety of models) can provide readily-available information when more locally relevant data are lacking (*4*). Though, it is questioned whether global maps provide accurate or useful information given their lack of sensitivity to local context (*5*). It is difficult to answer this question for most large-scale ES modelling exercises due to the lack of information on model accuracy - the closeness of the agreement between the modelled value and a reference value (*6*), the latter being considered 'true' (*7*) even though the validation data are also often uncertain (*8*). Individual model performance varies, validation with empirical data is sometimes lacking, and results are typically reported without estimates of accuracy (*8*). Two key advantages of global maps are that they can fill gaps in data-poor contexts until local data can be collected or created, and they are consistent among countries (*4*). For example, at a local level, the Critical Ecosystem Partnership Fund made conservation investment decisions in Madagascar based, in part, on local information on the relative importance of sites for ES derived from models and globally available data (*9*). At a global scale, consistent data can be used for international policy and decision making [e.g. informing targets and investments in the united Nations (UN) Sustainable Development Goals, the Convention on Biological Diversity post-2020 Biodiversity Framework, the UN's System of Environmental-Economic Accounting-Ecosystem Accounting (*10*)]. Global data can also provide consistent and comparable local reporting for these international agreements, as well as broader context for local decisions by revealing wider regional, continental and global patterns in ES status and trends (*4*).

Several studies have validated models of single ES (e.g. *11*, *12*), and rarely multiple ES (e.g. *8*, *13*). Independent evaluations of models have often been unable to demonstrate the consistently superior accuracy of any individual model (*8*, *13*). While a few studies find that, on average, more complex ES models show better fit to validation data, the best-fit model varies regionally and often according to the validation data used (*8*, *13*). Thus, decisions based on a single model for an ES are less likely to be robust and, when models are in disagreement, it is difficult for practitioners (those engaging with information from ES models) to know which model should be used to support decisions (*14*). In fact, projections by alternative models can be so variable as to compromise even the simplest assessment and therefore challenge the common practice of relying on a single method (*15*). This 'certainty gap' greatly reduces the confidence that practitioners have in projections from ES models (*16*).

The certainty gap is unlikely to be uniformly distributed across the globe. In developing countries, reliable information about ES is critically important because the rural and urban poor are often the most dependent on ES (directly or indirectly), both for their livelihoods and as a coping strategy for buffering shocks (*17*). ES declines driven by over-exploitation, habitat conversion or climate change therefore undermine 80% (35 of 44) of the Sustainable Development Goals (SDGs) (*18*). However, ES data and accuracy estimates are often unavailable in developing nations, or in less affluent regions within nations, where they are most needed (*17*). There is an urgent need for evaluations of model accuracy to better inform decision making – a need that has been emphasised by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) (*19*). To address this, researchers have established standards for best practice using model-data (*8*) and model-model (*13*, *20*) comparisons to provide robust and transparent evaluations of accuracy. For example, an ensemble of models is more accurate, on average, than one model for any location, although the amount of improvement depends on the local context and the models used (*13*, *15*, *20*). However, whilst model ensembles are common in climate modelling and other disciplines (*15*, *21*), they have been largely neglected in ES studies (*22*). Indeed, simple ('committee average') ensembles have been found to be at least 5% more accurate than individual ES models (*13*), while more complex, weighted ensembles provide even better predictions (up to 27% more accurate) (*20*). Furthermore, variation among models can provide an indicator of the uncertainty of the modelled ES estimate when no other information is available (*13*).

Whilst using ensembles of ES models is possible, there are barriers that need to be overcome before it can become standard practice within ES science. Implementing multiple ES models remains a difficult undertaking for many researchers and practitioners (*13*). Barriers include lack of input data, resources, and capacity for data collection or collation and for modelling (*13*, *14*). As with the certainty gap, these barriers are typically more substantial in poorer nations. For example, to create ensembles of carbon storage models across three major platforms – ARIES (*23*), InVEST (*24*) and Co$ting Nature (*25*) – requires access to the internet, high quality input data, computational power and GIS proficiency, as well as funds to support model subscription fees (where required) and the person-time required to learn and run three different models (*13*). Such resources can be out of reach for many researchers and practitioners. Furthermore, if practitioners must choose between running multiple models for a single service versus modelling additional services, the former may be of low priority; thus the widespread use of ES ensembles may be an unrealistic goal (*13*, *14*, *20*). We refer to the lack of these resources as the 'capacity gap'. One potential solution to the capacity gap is that those who have the resources to create ES ensembles make the resulting data, as well as estimates of uncertainty, freely available (e.g. *13*, *20*).

To address the certainty and capacity gaps, we developed ensembles of models for five ES (Figure 1) of high global and local policy relevance (*14*), and for which there are both: i) a variety of models available that are feasible to run at a global scale; and ii) accessible, independent validation data to assess ensemble accuracy. We included three material services (water supply: eight available models; fuelwood production: nine models; and forage production: 12 models); one regulating service (above ground [AG] carbon storage: 14 models); and one non-material service (recreation: five models). Some of these ES are potential services (e.g. water, fuelwood, forage) and some are realised (e.g. carbon recreation); where potential ES are 'the outcomes from ecosystems that directly lead to good(s) that can be used and valued by people (e.g. harvestable products, water supply), noting that some ecosystem services can be both ecosystem processes and potential ecosystem services', and realised ES are 'all use and non-use, material and non-material outputs from ecosystems that are used and valued by people' (*26*, *27*). Both potential and realised service metrics are useful to support decision making; with the latter providing insight into how the wellbeing of people is improved by nature, and the former indicating the maximum capacity of these potential wellbeing increases (*14*). We used model output predictions and created ES ensembles at an unprecedented global extent and at a 0.008333° resolution (approximately 1 km at the equator). We address the capacity gap by making the ensemble model outputs freely available (https://doi.org/10.5285/bd940dad-9bf4-40d9-891b-161f3dfe8e86), as well as providing the code (github.com/GlobalEnsembles) to make the overall approach more accessible. To address the certainty gap, we tested the accuracy of these ensembles against independent validation data (including country-level statistics and actual biophysical measurement), and investigated spatial patterns in ensemble accuracy.

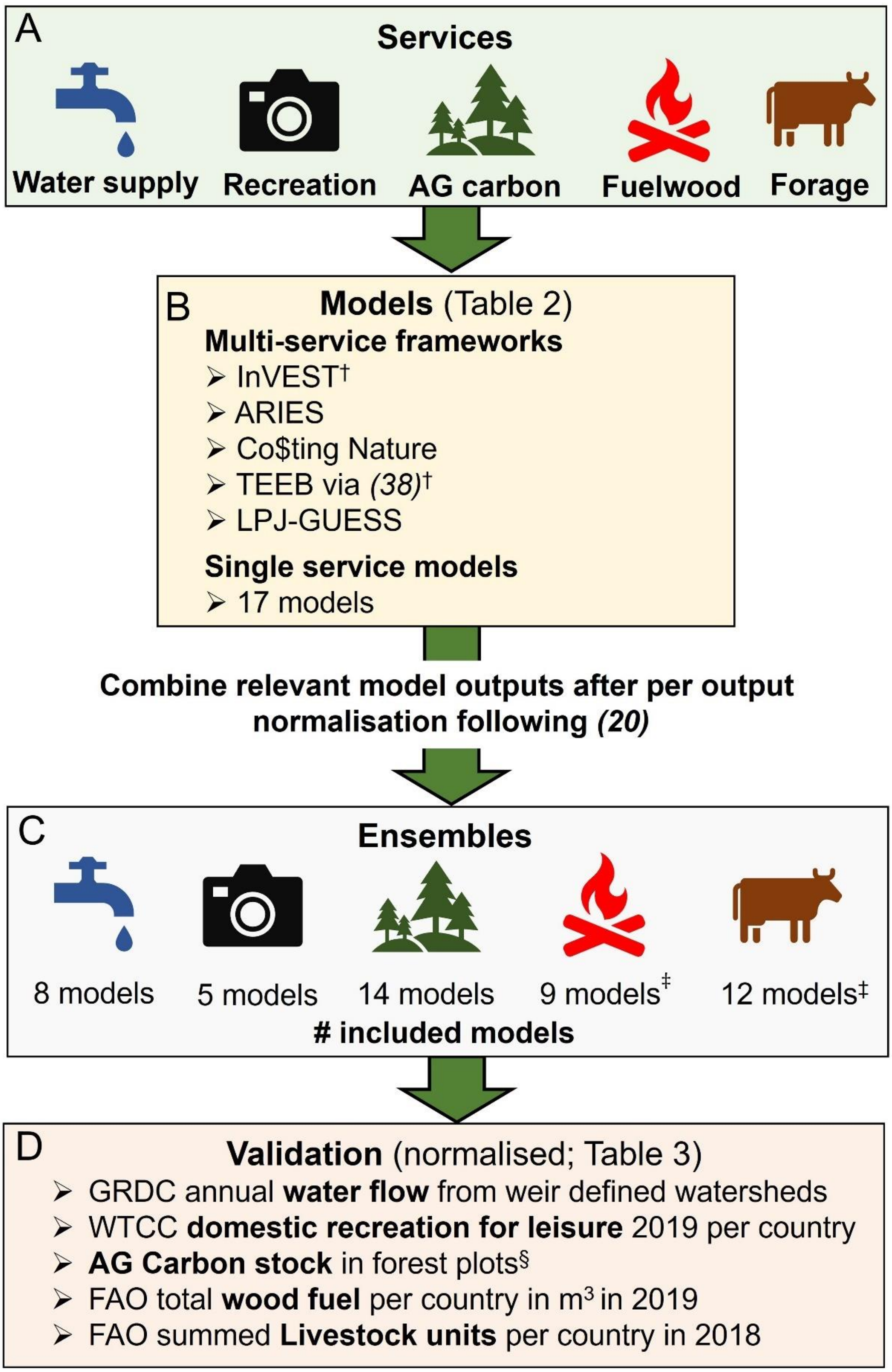

Figure 1: Schematic overview of the model flow, ensemble creation and validation processes implemented in this study. We modelled five ecosystem services (ES): potential water supply as flow in rivers, recreation as the number of visitors, potential above ground (AG) carbon stock, potential fuelwood, and potential forage production capacity (A). We used models for these ES from five multi-service frameworks (i.e. multiple ES per modelling framework) and 17 individual ES models (Table 2; B). These models are combined into model ensembles following Hooftman et al. (*20*), with the number of models in the ensemble for each ES shown in C. We use validation data on each ES to test accuracy of the ensembles to both their own service and as proxy for other services (D). Symbol key: †Including choice of input data; ‡ Including models created by masking of above ground (AG) carbon models with woody (fuelwood) or grassland land use masks [see (*8*); SI-3]; § combined pan-tropical biomass reference data (*28*) and United Kingdom (temporal) AG biomass stocks in forest estates (*20*). See Methods for full details.

Results

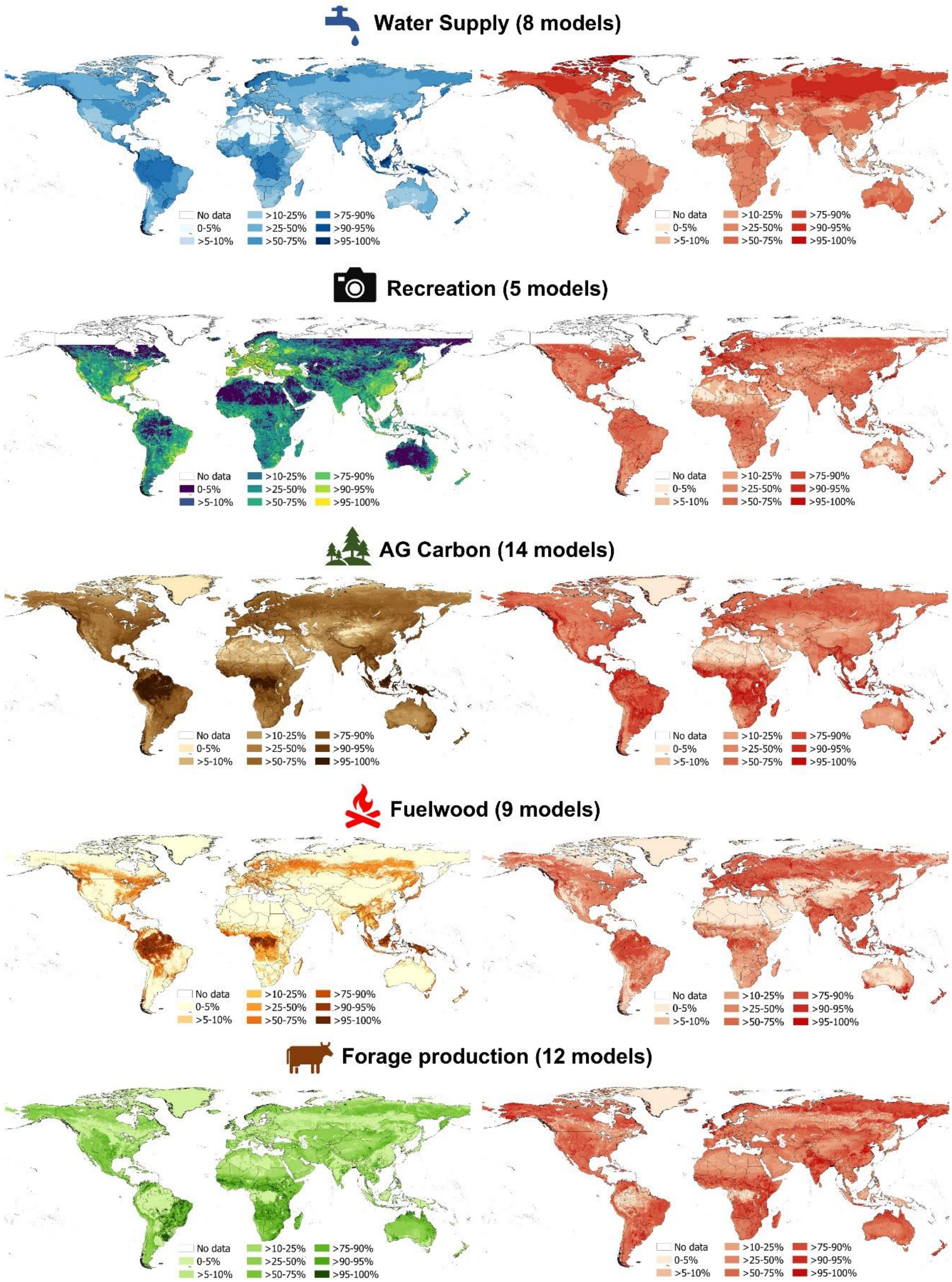

Figure 2: Left) Median ensembles values from models for five ecosystem services (ES) of high policy relevance (*14*)**.** We created ensembles for water, recreation, AG carbon, fuelwood, and forage production at global scale and at an 0.008333° resolution by taking the median value of multiple models for each grid cell. Addressing the capacity gap, we make these freely available via https://doi.org/10.5285/bd940dad-9bf4-40d9-891b-161f3dfe8e86, as well as maps produced using alternative ensemble approaches (including mean, PCA, correlation coefficient, and regression to the median and leave-one-out cross validation log-likelihood approaches – see SI). **Right) Addressing the certainty gap, we show the standard error of the mean associated with each ES ensemble output** which, in accordance with previous research (*13*), our investigations show can be used a proxy for ensemble accuracy in absence of validation data (Figure S4). All maps scaled in deciles 0-100%. True zero values (coloured) are distinguished from no-data (white). Selected case study regions are shown in SI-6. The figures are available via https://github.com/GlobalEnsembles/Maps, and the data are available via https://doi.org/10.5285/bd940dad-9bf4-40d9-891b-161f3dfe8e86 .

Here, we present results using an unweighted median ensemble (*8*) approach (i.e. taking the median value of multiple models for each grid cell; Figure 2). Other ensemble approaches, including unweighted (mean), and weighted (deterministic consensus: PCA & correlation coefficient; iterated consensus: regression to the median and leave-one-out cross validation log-likelihood) approaches) (*20*), which give consistent conclusions, are described in the SI. When compared to independent validation data (Figure 1, Table 3), global ES ensembles were more accurate than an individual model chosen at random (Table 1, Figure 3). Median ensemble improvement per validation data point for each ES was 14% for water (resolution of the validation data: weir defined watersheds), 6% for recreation (national-scale), 6% for above ground (AG) carbon (plot-scale), 3% for fuelwood (national-scale), and 3% for forage production (national-scale; Table 1, Figure 3). Thus, using global ES ensembles rather than an individual ES model reduces the certainty gap for practitioners with no *a priori* information on model accuracy. In general, the weighted ensembles provided more accurate predictions than unweighted ensembles (Figure S15 and S16), and so should be favoured by practitioners. Ensembles further address the certainty gap by transparently conveying any spatial variation in accuracy. For example, the standard error of the mean associated with each ES ensemble (Figure 2) correlates with the accuracy of the ensemble and so can be used as a proxy for ensemble accuracy in absence of validation data [(*13*) and Figure S4], indicating the accuracy of the ensembles in any specific geographic location. Our results are consistent when using alternative accuracy metrics (e.g. Spearman's ρ; see SI-5).

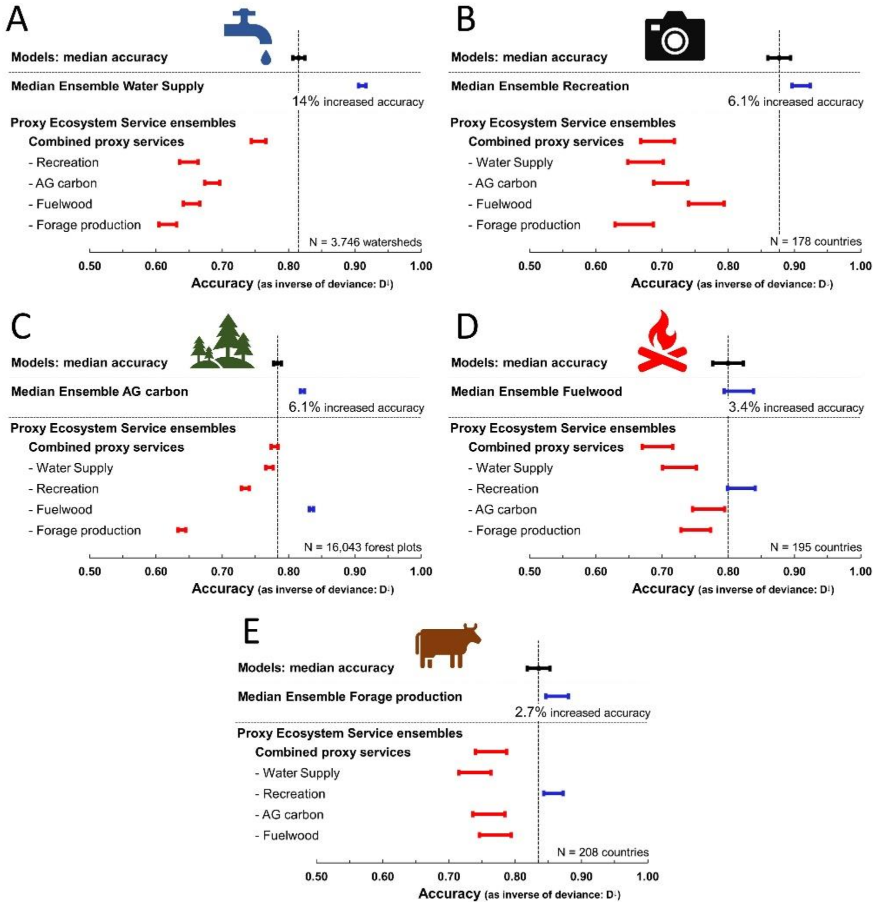

Figure 3: Ecosystem service (ES) ensembles show increased accuracy when compared to individual models. Shown are the median ensembles for: A) water, B) recreation, C) AG carbon, D) fuelwood, E) forage production. ES theory on bundles suggest that values for different ES can be spatially related to each other, either positively or negatively (*2*). However, spatial correlations among ES, while they do occur, may vary geographically, meaning there is no consistent correlative relationship among ES over large spatial scales (*29*). To test this, we spatially correlated each global ES ensemble output with the output of all other ES ensembles, both as a group (or 'bundle'; i.e. for all ES ensembles combined) and for each ES individually. Our results showed ES 'bundles' to be a relatively poor predictor of an additional ES and that most ES ensembles were not well correlated with other ES on an individual ES basis. Vertical dashed lines indicate the among-run median accuracy of an individual model chosen with no *a priori* information (i.e. at random). Blue bars indicate a model (or ensemble) accuracy was significantly higher than the median accuracy of the models (length of bars represent among-model standard deviation). Red bars indicate accuracy was significantly lower than the median of the models.

Whilst the results presented here show ES ensembles reduce the certainty gap, differences in ensemble performance between regions or countries might be expected. For example, nations investing more in research capacity might have better input data or have more researchers who develop and test ES models, potentially resulting in model outputs that are more locally relevant in those areas (*30*). Thus we might expect that ES ensembles perform better in countries with higher GDP, Human Development Index scores, or research capacity. After accounting for spatial autocorrelation (see Methods) and applying the Hoghberg correction to account for multiple tests, we found no evidence that ensembles are more accurate in countries with higher GDP (even when accounting for within-country variability using Gini metrics of inequality), with higher Human Development Index, or with higher research capability (expressed as the percentage of people who are researchers and proportion of GDP invested in research; Table 1). The results are consistent when using alternative statistical approaches (Tables S7-9). These findings suggest global consistency in ensemble accuracy, in relation to the potential drivers of variation that we tested (Table 1). A potential caveat is that if the validation data themselves are biased (for example, less accurate across developing countries) then true patterns in ensemble model accuracy could exist undetected.

Table 1: One tailed correlations as F-values with significance of the inverse of deviance per validation datapoint (where increasing accuracy is represented by increasing the inverse of deviance) of the five ecosystem service (ES) ensembles against globally available metrics that could potentially impact model accuracy. One-tailed tests were applied to test the hypothesis that the ensemble accuracy increases with higher values of each development/equality measure (two-tailed is presented in Table S7, including effect sizes). Degrees of freedom were standardised at 178 following a bootstrap convergence model for all services. Significance of the presented F-values were assessed taking account of multiple tests, using Hochberg's step-up correction with 8 tests per ES. An interaction model is added testing for interactions between GDP per capita and income equality, reflecting that income may be better represented using both mean and variance. To conform to the normality assumptions of the analysis, all metrics were arcsine transformed, with the exception of GDP per capita, which was $\log_{10}$-transformed, and the Human Development Index, which was not transformed. See Table 3 for the sources of each validation dataset.

	Water Supply	Recreation	AG Carbon	Fuelwood Production	Forage Production
Accuracy Improvement (inverse of deviance) Ensemble vs. a random selected model (median among models)†	14%	6.1%	6.1%	3.4%	2.7%
Spatial Autocorrelation‡	15.3***	14.6***	211***	0.47	0.14
Development/Equality per country					
GDP per capita	1.38	<0.01	1.21	3.58	0.24
Human Development Index	1.51	<0.01	0.14	6.43	0.25
Income Equality (Gini index)	0.17	6.69	1.37	<0.01	0.71
% People in R & D	1.44	<0.01	0.15	4.85	0.08
% GDP to R & D	0.08	<0.01	0.14	3.79	0.37
Interaction model					
GDP per capita	1.76	0.18	0.16	1.29	0.02
Income Equality	1.67	0.22	0.16	0.50	0.02
GDP x Income Equality	0.06	0.34	1.04	0.16	2.67

†Mean of pairwise comparisons per 1000 bootstrap runs; ‡Two sided tested without direction; *** P <0.001 corrected.

Finally, whilst the five ES ensembles made available here contribute to addressing the capacity gap, practitioners will often require accurate ES information on many additional services, including many for which there are no models (*14*). ES theory on bundles suggest that values for different ES can be spatially related to each other, either positively or negatively (*2*). However, spatial correlations among

ES, while they do occur, may vary geographically, meaning there is no consistent correlative relationship among ES over large spatial scales (*29*). To test this, we spatially correlated each global ES ensemble output with the output of all other ES ensembles, both as a group (or 'bundle'; i.e. for all ES ensembles combined) and for each ES individually. Our results showed ES 'bundles' to be a relatively poor predictor of an additional ES (Figure 3). Similarly, most ES ensembles were not well correlated with other ES on an individual ES basis.

Discussion

To help fill a major capacity gap in terms of available ES information for many countries, we have provided globally consistent ensemble data on five ES (https://doi.org/10.5285/bd940dad-9bf4-40d9-891b-161f3dfe8e86), as well as the code required to produce them (github.com/GlobalEnsembles). Finding increased performance through use of ensemble approaches is common in other fields (*20*), although an increase is not universal (*31*). Due to underlying assumptions, model predictions (including those from ES models) are all potentially biased in direction and amount, with biases varying among models due to their specific construction and available input data (*20*). The improvement in accuracy when using ensembles likely derives from suppression of idiosyncratic differences by inclusion of multiple possible system representations (termed a 'portfolio effect'), providing a more reliable average estimate (*20*, *32*). However, this effect is lessened if assumptions, and therefore concomitant biases, are shared across models (*20*). This highlights the importance of including: i) multiple model outputs in model ensembles (*33*), including from models not explicitly identified as ES models, such as hydrological models (*20*); and, ii) where data are available, model validation (*8*) - see Dormann *et al.* (*32*) and Hooftman *et al.* (*20*) for further theoretical explorations. Using ensembles also improves consistency across independent studies. For example, considering two studies applying different models in different locations, it is uncertain how comparable the findings are (*4*). However, if both studies use model ensembles, even if the ensemble approaches are not identical, results will be more comparable. This is because variation among ensemble approaches is substantially lower than among individual models (*20*) - resulting in greater consistency and coherence. Thus, potential applications of ES ensembles include supporting nations' efforts to implement natural capital accounting (*3*).

Our finding that global ES ensembles perform just as well in less wealthy regions with lower research capacity, where this information is often most needed, emphasises the utility of these modelled data. This might reflect that ES models are increasingly tested and parameterized using global-scale Earth Observation data. In addition to the ensemble maps themselves, we provide estimates of accuracy (https://doi.org/10.5285/bd940dad-9bf4-40d9-891b-161f3dfe8e86). The ability to quantify accuracy when it comes to ES is often lacking and, at worst, this can result in perverse outcomes – with the 'pot luck' associated with model selection (i.e. without *a priori* accuracy information) sometimes resulting in implementation of low-accuracy outputs and suboptimal decisions (*8*, *19*). For policy and decision making, accuracy estimates are as important as the ES maps themselves, and the lack of information about uncertainty is one driver of the 'implementation gap' between ES research and its incorporation into policy and decision making (*16*). By providing accuracy maps we are directly addressing this certainty gap. However, future work should seek to improve on these accuracy maps, particularly through the collection and inclusion of additional validation data at local scales, as using the national- and watershed-scale validation data that is currently available may be a poor proxy of model accuracy at local-scales.

Important capacity gaps remain. Most ES research predominantly focusses on a limited set of material and regulating services because the data are widely available, and their underlying processes are relatively well understood (*34*). This means our current ability to assess or predict unmodelled ES is low. We found that ensembles, whether as an individual ES or as a bundle, do not accurately predict other ES at global scales. It could be that as more ES are included in a bundle, predictive power of the bundle for unmodelled ES improves; in a recent analysis global maps resulting from individual models for 12 ES show high correlations between any one service and the remaining 11 (*2*). This is possibly because the more and more diverse ES that are included, the more likely that unmodelled ES will also be represented by the same set of ecosystems, either because they are similar to modelled ES or simply by chance. In general, the utility of the bundle approach is debated, with Spake et al (*29*) suggesting that a hypothesis-driven approach is required to predict relationships between ES. Ultimately, whilst individual models are available for more ES than are presented here, model development is urgently required before ensembles of additional ES can be assessed.

Practitioners show both capacity and willingness to engage with accuracy information when it is made available (*14*). Accuracy estimates allow practitioners to determine what level of confidence is acceptable to them and to use their own expertise to make potentially contentious decisions (*35*). Given limited resources, accuracy information can play an important role in prioritisation. For example, the accuracy of estimates may be vital in distinguishing between two sites with high levels of ES production. Another example could be a decision to give a site with high accuracy of medium ES levels lower priority over a potentially high-value site with medium or low accuracy; this is contentious, but defensible if accuracy information is transparently conveyed to practitioners. Thus, providing estimates of accuracy should become standard practice within the ES community (*22*). High levels of inaccuracy or uncertainty of ES estimates should not lead to inaction, but instead highlights the risks of making decisions using poor data, what data may need to be gathered to improve model inputs, or the need to develop new or improve existing ES models. The model-estimated quantity of ES and its accuracy should not be the only metrics considered in decision-making. For example, as the wellbeing of some marginalised groups may depend on ES where models or data are lacking, or uncertainty is high, therefore it is critical to incorporate local knowledge and values in any decision making process (*2*). Indeed, model accuracy is one of a range of metrics considered by practitioners when determining whether model outputs can be used to support decision-making, with others including spatial resolution and the ability to incorporate scenarios (*14*). Thus, simply reducing uncertainty is not necessarily going to lead to better policy decisions. However, in regions with a large capacity gap, practitioners lack any comprehensive spatial data on most ecosystem services. For these regions, our 1 km^2 resolution ES ensemble outputs provide, at a minimum, some data with a level of validation and associated accuracy at little to no cost to the practitioner (*14*).

We conclude that ensemble modelling of ES can help reduce capacity and certainty gaps by, for example, making more accurate ES estimates freely available. We suggest ES scientists adopt ensemble approaches (shown here to be, on average, a more accurate approach than using individual models), and accompany model outputs with estimates of uncertainty. These changes may help reduce the implementation gap between ES research and policy and decision making (*14*, *34*), in particular for assessments by IPBES and the Intergovernmental Panel on Climate Change.

Materials and Methods

We developed and tested (against validation data) ensembles of models for five ecosystem services (ES; Figures 1 & 4) for which there are both a variety of models which are feasible to run at a global-scale (*8*, *20*) and accessible independent validation data. We used model output estimates of ES (listed in Table 2) to create ensembles, and then validated them against independent data (Table 3) using methods developed previously for the UK (*20*) and sub-Saharan Africa (*8*). To ensure comparability among model outputs, we standardised them by normalising outputs from individual models prior to creating ensembles, following the same procedure for the validation data. We explored the spatial variation in accuracy of ES ensembles, using a variety of metrics. Finally, we investigated the use of ES ensemble 'bundles' as proxies for other ES. We depict our overall process in Figure 4 in 6-steps. Our calculations were performed using Matlab v7.14.0.739, ArcMap 10.7 and ArcPro 2.7, employing Arcpy coding for loops. Relevant code can be found at github.com/GlobalEnsembles.

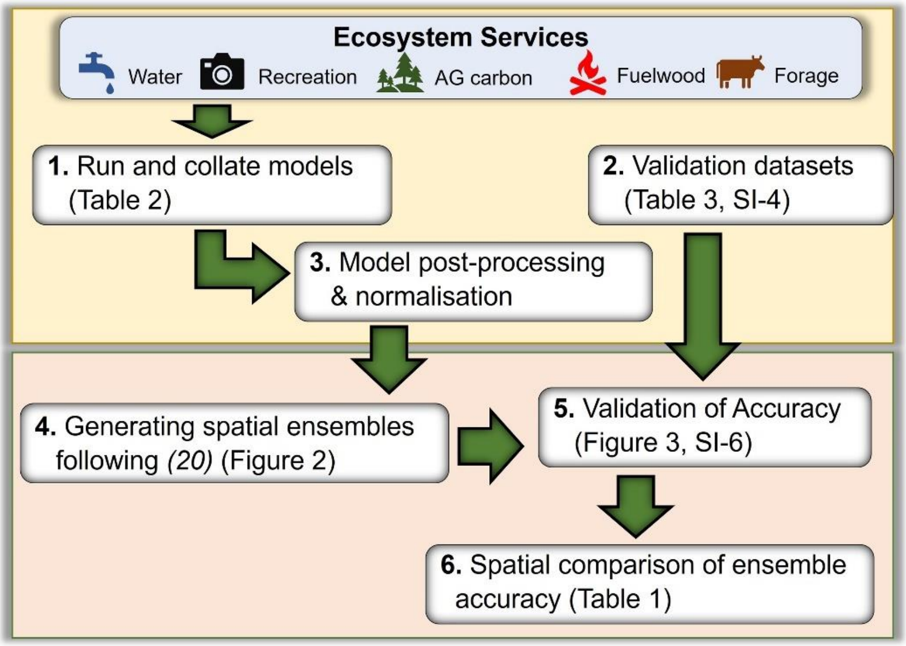

Figure 4: Schematic representation of our analysis, with arrows showing information flows. Numbers represent the steps within our methods; input tables and result figures are indicated.

1. Run and collate models

We collated models for this study according to their availability and feasibility to be run at a global scale, and to reflect different approaches to modelling ES, obtaining appropriate registrations and licenses if necessary. The collated models are summarised in Table 2, including their output grid sizes (spatial resolution) – as well as whether the model outputs are existing (*i.e.* can be found online; e.g. 10, 31), are generated online (ARIES, Co$ting Nature, WaterWorld), or can be calculated with a desktop tool (InVEST) or in local ArcGIS environment (Scholes, TEEB). For models that require input data choices (InVEST, Scholes, TEEB), we refer to SI-1 for details and supporting data. For models that were taken from Willcock *et al.* (*8*) and Hooftman *et al.* (*20*), we refer to the descriptions in those papers.

Table 2: Summary information for the individual ecosystem service models used in this study.

Model	Ecosystem Service	Details	Model Output Resolution
Multi service frameworks			
ARIES k.explorer (*23*) for year = 2020 integratedmodelling.org/modeler	- Recreation[†] - AG carbon - Forage production[‡] - Fuelwood production[¶]	Recreation run online per country; carbon follows (*36*); all in tonnes per hectare, except recreation in normalised # of people (SI-1-4)	0.008333° Mostly worldwide
Co$ting Nature (*25*) policysupport.org/costingnature	- Water Supply - AG carbon - Recreation[†§] - Forage production - Fuelwood production	Run online as 10° tiles; subsequent among tile normalisation; all unitless normalised indexes, except water in m^3 per year.	0.008333° Not above 60° North
InVEST v3.8.7 (*24*) naturalcapitalproject.stanford.edu/-software/invest	- Water Supply - AG carbon - Recreation[§] - Forage production[‡] - Fuelwood production[¶]	Desktop tool, parameterised for this project (SI-1-1). Water supply in m^3 per gridcell; recreation in number of photo uploads; carbon/forage/fuelwood in tonnes per hectare.	0.008333° Worldwide
Lund-Potsdam-Jena General Ecosystem Simulator (LPJ-GUESS) (*37*)	- Water Supply	Data set from (*8*) and run as described therein. Water supply in	0.5° Worldwide

	- AG carbon - Forage production© - Fuelwood production¶	m^3 per gridcell; carbon/forage/ fuelwood in tonnes per gridcell.	
TEEB via Costanza *et al*. (*38*)	- Water Supply - AG carbon - Recreation - Forage production‡ - Fuelwood production	In local GIS environment (SI-1-2) , all in *US-$* for the year 2007 as provided by (*38*)	0.002778° Worldwide
Scholes(*39*) via Willcock *et al*. (*8*), livestock distributions extended worldwide	- Water Supply - Forage production	In local GIS environment extended from (*8*); SI-1-3. Water supply in positive growth days; forage in livestock units per hectare.	0.008333° Worldwide
Single service models			
Aqueduct Global Maps 2.1 (WRI) (*40*): accumulated water run-off wri.org/data/aqueduct-global-maps-21-data	- Water Supply	Existing data; as available blue water (m^3) per catchment outlet	Watershed Polygons Worldwide
European map of above ground biomass stocks (*41*)	- AG carbon	Existing data, from (*20*); as tonnes per hectare	0.008333° Europe only
ESA CCI Biomass Climate Change Initiative (*42*) data.ceda.ac.uk/neodc/esacci/-biomass/data/agb/maps/v2.0/geotiff/2018	- AG carbon	Existing data; as tonnes per hectare	0.0008888° Worldwide Forest only
FAO combined gridded livestock distributions fao.org/livestock-systems/globaldistributions	- Forage production	Existing data, summed LSUs among types (SI-2) per gridcell	0.08333° Worldwide
Integrated GEOCARBON global forest biomass (*28*) lucid.wur.nl/datasets/high-carbon-ecosystems	- AG carbon	Existing data; as tonnes per hectare	0.01° Worldwide. Forest only
Gilbert *et al*. (*43*); Combined gridded livestock distributions dataverse.harvard.edu/dataverse/glw_3	- Forage production	Existing data, summed LSUs among types (SI-2) per gridcell	0.08333° Worldwide
Global Forest Watch, above ground biomass (*44*) data.globalforestwatch.org/-datasets/above-ground-live-woody-biomass-density/data	- AG carbon	Existing data; as tonnes per hectare	0.00025° Worldwide Forest only
JRC Above ground Biomass (*45*) data.jrc.ec.europa.eu/dataset/biomass	AG carbon	Existing data; as tonnes per hectare	0.0008333° Europe only
Chaplin-Kramer *et al*. (*2*)	- Recreation	Existing data in number of people per gridcell	0.01667° Not above 60° North
WaterWorld (*46*): Accumulated water run-off policysupport.org/waterworld	- Water Supply	Run online per available catchment in m^3 per catchment outlet	0.008333° Partially Worldwide
WaterWorld (*46*): Water Budget per cell policysupport.org/waterworld	- Water Supply	Run online per available catchment in m^3 per gridcell	0.008333° Partially Worldwide
Single service Carbon models with masked use for Grazing and Fuelwood			
Avitabile *et al*. (*28*): carbon in vegetation lucid.wur.nl/datasets/high-carbon-ecosystems	- AG carbon - Forage production‡ - Fuelwood production¶	Existing data; as tonnes per hectare	0.008333° Tropics only
Conservation International Total Carbon in vegetation (*47*) conservation.org/projects/-irrecoverable-carbon	- AG carbon - Forage production‡ - Fuelwood production¶	Existing data; as tonnes per hectare	0.002695° Worldwide
Kindermann *et al*. (*48*) above ground biomass stocks	- AG carbon - Forage production‡ - Fuelwood production¶	Existing data, from (*20*) ; as tonnes per hectare	0.008333° Worldwide
ORNL DAAC (NASA), above ground biomass density (*49*) daac.ornl.gov/cgi-bin/	- AG carbon - Forage production‡ - Fuelwood production¶	Existing data; as tonnes per hectare	0.002778° Worldwide

†including post-processing with a further data set (SI-1-4); based on above ground (AG) carbon with an ‡grassland and ¶woodland MODIS land cover mask following[9] (SI-3); §realised service based on photo uploads; © combined C_3 and C_4 carbon.

All model outputs were projected to WGS 1984 (EPSG 4326) and rescaled to a 0.008333° grid (approximately 1km at the equator), resampling models where necessary (Table 2). Generally, when upscaling, cells were aggregated by calculating the mean of the grid cell values with no-data cells ignored; when downscaling ArcPro's bilinear recalculation algorithm was used for resampling. This latter resampling resulted in a smooth transition by assuming values of smaller cells via linear extrapolations from neighbouring cells (e.g. for LPJ, gridded livestock). Small-scale non-linearity (e.g. as a result of unmodelled features) is not included in this downscaling, as such an output would heavily depend on post-processing assumptions and inputs and be a model in its own right. Rescaling factors are not needed during these calculations since these will not change relative values (*i.e.* resulting from subsequent normalisation; Step 3). All outputs were clipped and aligned to the exact same extent with standard number of rows and columns (43,200 x 18,600), using ArcPro's bilinear recalculation algorithm.

Whilst all model outputs were obtained at global scale, not all cover the entire globe (Table 2). Only the terrestrial globe was considered, but there were other specific constraints. For example, servers for certain online models restricted overly large data flows. Specifically, ARIES k.explorer was not able to run the recreation module per country for North America and parts of Europe because of the high level of detail in the supporting maps (*23*); WaterWorld was not able to run the largest watersheds such as the Amazon basin, the Mississippi and the Yangtze (*46*). Furthermore, Co$ting Nature is limited to latitudes below 60° north due to lack of input data for northern regions (*25*). We used above ground (AG) carbon models that were region specific; two for Europe (*41*, *45*), one for the tropics (*28*) and three that were forest vegetation specific (*28*, *42*, *44*).

2. Validation data sets

Our validation data sets are listed in Table 3 (and mapped in SI-4). Broadly, they include either informed expert statistics (such as country-level statistics from the FAO [forage production and fuelwood] and recreation values from the World Travel and Tourism Council), or actual biophysical measurement (tree inventory plots for AG carbon, and weir data for water flow):

- Our water supply validation data set is catchment-based. Specifically, we used a Global Runoff Data Centre (GRDC) data set with 3746 weirs (Table 3; Figure S1), covering all regions but not all land area. For each weir, bespoke catchments polygons were delineated using the a 90m SRTM Digital Elevation Map (*50*), following Willcock *et al.* (*8*).
- The recreation validation data consisted of the 178 available country sheets for economic and employment impact of travel & tourism of the World Travel and Tourism Council for 2019 (*i.e.* pre-Covid), providing the estimated total GDP of Tourism and Travel in US$. It also contains estimates for the proportion spent on business and leisure, and the proportion that is from domestic and international tourism. Three of the five recreation models represent leisure-oriented local access to nature, including gravity models (*51*). Therefore, to use validation data comparable to our modelled outputs, we multiplied the Tourism and Travel GDP with the proportions for leisure and domestic to get to 'GDP of domestic recreation for leisure'.
- The AG carbon validation data is a combination of pan-tropical biomass in forest plots from ForestPlots.Net (*28*), and from the United Kingdom assessment of carbon in all forest estates (*20*). By using both data sets, we are able to validate the models in both temperate and tropical contexts. See Avitabile *et al.* (*28*) and Hooftman *et al.* (*20*) for further details.
- The fuelwood and forage production validation data are country-level statistics from FAO for 2019 (195 countries available) and 2018 (208 countries) respectively (Figure S2).

Each data set has associated uncertainties (*8*) but after an extensive review of data, we identified these as the best-suited reference values for validation (i.e. metrics that corresponded most closely to those modelled, have been published in the peer-reviewed literature and/or widely accepted as an authoritative source [e.g. FAO statistics], and are available globally or for a large number of countries). Both model and validation data are normalised (see below) to ensure comparability and remove unavoidable differences in exact units. The validation data are as independent as feasibly possible from the models; however, due to data deficiency, some aspects of an individual model may have been trained with local census data which could, in part, relate to validation data. For example, gridded livestock from FAO and Gilbert *et al.* (*43*) are trained on various regional census data, some of which may have been included in the national-scale forage production validation data, and some of the plots from Avitabile *et al.* (*28*) may have been used to estimate the carbon stocks per land cover class per ecofloristic zone as used in ARIES (*23*).

Table 3: The empirical validation data sets used in this study (mapped in SI-4). ES models need to be evaluated against the real world to determine if they are able to provide sufficiently accurate information for regional- or local-scale policy- and decision-making. Since the 'true value' can never be absolutely determined, acceptable reference values must be used. Empirical data can be used as reference values to evaluate ES model accuracy (*8*). Although such reference values are likely to have errors associated with them and may not be totally representative of the true values (*8*), this approach is widely accepted in environmental sciences (*52*).

Service	Validation Set	Description	Original Resolution	Details
Water Supply	Global Runoff Data Centre: river discharges	3746 selected stations. Mean annual water flow per hectare catchment ($m^2\ ha^{-1}$)	Catchments as polygons (SI-4)	Selected on still running after 2000 and containing at least 25 years of data. bafg.de/GRDC/EN/Home
Recreation	WTCC: Tourism Economic Impact Reports per country	Total GDP of domestic recreation for leisure in US$ for 178 countries	Country (GAUL-2) polygons.	Country sheets for 2019, calculated as recreation GDP contribution (US$) x [% domestic spending x % leisure spending]. wttc.org/Research/Economic-Impact.
AG carbon	1. Pan-tropical biomass in forest plots	ABG stock in tonnes per hectare for 14,478 forest plots[35]	Point data	Tropical region. Pan-tropical biomass reference data lucid.wur.nl/datasets/high-carbon-ecosystems.
	2. United Kingdom carbon in forest estates	Mean ABG stocks in tonnes per hectare from 1606 estates.	Forest polygons	Temperate region. Modified data taken from (20), original data from UK Forest Research
Fuelwood production	FAOStat: Wood fuel per country	Wood fuel in m^3, summed non-coniferous & coniferous for 2019 for 195 countries	Country (GAUL-2) polygons.	fao.org/faostat/en/#data/FO
Forage production	FAOStat: Livestock per country	Summed livestock units per country for 2018 for 208 countries	Country (GAUL-2) polygons.	Animals are: Asses, Buffaloes, Camels, Cattle, Chickens, Goats, Horses, Mules, Pigs & Sheep. fao.org/faostat/en/#data/EK

3. Model postprocessing and normalisation

General model output postprocessing included projecting to WGS 1984, rescaling and clipping to the specified extent (step 1), as well as detecting and masking no-data values. The latter was especially applicable for forest only biomass/carbon data sets (Table 2), as well for sea/large water bodies in various model outputs. In making ensembles, true zeros contribute to the average, whereas no-data are ignored. Postprocessing of ARIES and Co$ting Nature model outputs with additional data (marked † in Table 2) is discussed in SI-1-4. This includes the procedure of among tile rescaling of Co$ting Nature, as the framework produces outputs in 10°-tiles which are individually normalised. Therefore, tiles need to be rescaled using other global-scale estimates (SI-1-4). AG carbon model output postprocessing with MODIS land cover (*53*) masks into forage production and fuelwood outputs following (*8*) as detailed in SI-3.

To ensure comparability among model outputs, we standardised by normalising each individual model output prior to making ensembles. This normalisation followed (*13*, *20*) and allowed us to address differences in units among models, such as monetary benefit transfer vs. satellite-based tree cover densities or water run-off, and negates the need for conversion factors (e.g. between biomass and AG carbon). To avoid impacts of extreme values without eliminating these data-points, we employed a double-sided Winsorising protocol for normalisation (*20*), using the values associated with the 2.5% and 97.5% percentiles to define the minimum (0) and maximum (1) values (values below or above these percentiles became 0 or 1 respectively). Winsorising loses the extremes and so does curtail skew, but avoids influences of very large and very small values (*20*). The Winsorising procedure can be found can be found at our GitHub account (github.com/GlobalEnsembles/Winsorising), both as Matlab and arcpy coding. The validation data sets were subjected to the same Winsorising protocol. It must be noted that, even when modelling the same ES, many of the ES models estimate different constructs to some extent, often with varying units (Table 2). However, since our statistical analyses focused on relative ranking, it is unlikely that these uncertainties impacted our findings greatly [see (*8*) for a full discussion].

4. Generating spatial ecosystem service ensembles.

The procedures to generate different types of ES model ensembles are discussed in Hooftman *et al*. (*20*). Here we focus on an unweighted ensemble, which is the median value of the model outputs calculated per grid cell. A selection of weighted methods developed by (*20*) (including mean, PCA, correlation coefficient, regression to the median, leave-one-out cross validation, and log-likelihood approaches) are reported in SI-7. These alternative ensemble approaches show consistent patterns and comparable accuracy to the relatively simple median ensemble.

For recreation, AG carbon, fuelwood and forage production our ensembles were based on per-grid cell estimates of the respective model outputs. Here, models for AG carbon, fuelwood and forage production are comparable point-based estimates of local resources, although differing in complexity and initial assumptions (*8*, *20*). Additionally, our recreation ensemble comprises different modelling methods which provide comparable predictions of potential recreational pressure: observations [Photo uploads: (*24*) and (*25*)], population movement through gravity functions [(*2*) and (*23*)], and benefit transfer (*38*); see SI-1-4 for a full discussion. For water supply, our ensembles are accumulated flow estimates following the global HydroSHEDS catchments definition (*54*). For grid cells, ensembles were created using ArcPro Cell Statistics module – with the median or standard deviation as the input statistic. Due to the way certain models accumulated water flows (WaterWorld, Aquastat) a per-grid cell approach was not possible for water supply, so the sum of grid values within catchment polygons was calculated for each catchment. In the case of accumulated flow models (WaterWorld, Aquastat), we used the maximum value per polygon assumed to be the flow out of the HydroSHED pour point. Since HydroSHEDS information do not contain the spatial location of the exact pour point we could not correct for differences in routing information as we do for the GRDC validation catchments (Step 5). We employed a forced 0.001° grid size to minimise edge effects.

As all models are normalised to the same 0-1 scale, calculations do not require any additional scaling factors. The spatial representation of the ensembles and variation are generated on the same extent and grid as described under Step 1, and can be downloaded from the Environmental Informatics Data Centre (https://doi.org/10.5285/bd940dad-9bf4-40d9-891b-161f3dfe8e86). The water ensembles are there available as HydroSHED (*54*) defined accumulated water flow (vector format), the other four ES as geotiffs (raster format). Since not all model outputs are globally comprehensive, variation is expressed by a standard error of means as ($\frac{\sigma_{(x)}}{\sqrt{n_{(x)}}}$), instead of standard deviation (σ), with n the number of models per grid cell x. The ensembles are renormalised to represent the full 0-1 range.

5. Validation of accuracy

After creating the ensembles, the model and ensemble outputs were calculated at the spatial resolution of the validation data. For recreation, fuelwood and forage production, the validation data are available on a per-country basis, so this was done by calculating the sum of all model ensemble grid cells within countries. Country definitions followed FAO Global Administrative Unit Layers (GAUL) level 2 with 2014 definition. This map includes separate polygons for overseas territories. When overseas territories were treated separately in one of the validation data sets (*e.g.* Martinique [FR] or British Virgin Island [UK]) those values were extracted as separate data-points from the ensembles. We refer to all these spatial units as 'countries', although not all units have that designation. For each individual model, outputs were obtained for each country polygon with the ArcGIS Zonal tool with a forced 0.001° grid size to minimise edge effects – *i.e.* all predicted values were obtained by down-sampling into 0.001° grid cells. For AG carbon plots, the point-based location of the forest plot was used as the mean value of underlying 0.001° gridcells. For grid-based water supply estimates the sum of grid values per watershed polygon was employed. In the case of accumulated flow models (WaterWorld, Aquastat), we corrected for potential small scale differences in flow routing among these models by taking the maximum flow value within a 0.041665° range (5 cell widths) of the GRDC reported location of the weir station (*20*), without exceeding the aligned watersheds. We note that, these validation data are diverse (Table 3), being collected using a range of methods of varying reliability, including: expert opinion (e.g. country-level statistics from the FAO) and biophysical measurement (e.g. tree inventory plots, and weir data on water flow). As such, each dataset has associated uncertainties (*55*) but, since the 'true value' can never be absolutely determined, provides useful reference values for validation (*8*, *13*, *52*). However, given that the datasets covered a wide range of methods and our focus was on ranked correlative relationships (below), there is unlikely to be systematic bias and so data quality issues should have a low impact on our results. We refer to references (*8*) and (*13*) for a full discussion of ES model validation.

To create ES ensemble proxy services, we followed the procedure as above – *e.g.* AG carbon summed per country to compare to national-scale validation data; recreation, forage production and fuelwood summed within catchments (for comparison to global runoff data) and at the point location of the forest plots (for comparison to AG carbon data). To be able to use accumulated water flow as proxy for country-validated services we split the HydroSHEDS by countries, generating sub-catchments where they crossed borders. Following this, data extraction and ensemble procedure was followed anew as described above. Similarly, for forest plot locations water flow ensembles were generated for the plot locations only.

Ensemble, bundle and model output accuracy was assessed following the inverse of the deviance ($D^{\downarrow}$) as was developed in (*8*) following:

$$D^{\downarrow} = 1 - \left(\frac{1}{n} \times \sum_{x}^{n} \left| X_{(x)} - Y_{(x)} \right| \right) \qquad \text{Eqn. 1}$$

in which n = the number of spatial data points, x a spatial data point, X(x) the normalised validation value for x, and Y(x) the normalised value for the model or ensemble tested.

We also conducted rank-order comparisons using Spearman's ρ as an accuracy measure, which showed consistent results (SI-5).

To allow statistical comparisons we bootstrapped with 1000 runs for 10% of the data sets (AG carbon, water supply) or 100 data-points (country validations) reporting the mean and standard deviation across these bootstraps. We tested all accuracies within the same bootstrap run, allowing pairwise comparisons. We assessed accuracy differences with pairwise t-tests (Matlab *ttest*-tool). The mean of pairwise differences per run is generally larger than the difference between the averaged accuracies as shown in Figure 3. The pairwise combinations included median accuracy among models [*i.e.* indicating a random pick among models (*20*)], the median ensemble and the median ensembles of the other four service as proxies. Since we used the same statistical test five times per service per comparison, we employed a Hochberg's step-up correction (*56*) to account for multiple tests on the resulting average p-values. Hochberg's step-up correction is seen as more powerful than Sidak, Bonferroni and Holms correction methods, which are known to underestimate true effects (*56*). A comparison with six other approaches to creating ensembles from (*20*) are reported in SI-7.

6. Spatial comparison of ensemble accuracy with development and equality per country

We explored possible drivers of the spatial variation of ES ensemble accuracy, testing if ensembles are more accurate in more economically developed countries with relatively higher levels of data, research and model development. We used the following metrics:

- The Human Development Index (HDI) of 2019, as metric developed by the United Nations Development Programme being a summary measure of proxies for three important ends of development: access to health, education, and goods (*57*). Downloaded from hdr.undp.org/en/indicators/137506.
- The following World Development indicators were downloaded from The World Bank (databank.worldbank.org/home.aspx) using 2018 data (except GDP per capita) or the latest available entry before:
    - GDP per capita downloaded from World Bank 2019 in US$ Purchasing Power Parity, supplemented for missing countries by CIA data for 2018 (cia.gov/the-world-factbook/field/real-gdp-per-capita/country-comparison).
    - Income Equality following the Gini index measuring the extent to which the distribution of income among households within an country deviates from an equal distribution.
    - The number of researchers engaged in research and development (R&D), expressed as per million.
    - Gross domestic expenditures on research and development (R&D), expressed as a percent of GDP.

After exporting all above outputs to Matlab v7.14.0.739 we correlated these metrics one by one (*Metric*) with the per-validation point accuracy of the median ensemble, calculated as the inverse of deviance per point ($D^{\downarrow}_{(x)} = \left(1 - \left|X_{(x)} - Y_{(x)}\right|\right)$), using a SS-type I model with the Matlab *Anovan* tool:

$$D^{\downarrow}_{(x)} \sim \beta_0 + \beta_1 Auto_{(x)} + \beta_2 Metric_{(x)} + \varepsilon \qquad \text{Eqn. 2}$$

in which $D^{\downarrow}_{(x)}$ is the accuracy for polygon *x*, with effect sizes *β* and error *ε*.

We incorporated a correction for potential spatial autocorrelation through inclusion of a covariate (*Auto*) prior to estimating the correlation of the metric of interest, describing relatedness between individual outputs in deviance with the Euclidean distances among centroids of polygons/points (*13*, *58*). We used a maximum spatial autocorrelation effect range of 5°. To equalise degrees of freedom across services and avoid high degrees of freedom (df) inflation of F-values for AG carbon and water supply – resulting in near-zero p-values even for very weak effects – an iteration method was used taking a standard sample size of 178 datapoints (the minimum N across services). Not setting a default number of bootstraps, we used a convergence iterations method, stopping the iterations after the mean Sum of Squares of each factor over all iterations will not have changed by more than 0.05% with an extra iteration, consistently for 25 tries sequentially (see codes on github.com/GlobalEnsembles). Furthermore, we explicitly test for potential higher accuracy in more economically developed countries using a one-sided p-value distribution (two-sided is reported in SI-5). The presented F-values themselves are mirrored accordingly to represent the one-sided significance distribution. Since, for each ES, all metrics and the interaction (Eqn. 3) are calculated for the identical set of $D^{\downarrow}_{(x)}$ per point and hence the spatial autocorrelation among those, we employed a Hochberg's step-up correction (*56*) of significance to account for the use of 8 tests, as in step 5. Identical tests using Spearman's *ρ* as accuracy measure are reported in SI-5.

Since individual wealth may be better represented by the distribution of wealth around the mean (*i.e.* GDP per capita), we also ran Eqn. 2 as a two factor interaction model for GDP per capita and income equality with type I Sum of Squares between spatial autocorrelation and the tested factors and type III among factors and *interaction* following:

$$D_{(x)} \sim \beta_0 + \beta_1 Auto_{(x)} + \{\beta_2 Equity_{(x)} + \beta_2 Equality_{(x)} + Interaction + \varepsilon\} \qquad \text{Eqn. 3}$$

References

1. R. Chaplin-Kramer, R. P. Sharp, C. Weil, E. M. Bennett, U. Pascual, K. K. Arkema, K. A. Brauman, B. P. Bryant, A. D. Guerry, N. M. Haddad, M. Hamann, P. Hamel, J. A. Johnson, L. Mandle, H. M. Pereira, S. Polasky, M. Ruckelshaus, M. R. Shaw, J. M. Silver, A. L. Vogl, G. C. Daily, Global modeling of nature's contributions to people. *Science*. **366**, 255–258 (2019).
2. R. Chaplin-Kramer, R. A. Neugarten, R. P. Sharp, P. M. Collins, S. Polasky, D. Hole, R. Schuster, M. Strimas-Mackey, M. Mulligan, C. Brandon, S. Diaz, E. Fluet-Chouinard, L. J. Gorenflo, J. A. Johnson, C. M. Kennedy, P. W. Keys, K. Longley-Wood, P. B. McIntyre, M. Noon, U. Pascual, C. Reidy Liermann, P. R. Roehrdanz, G. Schmidt-Traub, M. R. Shaw, M. Spalding, W. R. Turner, A. van Soesbergen, R. A. Watson, Mapping the planet's critical natural assets. *Nat. Ecol. Evol.* **17**, 1–11 (2022).
3. C. Brandon, K. Brandon, A. Fairbrass, R. Neugarten, Integrating Natural Capital into National Accounts: Three Decades of Promise and Challenge. *Rev. Environ. Econ. Policy*. **15**, 134–153 (2021).
4. R. Chaplin-Kramer, K. A. Brauman, J. Cavender-Bares, S. Díaz, G. T. Duarte, B. J. Enquist, L. A. Garibaldi, J. Geldmann, B. S. Halpern, T. W. Hertel, C. K. Khoury, J. M. Krieger, S. Lavorel, T. Mueller, R. A. Neugarten, J. Pinto-Ledezma, S. Polasky, A. Purvis, V. Reyes-García, P. R. Roehrdanz, L. J. Shannon, M. R. Shaw, B. B. N. Strassburg, J. M. Tylianakis, P. H. Verburg, P. Visconti, N. Zafra-Calvo, Conservation needs to integrate knowledge across scales. *Nat. Ecol. Evol. 2021 62*. **6**, 118–119 (2021).
5. C. Wyborn, M. C. Evans, Conservation needs to break free from global priority mapping. *Nat. Ecol. Evol. 2021 510*. **5**, 1322–1324 (2021).
6. IOS, "ISO 5725-1:1994 Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions" (1994), (available at https://www.iso.org/standard/11833.html).
7. W. E. Walker, P. Harremoës, J. Rotmans, J. P. van der Sluijs, M. B. A. van Asselt, P. Janssen, M. P. Krayer von Krauss, Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integr. Assess.* **4**, 5–17 (2003).
8. S. Willcock, D. A. P. P. Hooftman, S. Balbi, R. Blanchard, T. P. Dawson, P. J. O'Farrell, T. Hickler, M. D. Hudson, M. Lindeskog, J. Martinez-Lopez, M. Mulligan, B. Reyers, C. Shackleton, N. Sitas, F. Villa, S. M. Watts, F. Eigenbrod, J. M. Bullock, A Continental-Scale Validation of Ecosystem Service Models. *Ecosystems*. **22** (2019), doi:10.1007/s10021-019-00380-y.
9. R. A. Neugarten, M. Honzak, P. Carret, K. Koenig, L. Andriamaro, C. A. Cano, H. S. Grantham, D. Hole, D. Juhn, M. McKinnon, A. Rasolohery, M. Steininger, T. M. Wright, W. R. Turner, Rapid Assessment of Ecosystem Service Co-Benefits of Biodiversity Priority Areas in Madagascar. *PLoS One*. **11**, e0168575 (2016).
10. G. Schmidt-Traub, National climate and biodiversity strategies are hamstrung by a lack of maps. *Nat. Ecol. Evol. 2021 510*. **5**, 1325–1327 (2021).
11. L. A. Bruijnzeel, M. Mulligan, F. N. Scatena, Hydrometeorology of tropical montane cloud forests: emerging patterns. *Hydrol. Process.* **25**, 465–498 (2011).
12. J. W. Redhead, L. May, T. H. Oliver, P. Hamel, R. Sharp, J. M. Bullock, National scale evaluation of the InVEST nutrient retention model in the United Kingdom. *Sci. Total Environ.* **610–611**, 666–677 (2018).
13. S. Willcock, D. A. P. Hooftman, R. Blanchard, T. P. Dawson, T. Hickler, M. Lindeskog, J. Martinez-Lopez, B. Reyers, S. M. Watts, F. Eigenbrod, J. M. Bullock, Ensembles of ecosystem service models can improve accuracy and indicate uncertainty. *Sci. Total Environ.* **747**, 141006

(2020).

14. S. Willcock, D. Hooftman, N. Sitas, P. O'Farrell, M. D. Hudson, B. Reyers, F. Eigenbrod, J. M. Bullock, Do ecosystem service maps and models meet stakeholders' needs? A preliminary survey across sub-Saharan Africa. *Ecosyst. Serv.* **18**, 110–117 (2016).

15. M. B. Araújo, M. New, Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**, 42–47 (2007).

16. L. Olander, S. Polasky, J. S. Kagan, R. J. Johnston, L. Waigner, D. Saah, L. Maguire, J. Boyd, D. Yoskowitz, So you want your research to be relevant? Building the bridge between ecosystem services research and practice. *Ecosyst. Serv.* **26**, 170–182 (2017).

17. H. Suich, C. Howe, G. Mace, Ecosystem services and poverty alleviation: A review of the empirical links. *Ecosyst. Serv.* **12**, 137–147 (2015).

18. IPBES, E. S. Brondizio, J. Settele, S. Díaz, H. T. Ngo, Eds. (Bonn, Germany, 2019; https://ipbes.net/global-assessment), pp. 1–1148.

19. IPBES, in *Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*, S. Ferrier, K. N. Ninan, P. Leadley, R. Alkemade, L. A. Acosta, H. R. Akçakaya, L. Brotons, W. W. L. Cheung, V.Christensen, K. A. Harhash, J. Kabubo-Mariara, C. Lundquist, M. Obersteiner, H. Pereira, G. Peterson, R. Pichs-Madruga, N. Ravindranath, C. Rondinini, B. A. Wintle, Eds. (Secretariat of the Intergovernmental Platform for Biodiversity and Ecosystem Services, Bonn, Germany, 2016; https://www.ipbes.net/sites/default/files/downloads/pdf/2016.methodological_assessment_report_scenarios_models.pdf), p. 370.

20. D. A. P. Hooftman, J. M. Bullock, L. Jones, F. Eigenbrod, J. I. Barredo, M. Forrest, G. Kindermann, A. Thomas, S. Willcock, Reducing uncertainty in ecosystem service modelling through weighted ensembles. *Ecosyst. Serv.* **53**, 101398 (2022).

21. K. He, L. Yu, K. K. Lai, Crude oil price analysis and forecasting using wavelet decomposed ensemble model. *Energy*. **46**, 564–574 (2012).

22. B. P. Bryant, M. E. Borsuk, P. Hamel, K. L. L. Oleson, C. J. E. Schulp, Transparent and feasible uncertainty assessment adds value to applied ecosystem services modeling. *Ecosyst. Serv.* **33**, 103–109 (2018).

23. F. Villa, K. J. Bagstad, B. Voigt, G. W. Johnson, R. Portela, M. Honzák, D. Batker, A methodology for adaptable and robust ecosystem services assessment. *PLoS One*. **9**, e91001 (2014).

24. R. Sharp, J. Douglass, S. Wolny, K. Arkema, J. Bernhardt, W. Bierbower, N. Chaumont, D. Denu, D. Fisher, K. Glowinski, R. Griffin, G. Guannel, A. Guerry, J. Johnson, P. Hamel, C. Kennedy, C. K. Kim, M. Lacayo, E. Lonsdorf, L. Mandle, L. Rogers, J. Silver, J. Toft, G. Verutes, A. L. Vogl, S. Wood, K. Wyatt, "InVEST 3.11.0.post88+ug.gbbddbb6 User's Guide" (2020).

25. M. Mulligan, in *Impact of Climate Change on Water Resources in Agriculture*, C. . Zolin, R. de A. R. Rodrigues, Eds. (CRC Press, Boca, Raton, 2015), pp. 184–204.

26. M. Potschin-Young, R. Haines-Young, C. Görg, U. Heink, K. Jax, C. Schleyer, Understanding the role of conceptual frameworks: Reading the ecosystem service cascade. *Ecosyst. Serv.* **29**, 428–440 (2018).

27. UK National Ecosystem Assessment, "The UK National Ecosystem Assessment: Synthesis of the Key Findings" (Cambirdge, UK, 2011).

28. V. Avitabile, M. Herold, G. B. M. Heuvelink, S. L. Lewis, O. L. Phillips, G. P. Asner, J. Armston, P. S. Ashton, L. Banin, N. Bayol, N. J. Berry, P. Boeckx, B. H. J. de Jong, B. Devries, C. A. J. Girardin, E. Kearsley, J. A. Lindsell, G. Lopez-Gonzalez, R. Lucas, Y. Malhi, A. Morel, E. T. A. Mitchard, L. Nagy, L. Qie, M. J. Quinones, C. M. Ryan, S. J. W. Ferry, T. Sunderland, G. V. Laurin, R. C. Gatti, R. Valentini, H. Verbeeck, A. Wijaya, S. Willcock, An integrated pan-tropical biomass map using multiple reference datasets. *Glob. Chang. Biol.* **22**, 1406–1420 (2016).

29. R. Spake, R. Lasseur, E. Crouzat, J. M. Bullock, S. Lavorel, K. E. Parks, M. Schaafsma, E. M. Bennett, J. Maes, M. Mulligan, M. Mouchet, G. D. Peterson, C. J. E. Schulp, W. Thuiller, M. G. Turner, P. H. Verburg, F. Eigenbrod, Unpacking ecosystem service bundles: Towards predictive mapping of synergies and trade-offs between ecosystem services. *Glob. Environ. Chang.* **47**,

37–50 (2017).
30. J. W. Redhead, C. Stratford, K. Sharps, L. Jones, G. Ziv, D. Clarke, T. H. Oliver, J. M. Bullock, Empirical validation of the InVEST water yield ecosystem service model at a national scale. *Sci. Total Environ.* **569**–**570**, 1418–1426 (2016).
31. T. Hao, J. Elith, J. J. Lahoz-Monfort, G. Guillera-Arroita, Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography (Cop.).* **43**, 549–558 (2020).
32. C. F. Dormann, J. M. Calabrese, G. Guillera-Arroita, E. Matechou, V. Bahn, K. Bartoń, C. M. Beale, S. Ciuti, J. Elith, K. Gerstner, J. Guelat, P. Keil, J. J. Lahoz-Monfort, L. J. Pollock, B. Reineking, D. R. Roberts, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, S. N. Wood, R. O. Wüest, F. Hartig, Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecol. Monogr.* **88**, 485–504 (2018).
33. H. Ding, J. M. Bullock, *A Guide to Selecting Ecosystem Service models for Decision-Making: Lessons from Sub-Saharan Africa* (Washington DC, USA, 2018; https://www.wri.org/research/guide-selecting-ecosystem-service-models-decision-making-lessons-sub-saharan-africa).
34. C. P. Wong, B. Jiang, A. P. Kinzig, K. N. Lee, Z. Ouyang, Linking ecosystem characteristics to final ecosystem services for public policy. *Ecol. Lett.* **18**, 108–118 (2014).
35. S. Willcock, J. Martínez-López, D. A. P. D. A. P. Hooftman, K. J. K. J. Bagstad, S. Balbi, A. Marzo, C. Prato, S. Sciandrello, G. Signorello, B. Voigt, F. Villa, J. M. Bullock, I. N. I. N. Athanasiadis, Machine learning for ecosystem services. *Ecosyst. Serv.* (2018), doi:10.1016/j.ecoser.2018.04.004.
36. A. Ruesch, H. K. Gibbs, New IPCC Tier1 Global Biomass Carbon Map For the Year 2000. *Available online from Carbon Dioxide Inf. Anal. Cent. [http//cdiac.ornl.gov/ - accessed 15/01/12], Oak Ridge Natl. Lab. Oak Ridge, Tennessee.* (2008).
37. B. Smith, D. Wårlind, A. Arneth, T. Hickler, P. Leadley, J. Siltberg, S. Zaehle, Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. *Biogeosciences*. **11**, 2027–2054 (2014).
38. R. Costanza, R. de Groot, P. Sutton, S. van der Ploeg, S. J. Anderson, I. Kubiszewski, S. Farber, R. K. Turner, Changes in the global value of ecosystem services. *Glob. Environ. Chang.* **26**, 152–158 (2014).
39. R. J. Scholes, “The South African 1: 250 000 maps of areas of homogeneous grazing potential” (1998).
40. F. Gassert, M. Landis, M. Luck, P. Reig, T. Shiao, Aqueduct Global Maps 2.1. *World Resour. Inst.* (2014).
41. J. I. Barredo, J. San Miguel, G. Caudullo, L. Busetto, “A European map of living forest biomass and carbon stock: Executive report, EUR 25730 EN” (Luxembourg, 2012), , doi:10.2788/780.
42. S. Quegan, T. Le Toan, J. Chave, J. Dall, J. F. Exbrayat, D. H. T. Minh, M. Lomas, M. M. D’Alessandro, P. Paillou, K. Papathanassiou, F. Rocca, S. Saatchi, K. Scipal, H. Shugart, T. L. Smallman, M. J. Soja, S. Tebaldini, L. Ulander, L. Villard, M. Williams, The European Space Agency BIOMASS mission: Measuring forest above-ground biomass from space. *Remote Sens. Environ.* **227**, 44–60 (2019).
43. M. Gilbert, G. Nicolas, G. Cinardi, T. P. Van Boeckel, S. Vanwambeke, W. G. R. Wint, T. P. Robinson, “Gridded Livestock of the World in 2010 (5 minutes of arc) V3” (2018), (available at https://dataverse.harvard.edu/dataverse/glw_3).
44. N. L. Harris, D. A. Gibbs, A. Baccini, R. A. Birdsey, S. de Bruin, M. Farina, L. Fatoyinbo, M. C. Hansen, M. Herold, R. A. Houghton, P. V. Potapov, D. R. Suarez, R. M. Roman-Cuesta, S. S. Saatchi, C. M. Slay, S. A. Turubanova, A. Tyukavina, Global maps of twenty-first century forest carbon fluxes. *Nat. Clim. Chang. 2021 113*. **11**, 234–240 (2021).
45. European Commission, Joint Research Centre, “Forest Biomass Map of Europe” (2020), (available at http://data.europa.eu/89h/d1fdf7aa-df33-49af-b7d5-40d226ec0da3).
46. M. Mulligan, WaterWorld: a self-parameterising, physically based model for application in

data-poor but problem-rich environments globally. *Hydrol. Res.* **44** (2013) (available at http://hr.iwaponline.com/content/44/5/748).
47. M. L. Noon, A. Goldstein, J. C. Ledezma, P. R. Roehrdanz, S. C. Cook-Patton, S. A. Spawn-Lee, T. M. Wright, M. Gonzalez-Roglich, D. G. Hole, J. Rockström, W. R. Turner, Mapping the irrecoverable carbon in Earth's ecosystems. *Nat. Sustain. 2021 51*. **5**, 37–46 (2021).
48. G. E. Kindermann, I. McCallum, S. Fritz, M. Obersteiner, A global forest growing stock, biomass and carbon map based on FAO statistics. *Silva Fenn.* **42**, 387–396 (2008).
49. S. A. Spawn, C. C. Sullivan, T. J. Lark, H. K. Gibbs, Harmonized global maps of above and belowground biomass carbon density in the year 2010. *Sci. Data 2020 71*. **7**, 1–22 (2020).
50. A. Jarvis, H. I. Reuter, A. Nelson, E. Guevara, "Hole-filled seamless SRTM data V4" (2008), (available at http://srtm.csi.cgiar.org).
51. M. Schläpfer, L. Dong, K. O'Keeffe, P. Santi, M. Szell, H. Salat, S. Anklesaria, M. Vazifeh, C. Ratti, G. B. West, The universal visitation law of human mobility. *Nat. 2021 5937860*. **593**, 522–527 (2021).
52. S. Eker, E. Rovenskaya, M. Obersteiner, S. Langan, Practice and perspectives in the validation of resource management models. *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-07811-9.
53. P. D. Broxton, X. Zeng, D. Sulla-Menashe, P. A. Troch, A Global Land Cover Climatology Using MODIS Data. *J. Appl. Meteorol. Climatol.* **53**, 1593–1605 (2014).
54. B. Lehner, K. Verdin, A. Jarvis, New Global Hydrography Derived From Spaceborne Elevation Data. *Eos (Washington. DC).* **89**, 93–94 (2008).
55. A. Grainger, Difficulties in tracking the long-term global trend in tropical forest area. *Proc. Natl. Acad. Sci.* **105**, 818–823 (2008).
56. Y. Huang, J. C. Hsu, Hochberg's Step-Up Method: Cutting Corners Off Holm's Step-Down Method. *Biometrika*. **94**, 965–975 (2007).
57. E. A. Stanton, "The Human Development Index: A History. Working Paper Series Number 127" (Amherst, USA, 2007), (available at www.peri.umass.edu).
58. C. F. Dormann, J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. Daniel Kissling, I. Kühn, R. Ohlemüller, P. R. Peres-Neto, B. Reineking, B. Schröder, F. M. Schurr, R. Wilson, C. F. Dormann, J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. Daniel Kissling, I. Kühn, R. Ohlemüller, P. R. Peres-Neto, B. Reineking, B. Schröder, F. M. Schurr, R. Wilson, Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography (Cop.).* **30**, 609–628 (2007).
59. J. H. Goldstein, G. Caldarone, T. K. Duarte, D. Ennaanay, N. Hannahs, G. Mendoza, S. Polasky, S. Wolny, G. C. Daily, Integrating ecosystem-service tradeoffs into land-use decisions. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7565–70 (2012).
60. C. Byczek, P. Y. Longaretti, J. Renaud, S. Lavorel, Benefits of crowd-sourced GPS information for modelling the recreation ecosystem service. *PLoS One*. **13**, e0202645 (2018).
61. K. E. Saxton, W. J. Rawls, Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions. *Soil Sci. Soc. Am. J.* **70**, 1569 (2006).
62. C. DiMiceli, M. Carroll, R. Sohlberg, D. Kim, M. Kelly, J. Townshend, MOD44B MODIS/Terra Vegetation Continuous Fields Yearly L3 Global 250m SIN Grid V006. *Distrib. by NASA EOSDIS L. Process. DAAC* (2015), , doi:10.5067/MODIS/MOD44B.006.
63. R. de Groot, L. Brander, S. van der Ploeg, R. Costanza, F. Bernard, L. Braat, M. Christie, N. Crossman, A. Ghermandi, L. Hein, S. Hussain, P. Kumar, A. McVittie, R. Portela, L. C. Rodriguez, P. ten Brink, P. van Beukering, Global estimates of the value of ecosystems and their services in monetary units. *Ecosyst. Serv.* **1**, 50–61 (2012).
64. FAO, "Guidelines for the measurement of productivity and efficiency in agriculture" (Rome, Italy, 2018), (available at https://www.fao.org/3/ca6395en/ca6395en.pdf).
65. S. L. Lewis, G. Lopez-Gonzalez, B. Sonke, K. Affum-Baffoe, T. R. Baker, L. O. Ojo, O. L. Phillips, J. M. Reitsma, L. White, J. A. Comiskey, M. N. Djuikouo, C. E. N. Ewango, T. R. Feldpausch, A. C. Hamilton, M. Gloor, T. Hart, A. Hladik, J. Lloyd, J. C. Lovett, J. R. Makana, Y. Malhi, F. M.

Mbago, H. J. Ndangalasi, J. Peacock, K. S. H. Peh, D. Sheil, T. Sunderland, M. D. Swaine, J. Taplin, D. Taylor, S. C. Thomas, R. Votere, H. Woll, B. Sonké, E. Gloor, M. N. D. Kamdem, H. Wöll, Increasing carbon storage in intact African tropical forests. *Nature*. **457**, 1003–1006 (2009).

Acknowledgments

This work took place under the following UKRI grants: NE/W005050/1, NE/T00391X/1, ES/R009279/1, ES/T007877/1, ES/V004077/1, ES/R006865/1, as well as UKCEH project 08695 and NSF GRFP Grant number DGE-2139899. JML was funded by a María Zambrano postdoctoral grant by the Spanish Ministry of Universities and the Next Generation European Union funds. SW, DAPH, RAN, RCH & JMB conceived the project. DAPH, JIB, TH, GK, ARL, ML, & JML provided ES model descriptions and outputs. DAPH conducted all analyses. SW, DAPH, RAN, RCH & JMB wrote the manuscript, with comments from all other authors. The authors declare that they have no competing interests. We also thank the anonymous reviewers for their insightful comments on the manuscript. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.