Supporting Information to

Hooftman *et al.*

Ensemble outputs among contemporary ecosystem service models for water supply and aboveground carbon storage in the UK following 10 different methods

Contents:

	Page
• Details of the data-structure and nature of values	1
• Origin of individual model outputs	2
• Individual model output data processing	2
• Models and existing outputs used (Table 1)	3
• Calculation polygons and mapping	4
• Analytical approaches of generating Ensemble Models	4
• Quality control	7
• Funding Information	7
• References	7

Details of data-structure and nature of values

This data set contains:

1. 14 tiff-file raster layers including world-files of 10 different among-model ensemble approaches among normalised outputs of modelling frameworks for above ground Carbon stock –with a double option for two of approaches– together with a maps of variation among models and among ensembles.
2. One shapelayer for water supply service with polygons per watershed, each polygon containing these 14 estimates. Catchments are associated to a selection gauging stations of The National River flow Archive (nrfa.ceh.ac.uk/).

Naming: rasters: <Carbon_*Ensemble approach*>; shapefile: < UKWaterEnsembles>

Grid: Raster have an exact 1000m x 1000m grid size containing 1-band with 32-bit floating point grid values.

Projection: British National Grid transverse Mercator projection (EPSG 27700) with a 0.9996 scale factor and units in metres. Prior to our calculations all underlying model outputs were projected to this definition.

Area covered:

- Above ground carbon stock for Great Britain, Northern Ireland, including the Isles of Man, Scilly, and Wight, Orkney- and Shetland islands, Inner- and Outer Hebrides.
- Water supply covers 519 selected catchments on mainland Great Britain and Northern Ireland.

Units: Relative Service delivery, linearly normalised between 0 and 1 (see below).

Raster No Data value: -999

Origin of individual model outputs

In this data-set we present maps of different ES ensembles approaches originating from the collaborative EnsemblES project –*Using ensemble techniques to capture the accuracy and sensitivity of ecosystem service models*– of Bangor University, UK-CEH, and subcontractor Lactuca: Environmental Data analyses and Modelling, funded by the UKRI Landscape Decisions programme (project: NE/T00391X/1). The methods and depiction of these outputs are under consideration of publishing by Ecosystem Services (as *dd* 16/11/2021) as Hooftman *et al.* (2022), *Weighted Ensembles Reduce Uncertainty in Ecosystem Service Modelling*, which has a methodological focus and depicts the maps presented here in the Supplementary Information only.

We generated 10 different ensemble approaches among a variety of model framework outputs for above ground carbon stock and water supply in the UK. These model outputs were collated for this study according to their availability and to reflect different approaches to modelling ecosystem services (ES). Not all input models covered the whole area (see Table 1), *e.g.* some excluded Northern Ireland, Wales or Scotland, or are mainland GB only. The model frameworks are described in Table 1 below. Part of the model frameworks were generated for this work, but in majority these are existing outputs, freely available or under non-payment licenses. For all further details about individual model approaches, input data, bespoke model generation and their validation we refer to Hooftman *et al.* (2022).

For all the water ensembles approaches, the maps present the mean normalised value across jack-knifed runs as catchment polygons (N = 519, details see below). All carbon ensemble approaches are mapped as 1-km^2 grid cells. Here, for each ensemble approach, the estimated weights as calculated for the validation polygons (below) –averaged among jack-knife runs– were transferred to the full area, with the result aggregated to a 1 km^2 resolution based of an calculated grain of 1 hectare gridcells. In total, this carbon dataset has 253,802 cells that (partially) contain non-sea land cover. We transferred the weights calculated for the forests since running cross-validation approaches on over 250K data points would extremely time consuming to compute. However, we are aware of introducing a potential bias that could skew non-forested areas to lower values. Furthermore, we generated UK-scale maps of variation among the different input models and among the untrained ensemble approaches, by calculating the Standard Error of the Mean (SEM) among these spatial outputs.

Individual model output data processing

For each individual model outputs, listed in Table 1, predictions were obtained for each polygon in the validation datasets (see below) using the ArcGIS spatial analyst Zonal tool with a forced 2.5-meter grid size environmental setting to minimise edge effects; *i.e.* all predicted values were obtained by nearest point resampling into 2.5 × 2.5 meter gridcells. In most cases the modelled value per polygon was obtained by taking the sum of all constituent gridcell values, corrected for both actual grid size and the resampling to 2.5-meter. In the case of accumulated flow models, we corrected for potential small scale differences in flow routing among these models by taking the maximum flow value within both a 2 km range of the NRFA reported location of the gauging station and the polygon associated with that gauging station. To ensure comparability among model outputs, we standardised by normalising among the outputs for each individual model and for the validation data-sets. Prior to this step all outputs were area corrected as either mean carbon stock – or proxy thereof – per hectare or water supply per hectare of catchment. This normalisation followed Willcock *et al.* (2019), and allowed us to address differences in units among models (such as monetary benefit transfer vs. satellite-based tree cover densities or run-off, and equalised carbon and biomass). To avoid impacts of extreme values without eliminating data-points, we employed a double-sided Winsorising protocol for normalisation (Willcock *et al.* 2019), using the values associated to the 2.5% and 97.5% percentiles of number of datapoints to define the 0 and 1 values (values below or above these percentiles became 0 or 1 respectively). For each model, normalisation was done prior to creating ensembles.

Table. 1. Models and existing outputs used, with their original grain size and coverage; where applicable EIDC links are provided.

Model	Description	Grid size (grain)	Coverage
InVest v3.7.0[1†]	Carbon module: above ground stocks	25 × 25 meters	Full Area
	Water yield module: run-off per cell		Full Area
LPJ-GUESS[2,3†]	Vegetation biomass stocks per cell, mean for years 2009-2018	0.5° (≈ 46 × 46 km)	Full Area
	Water run-off per cell, mean for years 2009-2018		
LUCI[4†]	Above ground carbon stocks	10 × 10 meters	Great Britain
	Accumulated water run-off	5 × 5 meters	England & Wales
$-benefit transfer using The Economics of Ecosystems and Biodiversity database[5,6†]	Above ground carbon stock as monetary value	25 × 25 meters	Full Area
	Water run-off as monetary value per cell		
Aqueduct v2.1 Total Blue Water[7§]	Accumulated water run-off	138 flow areas	Full Area
ARIES k-Explorer[8‡]	Joined above and below ground carbon stocks	1-hectare	Full Area
Barredo *et al.* (2012)[§]	A European map of above ground biomass stocks	1 km^2	Full Area
Copernicus, Tree Cover Density[9§]	Proxy for carbon: tree Cover Density 2015 from MODIS satellite imagery.	20 × 20 meters	Full Area
DECIPHeR[10§], via EIDC: doi.org/10.5285/d770b12a-3824-4e40-8da1-930cf9470858	Accumulated water run-off through NRFA delineated catchment outlets, mean for years 1995-2015	387 catchments in common with selected	Great Britain
Grid-to-Grid[11§], via EIDC: doi.org/10.5285/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0	Accumulated water run-off, mean for years 1995-2015	1 km^2	Great Britain
Henrys *et al.* (2016)[§], via EIDC: doi.org/10.5285/9be652e7-d5ce-44c1-a5fc-8349f76f5f5c	Above ground carbon stocks	1 km^2	England
Kindermann *et al.* (2008)[§]	A global map of above ground forest biomass stocks	1 hectare	Full Area
National Forest Inventory (2018)[12†]	Woodland Land Cover Map[15] with above ground carbon stocks based on added Look-up table	20 × 20 meters	Great Britain
Scholes Growth Days[13,14†]	Proxy for water run off per cell: # Days precipitation exceeds evapotranspiration	1 km^2	Full Area
WaterWorld v2[15‡]	Accumulated water run-off	0.0083° (≈ 1 km^2)	Full Area

[†]Output generated for this work; [‡]online tool; [§]existing dataset; [1]Kareiva *et al.* (2011); [2]Smith *et al.* (2014); [3]Ahlström *et al.* (2015); [4]Thomas *et al.* (2020); [5]de Groot *et al.* (2012); [6]Costanza *et al.* (2014); [7]Gassert *et al.* (2015) [8]Martínez-López *et al.* (2019); [9]land.copernicus.eu/tree-cover-density/ status-maps/2015; [10]Coxon *et al.* (2019a; 2019b); [11]Bell *et al.* (2018a; 2018b); [12]Forestry Commission (2018); [13]Scholes (1998); [14]Willcock *et al.* (2019); [15]Mulligan (2013).

Calculation polygons and mapping

Calculations were done in polygons that contained the validation data – the validation data and validations themselves are not described here.

1) Our carbon stock validation dataset was provided by Forest Research and comprises species inventories in all forest estates in England and Scotland in 2019 (data-forestry.opendata.arcgis.com/). This data-set contains 201,143 forest compartments of varying size (mean: 4.4 hectares. median 1.6 hectares, ± 22.1). To generate lager units, the compartments were spatially joined into 2078 polygons of 'forest' that were separated if more than 25 meters distance from each other. For the maps presented here we recalculated from those forests to the full area.
2) Our water supply validation dataset comprised 519 hydrometric gauging stations from the National River Flow Archive (NRFA; nrfa.ceh.ac.uk), with associated catchments representing a variety of sizes distributed across the whole of the UK. From the 1598 potential catchments in NRFA (as present in November 2019), we selected those that were >100 km^2 to get a robust mean run-off from the catchments. In cases where multiple gauging stations were found along the same river, based on name, only the largest was chosen to avoid pseudoreplication. An additional set of 41 Welsh catchments was included which did not meet this size criterion. Wales contains mainly small catchments due its geography – mountain ranges close to the sea – and so we selected catchments >25 km^2 to avoid this part of the UK being underrepresented.

Analytical approaches of generating Ensemble Models

All our ensemble calculations followed the list below. To generate variation estimates allowing statistical comparison between models and ensembles, and among ensembles themselves – which is not presented here– we used a bagging approach in which we jack-knifed with 50% of the spatial data polygons for 250 runs. To assure like-for-like comparisons, per run all model and ensemble calculations and their comparisons were calculated on the same set of jack-knifed spatial data points. Afterwards mean averaged weights were calculated among runs. All codes for generating ensembles were written in Matlab v7.0.14.739 with statistical and parallel toolboxes available; Codes can be found in our GitHub account github.com/EnsemblesTypes/EnsemblesTypes.

1) We selected per run 50% of data points, in case of unequal data points (water) this was the ceiling. This selection would remain for all model and ensemble calculations. The remaining 50% of data-points was stored to be used in the trained ensembles procedures.
2) We calculated the **unweighted ensembles** by taking per data-point the mean ('*Mean Ensemble'*) respectively the median ('*Median Ensemble'*), *i.e.* a datapoint represents one validation polygon. See for unweighted method discussions *e.g.* Marmion *et al.* (2009), Grenouillet *et al.* (2011), Refsgaard *et al.* (2014) and Willcock *et al.* (2020).
3) Subsequently, we calculated the **untrained weighted ensembles** simulating the situation without any or reliable validation data present. Ensembles are of the general form:

$$E_{(x)} = \sum_i^n \left(\frac{\omega_i}{\sum_i^n \omega_i} \times Y_i \right)_{(x)} \qquad \text{Eq. 1}$$

with positive weights ω_i for model i of validation polygon x, weights ω_i are normalised to sum to 1, Y the modelled values for i per polygon, and n the total number of models per service.

This implies every model has one weight assigned which will be used to multiply all its containing datapoints with. The difference in approaches is how the weights are generated. After assessing weights, Eq. 1 is performed including normalisation of weights to sum to 1. Weights per runs are stored and averaged for use in spatial mapping.

We calculated the following untrained weighted ensembles:

a) The '*PCA ensemble'* with PCA as consensus axis, being a deterministic approach. The suggestion of using PCA's comes from Marmion *et al.* (2009) and Grenouillet *et al.* (2011). Deterministic here means that the result is an inherent property of the dataset, *i.e.* the statistical outcome is identical given the same dataset. Principal components were calculated using the Matlab *princomp*-tool, the weights per model *i* outputted to Eq. 1 were the loadings to the first –main– pca axis. So models with the better correlation to the consensus axis are

assigned higher weights.

b) The '*correlation coefficient ensemble*' is our second deterministic consensus approach. Here we calculated the full [model × model] correlation matrix using the Matlab *corrcoef*-tool. Following the weight per model was the mean correlation of that individual model with all other models, not including itself. Hence the higher general correlation to the other models, the more weight a model has. This technique was developed to have a second deterministic approach and can be seen as further way to minimise variance among models (Dormann *et al.* 2018).

c) The '*regression to the median ensemble*' is our first iterative consensus approach using log-likelihood regression (Dormann *et al.* 2018). Using multivariate regression we assess weights such that the summed results maximises the explanation of an comparator. The resulting regression coefficients are used as weights. In this case the comparator is the median ensemble, asking which contribution of models would be most closely result to the median. The regression contains no constant, hence it can be represented as: $[E_{ii} \sim \omega_1 Y_1 + \omega_2 Y_2 \ldots. + \omega_n Y_n]$.

This approach is iterative, parameter space is step-wise systematically explored improving the maximum log-likelihood until convergence is reached, *i.e.* no better solutions is found. Theoretically, different outcomes would be possible by redoing the calculation. However, the used *nlmefit* Matlab tool is of such quality, including multiple replications, that noticeable difference are absent – which could be caused by local but not absolute maximum log-likelihood in parameter space. Multi-variate regression to the median was done using the *nlmefit*-tool, maximising log-likelihood with 200 iterations: repeating the regression 200 times), an output tolerance of $1.0000e^{-4}$ and naïve priors. The resulting regression coefficients (ω_i) per model were the weights that were used in Eq. 1

d) The "*leave-one-out cross-validation ensemble*' is our second iterative consensus approach following direct recommendation in Dormann *et al.* (2018), with the difference of not omitting sets of data points but entire models one-by-one. As for the above this is done using a no constant multi-variation regression with the same *nlmefit*-tool, with the same settings and naïve priors. However, in this approach we loop through the model outputs. One-by-one, a regression is performed using a single model output as comparator and the remaining model outputs as explanatory variables. For model 1 such would be the regression representation $[Y_1 \sim \omega_2 Y_2 + \omega_3 Y_3 \ldots. + \omega_n Y_n]$. The regression coefficients (ω_i) are stored as consensus weights. After looping through all models – 9 water models or 10 carbon models –, the mean is taken of all regression coefficients per model as weights (excluding itself), *i.e.* this represents the weights that would generate the highest mean consensus with all models. The resulting mean consensus per model were the weights that were used in Eq. 1.

e) Models that are generated on smaller scales (*i.e.* with smaller gridcells) could be more accurate since the information per cell could better represent the local situation whereas larger gridcells could be more averaged across larger areas (Willcock *et al.* 2019; 2020): '*the grain size ensemble*'. To include this we generated an ensemble in which we penalised model outputs that are generated at coarser spatial resolutions (Willcock *et al.* 2016). The weights taken were: $\omega_i = \frac{1}{\log_{10}(\text{grain}_i)}$, for which the resulting weights were normalised afterwards to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$. These weights were the weights that were used in Eq. 1.

f) One might a priori place value on particular model characteristics and use these to create weights (Masson & Knutti 2011; Englund *et al.* 2017; Willcock *et al.* 2019): '*the distinct ensembles*'. For example, up- or down-weighting more distinct model types emphasising models that may contain processes not captured in others, or by penalising those models that

go against the convention (Grenouillet *et al.* 2011). Attribute weighting could be done for many attributes, which are largely arbitrary in use. To not focus on one attribute but many at the same them we choose an overall attribute assessment into groups based on 17 categories per model. The grouping statistic used is a pairwise Spearman's rank correlation among binary classifications. The goal was to generate 4 or 5 groups of different amounts of models with similar attributes. The attributes, the resulting trees and model grouping are beyond this document and are explained in Hooftman *et al.* (2022) The output variable we included in our weighting is the *distinctiveness factor*, representing how proportional representation of a group of models among all models. By upweighting distinctiveness, models in minority groups (g) are assigned a higher weight compared to majority groups as $\omega_i = \left(\frac{n^g}{n}\right)$ when upweighted with $n^g = i \in$ group g and n the number of models. Alternatively, consensus is sought so distinctiveness is downweighted, assigning higher weights to majority groups as $\omega_i = \left(\frac{n}{n^g}\right)$. In all cases resulting weights are normalised afterwards to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$. These weights were the weights that were used in Eq. 1.

5) Subsequently, we calculated **the trained weighted ensembles** simulating the situation in which validation data are partly present or present for a very similar area. So this approximates the standard Species distribution modelling techniques were data is split in a training and a testing set following Marmion *et al.* (2009), Thuiller *et al.* (2009; 2019) and Djengdoh *et al.* (2020). In these ensembles the 50% data points selected in step 1 and their accompanying validation data for the same points is to train the model. Subsequently, the resulting weights are multiplied with the second set of 50% of the data following Eq. 1, so the part of data that was not used for training. This ensemble is than tested for accuracy against the validation comparator belonging to this second set of data-points. By doing this in a jack-knife loop, a good representation of all possible combinations of selected data is provided and accuracy by chance is avoided.

   g) '*Accuracy weighted ensembles*' are based on the per model accuracies over the same 50s% of datapoints assigned to all ensembles. Model accuracy is not discussed here. This weighting is similar to AUC weighting as suggested in Marmion *et al.* (2009), Crossman *et al.* (2012), Grenouillet *et al.* (2011) and Dormann *et al.* (2018). These "trained" accuracies per model (either $D^{\downarrow}$ or Spearman ρ) are, after normalisation to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$, used as weights ($\omega_i$) in Eq. 1, with as Y_i the second set of datapoints not used in the training. Accuracy is assessed against the corresponding set of comparator validation data points not used in the training.

   h) The '*Log-likelihood regression*' is identical to unweighted regression to the median with the difference that models are not regressed against their median but against their corresponding validation data points represented as: $[V \sim \omega_1 Y_1 + \omega_2 Y_2 \ldots + \omega_n Y_n]$, with V the validator. The resulting regression coefficients (ω_i) per model are used, after normalisation to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$, as weights in Eq. 1 with as Yi the second set of datapoints not used in the training. Accuracy is assessed against the corresponding set of comparator validation data points not used in the training.

Quality control

All approaches and their outputs have been subject to internal full team review (the nine authors of Hooftman *et al.* 2022) and in external peer review for publication in Ecosystem Services. Most input model data are from peer reviewed or well-accepted sources.

Funding Information

This work took place under the 'EnsemblES' project –'*Using ensemble techniques to capture the accuracy and sensitivity of ecosystem service models*' funded by the UKRI Landscape Decisions programme (project: NE/T00391X/1).

References

Ahlström, A. *et al.* (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. doi.org/10.1126/science.aaa1668

Barredo, J.I. *et al.* (2012). *A European map of living forest biomass and carbon stock.* (European Commission, Joint Research Centre). op.europa.eu/en/publication-detail/-/publication/b9345574-a96f-4417-87ed-1a85d2252834/language-en

Bell, V.A. *et al.* (2018a). The MaRIUS-G2G datasets: Grid-to-Grid model estimates of flow and soil moisture for Great Britain using observed and climate model driving data. *Geosci. Data J.* **5**, 63–72. doi.org/10.1002/gdj3.55

Bell, V.A. *et al.* (2018b). *Grid-to-Grid model estimates of monthly mean flow and soil moisture for Great Britain (1891 to 2015): observed driving data [MaRIUS-G2G-Oudin-monthly].* [Data Set] (NERC Environmental Information Data Centre). doi.org/10.5285/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0

Costanza, R. *et al.* (2014). Changes in the global value of ecosystem services. *Glob. Environ. Change* **26**, 152–158. doi.org/10.1016/j.gloenvcha.2014.04.002

Coxon, G. *et al.* (2019a). DECIPHeR v1: Dynamic fluxEs and ConnectIvity for Predictions of HydRology. *Geosci. Model Dev.* **12**, 2285–2306. doi.org/10.5194/gmd-12-2285-2019

Coxon, G. *et al.* (2019b). *DECIPHeR model estimates of daily flow for 1366 gauged catchments in Great Britain (1962-2015) using observed driving data*. [Data Set] (NERC Environmental Information Data Centre). doi.org/10.5285/d770b12a-3824-4e40-8da1-930cf9470858

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. doi.org/10.1111/j.1472-4642.2011.00851.x

Diengdoh, V.L. *et al.* (2020). A validated ensemble method for multinomial land-cover classification. *Ecol. Inform.* **56**, 101065. doi.org/10.1016/j.ecoinf.2020.101065

Dormann, C.F. *et al.* (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecol. Monogr.* **88**, 485–504. doi.org/10.1002/ecm.1309

Englund, O., Berndes, G. & Cederberg, C. (2017). How to analyse ecosystem services in landscapes–A systematic review. *Ecol. Indic.* **73**, 492–504. doi.org/10.1016/j.ecolind.2016.10.009

Forestry Commission, United Kingdom. (2018). *National Forest Inventory Woodland GB 2018.* [Data Set] (Forestry Commission Open Data). data-forestry.opendata.arcgis.com/datasets/d3d7bfba1cba4a3b83a948f33c5777c0_0

Gassert, F. *et al.* (2015). *Aqueduct Global Maps 2.1.* [Data Set] (World Resources Institute). wri.org/resources/data-sets/aqueduct-global-maps-21-data

Grenouillet, G. *et al.* (2011). Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography* **34**, 9–17. doi.org/10.1111/j.1600-0587.2010.06152.x

de Groot, R. *et al.* (2012). Global estimates of the value of ecosystems and their services in monetary units. *Ecosyst. Serv.* **1**, 50–61. https://doi.org/10.1016/j.ecoser.2012.07.005

Kareiva, P. *et al.* (2011). *Natural Capital: Theory and Practice of Mapping Ecosystem Services*. (Oxford University Press). oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199588992.001.0001/acprof-9780199588992

Kindermann, G.E. *et al.* (2008). A global forest growing stock, biomass and carbon map based on FAO statistics. *Silva Fennica* **42**, 397–396. pure.iiasa.ac.at/id/eprint/8616/

Marmion, M. *et al.* (2009). Evaluation of consensus methods in predictive species distribution modelling. *Divers. Distrib.* **15**, 59–69. doi.org/10.1111/j.1472-4642.2008.00491.x

Martínez-López, J. *et al.* (2019). Towards globally customizable ecosystem service models. *Sci. Total Environ.* **650**, 2325–2336. doi.org/10.1016/j.scitotenv.2018.09.371

Masson, D. & Knutti, R. (2011). Climate model genealogy. *Geophys. Res. Lett.* **38**. L08703. doi.org/10.1029/2011GL046864

Mulligan M. (2013). WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally. *Hydrol. Res.* **44**, 748–69. doi.org/10.2166/nh.2012.217

Refsgaard, J.C. *et al.* (2014). A framework for testing the ability of models to project climate change and its impacts. *Clim. Change* **122**, 271–282. doi.org/10.1007/s10584-013-0990-2

Scholes, R.J. (1998). *The South African 1: 250 000 maps of areas of homogeneous grazing potential.* (CSIR, South Africa). No internet reference

Smith, B. *et al.* (2014). Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. *Biogeosciences* **11,** 2027–2054. d-nb.info/1121909426/34

Thomas, A. *et al.* (2020). Fragmentation and thresholds in hydrological flow-based ecosystem services. *Ecol. Appl.* **30**, e02046. doi.org/10.1002/eap.2046

Thomas, A. *et al.* (2020). Fragmentation and thresholds in hydrological flow-based ecosystem services. *Ecol. Appl*. **30**, e02046. doi.org/10.1002/eap.2046

Thuiller, W. *et al.* (2009). BIOMOD–a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369-373. doi.org/10.1111/j.1600-0587.2008.05742.x

Willcock, S. *et al.* (2016). Do ecosystem service maps and models meet stakeholders' needs? A preliminary survey across sub-Saharan Africa. *Ecosyst. Serv.* **18**, 110–117. doi.org/10.1016/j.ecoser.2016.02.038

Willcock, S. *et al.* (2019). A Continental-Scale Validation of Ecosystem Service Models. *Ecosystems* **22**, 1902–1917. doi.org/10.1007/s10021-019-00380-y

Willcock, S. *et al.* (2020).Ensembles of ecosystem service models can improve accuracy and indicate uncertainty. *Sci. Total Environ.* **747***,* 141006. doi.org/10.1016/j.scitotenv.2020.141006

.