

TITLE PAGE

Reducing Uncertainty in Ecosystem Service Modelling through Weighted Ensembles

Hooftman, Danny A.P.^{1,2}, James M. Bullock², Laurence Jones³, Felix Eigenbrod⁴,
José I. Barredo⁵, Matthew Forrest⁶, Georg Kindermann⁷, Amy Thomas³ & Simon
Willcock^{8,9*}

* Corresponding author

Affiliations:

1. Lactuca: Environmental Data Analyses and Modelling, The Netherlands.
danny.hooftman@lactuca.nl
2. UK Centre for Ecology and Hydrology, Wallingford, OX10 8BB, United Kingdom.
jmbul@ceh.ac.uk
3. UK Centre for Ecology and Hydrology, Bangor, LL57 2UW, United Kingdom. lj@ceh.ac.uk,
athomas@ceh.ac.uk
4. Geography and Environment, University of Southampton, United Kingdom.
F.Eigenbrod@soton.ac.uk
5. Joint Research Centre of the European Commission, Brussels, Belgium.
Jose.BARREDO@ec.europa.eu
6. Senckenberg Biodiversity and Climate Research Centre, Frankfurt, Germany.
matthew.forrest@senckenberg.de
7. International Institute for Applied Systems Analysis, Laxenburg, Austria. kinder@iiasa.ac.at
8. School of Natural Sciences, Bangor University, United Kingdom. s.willcock@bangor.ac.uk
9. Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, United Kingdom.

Contributions: DAPH, JMB & SW conceived the project. DAPH, LJ, AT, MF, JB & GK provided ES model descriptions and outputs. DAPH conducted all analyses. DAPH, JMB & SW wrote the manuscript, with comments from AT, FE, JB, LJ, MF & GK.

Acknowledgements: This work took place under the Ensembles project – Using ensemble techniques to capture the accuracy and sensitivity of ecosystem service models ([NE/T00391X/1](#)). Land Cover Map 2015 is under UKCEH licence 1403. We acknowledge the help of Kevin Watts for guiding us through the Forest Research data and John Redhead for providing InVEST biophysical tables. We also thank the anonymous reviewers for their insightful comments on the manuscript.

37 **ANONYMISED MANUSCRIPT**

38 **Reducing Uncertainty in Ecosystem Service Modelling**
39 **through Weighted Ensembles**

40
41 **Highlights:**

- 42 • Ensembles of models are used for other disciplines but not ecosystem services
- 43 • How best to combine ecosystem service models into an ensemble is unknown
- 44 • We test ten contrasting ensemble approaches
- 45 • Ensembles had up to 27% higher accuracy than a randomly selected individual model
- 46 • Weighted ensembles provided better predictions

47
48 **Abstract:** (150 words)

49 Over the last decade many ecosystem service (ES) models have been developed to inform sustainable land
50 and water use planning. However, uncertainty in the predictions of any single model in any specific situation
51 can undermine their utility for decision-making. One solution is creating ensemble predictions, which
52 potentially increase accuracy, but how best to create ES ensembles to reduce uncertainty is unknown and
53 untested. Using ten models for carbon storage and nine for water supply, we tested a series of ensemble
54 approaches against measured validation data in the UK. Ensembles had at minimum a 5-17% higher
55 accuracy than a randomly selected individual model and, in general, ensembles weighted for among model
56 consensus provided better predictions than unweighted ensembles. To support robust decision-making for
57 sustainable development and reducing uncertainty around these decisions, our analysis suggests various
58 ensemble methods should be applied depending on data quality, for example if validation data are available.

59
60 **Graphical Abstract:**

Accuracy compared to mean ensemble	
Individual Models	Mean accuracy is always worse

Unweighted Ensembles	
Mean Ensemble	Reference ensemble type: up to 19% better than individual models
Median Ensemble	Mostly better, rarely worse

Untrained Weighted Ensembles	
Deterministic Consensus	Sometimes better, sometimes worse
Iterated Consensus	Mostly better, rarely worse
Attribute based	Sometimes better, sometimes worse

Trained Weighted Ensembles	
Accuracy weighted	Mostly better, rarely worse
Regressed consensus	Mostly better, never worse

61
62
63 **Keywords:** Carbon; Committee averaging; Prediction Error; Accuracy; United Kingdom; Validation;
64 Water supply; Weighted averaging

65
66 **Video Summary:** (see attached file)

67 1. Introduction

68 If the United Nations' sustainable development goals (SDG) are to be achieved worldwide (Griggs *et al.*
69 2013), it is vital to understand and manage “*nature's contributions to people*” (termed ecosystem services;
70 ES; Pascual *et al.* 2017). The empirical data needed to quantify ES are sparse in many parts of the world
71 (Suich *et al.* 2015; Willcock *et al.* 2016), which is problematic as ES need to be accurately assessed and
72 mapped to be incorporated in policy making and planning decisions (UKNEA 2011; de Groot *et al.* 2012).
73 Such decisions require assessment of multiple ES, and the synergies and trade-offs among these ES, in order
74 to estimate potential effects of land/water use change or other impacts (Willcock *et al.* 2016). Spatially-
75 explicit models produce maps of estimated ES – typically based on globally available datasets of land cover
76 combined with other predictor variables – and so can provide credible information of the spatial distributions
77 of multiple ES, particularly where empirical data are lacking (Malinga *et al.* 2015; Costanza *et al.* 2017).

78
79 Over the last 10 years, many ES models have been developed, by different teams, often using dissimilar
80 approaches, and with little reference to the other models (Bagstad *et al.* 2013; Ochoa & Urbina-Cardona
81 2017). For example, carbon stocks for climate change mitigation can be modelled by ‘look-up tables’
82 relating land cover to stocks, by deterministic statistical inference, or by simulating complex processes
83 (Willcock *et al.* 2019). However, most applications of ES models rely on only a single model for each ES
84 (Englund *et al.* 2017; Bryant *et al.* 2018). Furthermore, while models can only approximate reality, few
85 applications explicitly validate ES models against independent datasets (Chaplin-Kramer *et al.* 2019),
86 although there are notable exceptions (Redhead *et al.* 2016; Sharps *et al.* 2017; Willcock *et al.* 2019). This
87 is a particular issue as the results of location-specific validation (*e.g.* that performed during model
88 development) may not be transferable to new locations (Redhead *et al.* 2016), or up-scalable to the regional
89 and national extents over which ES model outputs are required to achieve the SDG (Willcock *et al.* 2016;
90 Willcock *et al.* 2019). From a user and stakeholder perspective, not knowing the accuracy of the available
91 ES models for the region of interest typically leads to either selection of a single suboptimal model – at
92 worst leading to perverse decision-making – or a reluctance to use ES models altogether, causing an
93 implementation gap between research, incorporation into policy and subsequent decision-making (Wong *et al.*
94 *et al.* 2014; Willcock *et al.* 2016).

95
96 Despite claims for predictive superiority of certain modelling techniques and platforms, independent
97 evaluations have been unable to demonstrate the pre-eminence of any single approach. In fact, while more
98 complex models on average perform better in terms of fit to validation data, the best-fit model varies
99 regionally and often according to the validation data used (Sharps *et al.* 2017; Willcock *et al.* 2019; Willcock
100 *et al.* 2020). So, if no single ES model is always the most accurate, how should a suitable approach be
101 selected?

102
103 Across the sciences, one solution to address uncertainty surrounding the accuracy of any single model is to
104 use an ensemble of models (Araújo & New 2007; Willcock *et al.* 2020) – using individual models as
105 replicates with different input parameters and boundary conditions (Araújo & New 2007; Dormann *et al.*
106 2018). Variation among models in their assumptions and formats can result in large differences in
107 predictions, in terms of predicted values and how they vary over space, especially when there is uncertainty
108 as to the state and processes of the system being modelled (van Soesbergen & Mulligan 2018; Willcock *et al.*
109 *et al.* 2019). Ensembles of models are hypothesised to have enhanced accuracy over individual models due to
110 fewer overall errors in prediction by reducing the influence of idiosyncratic outcomes from single models
111 (Araújo & New 2007; Dormann *et al.* 2018). Individual models rarely capture all potentially relevant
112 processes or are often tuned to particular ecosystem characteristics. A combination of models might provide
113 a more comprehensive coverage of processes and their forms, and avoids the chance of (unknowingly)
114 selecting a model with a high prediction error at the location and scale of interest for a particular study
115 (Willcock *et al.* 2020).

116

117 Model ensembles are common in other disciplines – *e.g.* in niche modelling (Araújo & New 2007,
118 Grenouillet *et al.* 2011), agroecology (Refsgaard *et al.* 2014), hydrology and water resources management
119 (Wang *et al.* 2019; He *et al.* 2021), and climate and weather modelling (Knutti *et al.* 2013), as well as market
120 forecasting (He *et al.* 2012). However, ensembles have been largely neglected in ES studies (Bryant *et al.*
121 2018). The only current exception is the simplest ensemble approach (*i.e.* ‘committee averaging’ – taking
122 the unweighted mean of a group of individual models per location –) which was applied to ES models in
123 Sub-Saharan Africa, and gave higher accuracy in terms of fit to validation data (Willcock *et al.* 2020).
124 Approaches that use more information might yield even more accurate estimates. Thus, here we explore the
125 outstanding question of “what are the best ways to build ES model ensembles to realise the benefits such
126 ensembles can bring to sustainability science?”

127
128 Approaches to building model ensembles vary across disciplines, ranging from committee averaging
129 (Marmion *et al.* 2009; Grenouillet *et al.* 2011) to complex Bayesian algorithms (Tebaldi & Knutti 2007).
130 For example, species distribution models are generally deterministic statistical models; their fit to the data
131 is often assessed with an accuracy metric and so ensembles are generally created using weighted averaging
132 based on accuracy (Araújo & New 2007). By contrast, climate models are often treated as equal replicates
133 with identical weights when making an ensemble (Tebaldi & Knutti 2007; Grenouillet *et al.* 2011) – we
134 refer to such ensembles as ‘unweighted’. This difference may stem from the availability of suitable
135 validation data, as well as different traditions. For example in species distribution models, biodiversity data
136 are readily available and are used to train through cross-validation (Araújo & New 2007), whereas validation
137 data on future climates obviously do not exist – although cross-validation against historic climate data is
138 possible.

139
140 As well as varying considerably in their underlying method, ES models often differ in the forms of their
141 outputs, even when modelling the same ES (*e.g.* summed monetary value of the ES (de Groot *et al.* 2012)
142 *vs.* specific biophysical predictions). By contrast, climate models generally have very similar forms of
143 outputs. An important knowledge gap is therefore how to combine distinct ES model outputs as
144 complementary inputs to provide a reliable ensemble. Outputs from different ES models can have different
145 units and it is challenging to decide the relative weighting to place on each model. Models for a particular
146 ES often have different structures, may include different processes, or may represent the same processes in
147 different ways (Ochoa & Urbina-Cardona 2017). As a result, the different ES models will most likely not
148 have equal accuracy, and so prediction errors (*i.e.* bias) may not be normally distributed among models
149 (Dormann *et al.* 2018). If ES models had equal overall accuracies, unweighted averaging may provide a
150 smoothing effect, reducing the impact of idiosyncratic outputs (*e.g.* at specific locations) of any particular
151 model to reveal useful signals (Araújo & New 2007, Knutti *et al.* 2013; Diengdoh *et al.* 2020). In cases of
152 varying overall accuracy, appropriate weighting of outputs based on model accuracy – *i.e.* models having
153 unequal assigned weights – might re-adjust the distribution of prediction errors, and so improve the accuracy
154 of the resulting ensemble (Refsgaard 2014; Dormann *et al.* 2018; Liu *et al.* 2020).

155
156 However for ES, the lack of *a priori* validation data in many cases means that the distributions of accuracy
157 among ES models are unknown. Furthermore, given that inferences about model accuracy at one location
158 may not be transferable to others (Willcock *et al.* 2019), weighting using validation results from a separate
159 study may not improve outcomes. Therefore where validation data are not available, the consensus among
160 models could be used to weight their individual contribution to the ensemble value (Marmion *et al.* 2009;
161 Grenouillet *et al.* 2011). This approach follows the logic that models whose output values are more different
162 to those of the other models (*i.e.* are more distinct) are more likely to be incorrect. Therefore, weighting by
163 consensus reduces the impact of outputs from more idiosyncratic models (*i.e.* those with extreme values,
164 outliers or badly comparable processes) by comparison with the other models (Araújo & New 2007;
165 Dormann *et al.* 2018), but does not exclude their information fully. The opposite may also be true – *i.e.*
166 more distinct models are more accurate – for example in cases where more similar models have common
167 inaccuracies.

168

169 Here, we implement 10 alternative ensemble methods, restricting ourselves to methods feasible for a wide
 170 range of users, to evaluate whether weighting provides higher accuracy and if so which type of method
 171 produces the most accurate predictions against validation data. We focus on two services, water supply and
 172 carbon storage, in the United Kingdom. To support decision-making, we map the results for potential further
 173 use, which are available via <https://doi.org/10.5285/a9ae773d-b742-4d42-ae42-2b594bae5d38>. We use
 174 post-processing – specifically normalisation and per area correction – developed in earlier work (Willcock
 175 *et al.* 2019; Willcock *et al.* 2020) to make outputs among models comparable.

177 2. Methods

178 We developed and validated unweighted average and weighted average ensembles of models for a
 179 provisioning service (water supply; subsequently referred to as ‘water’) and a regulating service
 180 (aboveground carbon storage; subsequently referred to as ‘carbon’), for which there is both a variety of
 181 models available (Bagstad *et al.* 2013; Ochoa & Urbina-Cardona 2017; Willcock *et al.* 2019) and the
 182 presence of accessible validation data. We applied the models and ensemble methods in the United Kingdom
 183 (UK), for which there is a large quantity of reliable validation data; allowing us to assess ensemble
 184 accuracies. We compared accuracy (*i.e.* fit to validation data) of these individual models with those of the
 185 ensembles generated from them via multiple approaches, assessed if weighted ensembles were an
 186 improvement on the unweighted mean-averaged ensemble, and identified the methods of weighting
 187 ensembles that gave the highest accuracy.

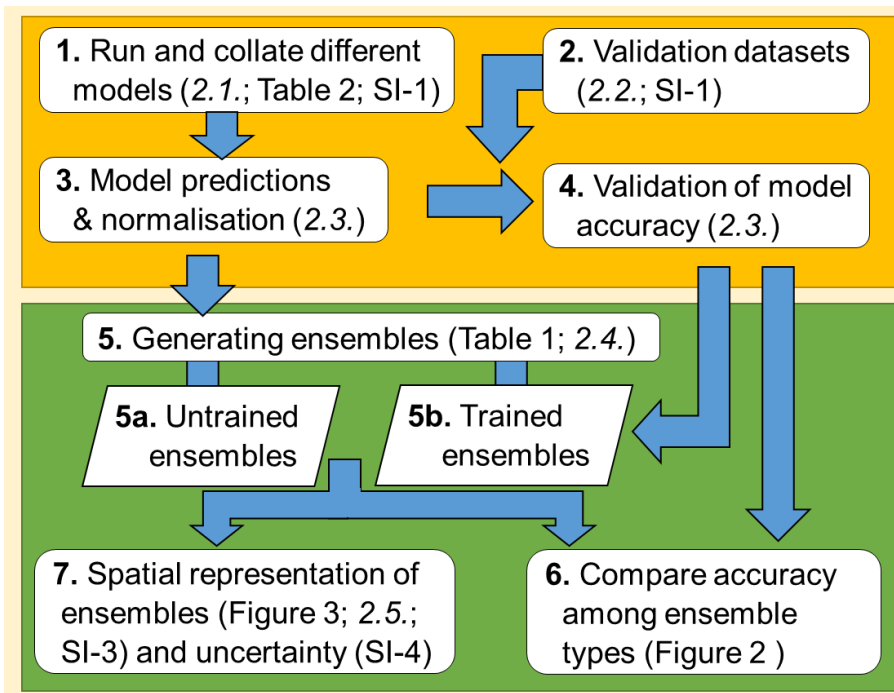
188
 189 We modelled each ES at a 1 ha (100 × 100 m) resolution, and subsequently assessed performance of the
 190 different ensemble approaches using weighting approaches we organised into three categories (Table 1):
 191 deterministic consensus (*i.e.* always providing the same result), iterated consensus (*i.e.* using structured
 192 trial-and-error approaches) and attribute-based (*e.g.* spatial resolution or distinctiveness). Finally, we
 193 assessed the transferability of our UK results using independent data and models from a very different study
 194 area – Sub-Saharan Africa (Willcock *et al.* 2019). We depict our overall process in Figure 1 in 7-steps. Our
 195 calculations were performed using Matlab v7.14.0.739 and ArcMap 10.7.1, employing ArcPy coding for
 196 loops. Relevant codes can be found at github.com/EnsemblesTypes, with flow among codes explained in
 197 SI-1-3.

199 **Table 1. Approaches used to calculate accuracy (A) and ensembles (B).** Ensemble approaches were
 200 applied to the outputs of ten models for carbon storage and nine for water supply (see Table 2). For weighted
 201 averaging, the procedure is described, and where applicable the Matlab tools used are mentioned; similar
 202 regression tools are available in most statistical packages (further explanation is provided in SI-1). Trained
 203 weighting (En-9 & En-10) uses validation data, whereas untrained weighting (En-3 to En-8) does not. En-1
 204 and En-2 are unweighted average ensemble approaches, and En-3 to En-10 are weighted average
 205 approaches; the latter comprising *deterministic* (En-3 & En-4), *iterated* (En-5, En-6 & En-10) and *attribute*
 206 *weighted* (En-7 to En-9) techniques. With ω_i : weight for model i ; $E_{(x)}$: the value of the ensemble; $V_{(x)}$: the
 207 normalised validation value; $Y_{i(x)}$ and $Y_{j(x)}$: the normalised value of model i or comparator j respectively, all
 208 for selected spatial point x ; ($y \neq x$) denoting a split dataset; $C_{(i,j)}$: the correlation coefficient between model
 209 i and j ; with n the # models, m the # spatial data points; n^g : the # models in distinctiveness group g (see SI-
 210 1 for distinctiveness grouping).

Approach	Description	Details & Matlab Tool
A. Accuracy approaches		
• Spearman ρ	Correlation coefficient between ranked variables V and T .	T is either Y_i or E , depending on ensemble method
• Inverse Deviance (D^\downarrow)	$D^\downarrow = 1 - \left(\frac{1}{m} \times \sum_x X_{(x)} - T_{(x)} \right)$	$T_{(x)}$ is either $Y_{i(x)}$ or $\underline{E}_{(x)}$
B. Ensemble approaches		
Unweighted Averaging:		
En-1. Mean	$E_{(x)} = (\bar{Y}_i)_{(x)}$	

En-2. Median		$E_{(x)} = (\bar{Y}_i)_{(x)}$	Hypothesised to perform better than mean for skewed distributions.
Untrained Weighted Ensembles: $E_{(x)} = \sum_i^n \left(\frac{\omega_i}{\sum_i^n \omega_i} \times Y_i \right)_{(x)}$ with ω_i following:			
Deterministic consensus	En-3. PCA	$\omega_i = \text{loadings of first Principal Component axis}$	Princomp-tool
	En-4. Correlation coefficients	$\omega_i = \frac{1}{n} \times \sum_j^n \frac{C_{(i,j)}}{\sqrt{C_{(i,i)} \times C_{(j,j)}}}$, for all $j \in i$ with $C_{(i,j)} = \frac{1}{m-1} \times \sum_x^m \left((Y_{i(x)} - \bar{Y}_i) \times (Y_{j(x)} - \bar{Y}_j) \right)$	
Iterated consensus	En-5. Regression to the median	$\bar{Y}_{(x)} \sim (\sum_i^n \omega_i Y_i)_{(x)}$	nlmefit-tool, maximising Log Likelihood
	En-6. Exhaustive leave-one-out cross-validation ²	$Y_{j(x)} \sim \sum_{i \neq j}^n \omega_i Y_{i(x)}$, for all $j \in i$ subsequently: $\omega_i = \frac{1}{n} \times \sum_i^n \left(\left(\frac{1}{n-1} \right) \times \sum_{i \neq j}^n \omega_{ij} \right)$	nlmefit-tool, maximising Log Likelihood
Attribute-based	En-7. Upweighted finer spatial resolution	$\omega_i = \frac{1}{\log_{10}(\text{spatial resolution})}$	Finer spatial resolution: smaller grid size in 1-dimensional meters (e.g. 25 m)
	En-8. Attribute weighting: distinctiveness	$\omega_i = \left(\frac{n^g}{n} \right)$ when upweighted with $n^g = i \in g$ $\omega_i = \left(\frac{n}{n^g} \right)$ when downweighted with $n^g = i \in g$	
Trained Weighted Ensembles: ω-transfer via jack-knife training			
Attribute-based	En-9. Accuracy-weighted	$\omega_i = A_i$, with $A_i(V_{(y \neq x)}, Y_{(y \neq x)})$	With A , either Spearman ρ or D^{\downarrow} accuracy
Iterated consensus	En-10. Log-likelihood regressions	$V_{(y \neq x)} \sim (\sum_i^n \omega_i Y_i)_{(y \neq x)}$	Using nlmefit-tool, maximising Log Likelihood

212



213

214

215

216

217

218

219

220

Figure 1. Schematic representation of our ensemble analysis with arrows showing information flows. Numbers represent the steps with the method chapters indicated in italics, with respective detailing SIs; result figures are indicated. Parallelograms highlight the 10 ensembles approaches (Table 1), using models described in Table 2.

2.1. Run and collate different models (step 1)

221 We used outputs from 10 models for above ground carbon stocks based on per grid cell estimates, and
222 outputs from nine models for annual water supply which provided accumulated flow estimates through
223 specific pour points, either directly or through summation of run-off estimates per grid cell. We list these
224 models in Table 2, including their output grid sizes (spatial resolution); we refer to SI-1-1 for full details,
225 scales and supporting data. Acknowledging that model outputs have different units and sometimes model
226 different constructs, we refer further to them in the general terms of carbon and water supply. Adhering to
227 the aim of this paper, we do not compare individual model outputs, but focus on ensemble methods. All
228 model outputs were set to the British National Grid transverse Mercator projection (EPSG 27700) with a
229 0.9996 scale factor and units in metres. Not all models covered the whole of the UK, *e.g.* some excluded
230 Northern Ireland or Scotland (see SI-1-1). Where applicable we corrected for this by using a standard error
231 of means as $(\frac{\sigma(x)}{\sqrt{n(x)}})$, instead of standard deviation (σ), with n the number of models per grid cell x . We
232 collated models for this study according to their availability and to reflect different approaches to modelling
233 ES.

234 **Table 2. Models and existing outputs used.** Full details, input data, post processing descriptions, and coverage are provided in SI-1-1. Model names are
 235 shown as acronyms and in full.
 236

Model	Description	Grid size (spatial resolution)	Model Type ¹⁶
InVest v3.7.0 ^{1†} (Integrated Valuation of Ecosystem Services and Trade-offs)	Carbon module: above ground stocks	25 × 25 meters	Look-up table
	Water yield module: run-off per cell		Process
LPJ-GUESS ^{2,3†} (Lund-Potsdam-Jena General Ecosystem Simulator)	Vegetation biomass stocks per cell, mean for years 2009-2018	0.5° (≈ 46 × 46 km)	Process
	Water run-off per cell, mean for years 2009-2018		
LUCI ^{4†} (Land Utilisation Capability Indicator)	Above ground carbon stocks	10 × 10 meters	Look-up table
	Accumulated water run-off	5 × 5 meters	Process
\$-benefit transfer using The Economics of Ecosystems and Biodiversity database ^{5,6†}	Above ground carbon stock as monetary value	25 × 25 meters	Look-up table
	Water run-off as monetary value per cell		
Aqueduct v2.1 Total Blue Water ^{7§}	Accumulated water run-off	138 flow areas	Deterministic
ARIES k-Explorer ^{8‡} (Artificial Intelligence for Environment & Sustainability)	Joined above and below ground carbon stocks	1-hectare	Look-up table
Barredo <i>et al.</i> (2012) [§]	A European map of above ground biomass stocks	1 km ²	Look-up table
Copernicus, Tree Cover Density ^{9§}	Proxy for carbon: tree Cover Density 2015 from MODIS satellite imagery.	20 × 20 meters	Deterministic
DECIPHeR ^{10§} (Dynamic fluxEs and ConnectIvity for Predictions of HydRology)	Accumulated water run-off through NRFA delineated catchment outlets, mean for years 1995-2015	387 catchments in common with validation	Process
Grid-to-Grid ^{11§}	Accumulated water run-off, mean for years 1995-2015	1 km ²	Process
Henrys <i>et al.</i> (2016) [§]	Above ground carbon stocks	1 km ²	Look-up table
Kindermann <i>et al.</i> (2008) [§]	A global map of above ground forest biomass stocks	1 hectare	Deterministic
National Forest Inventory (2018) ^{12†}	Woodland Land Cover Map ¹⁵ with above ground carbon stocks based on added Look-up table (Table. SI-1-4)	20 × 20 meters	Look-up table
Scholes Growth Days ^{13,14†}	Proxy for water run off per cell: # Days precipitation exceeds evapotranspiration	1 km ²	Deterministic
WaterWorld v2 ^{15‡}	Accumulated water run-off	0.0083° (≈ 1 km ²)	Process

237 [†]Output generated for this work; [‡]online tool; [§]existing dataset; ¹Kareiva *et al.* (2011); ²Smith *et al.* (2014); ³Ahlström *et al.* (2015); ⁴Thomas *et al.* (2020); ⁵de Groot *et al.*
 238 (2012); ⁶Costanza *et al.* (2014); ⁷Gassert *et al.* (2015) ⁸Martínez-López *et al.* (2019); ⁹land.copernicus.eu/tree-cover-density/status-maps/2015; ¹⁰Coxon *et al.* (2019a; 2019b);

239 ¹¹Bell *et al.* (2018a; 2018b); ¹²Forestry Commission (2018); ¹³Scholes (1998); ¹⁴Willcock *et al.* (2019); ¹⁵Mulligan (2013); ¹⁶following Ding & Bullock (2018), Willcock *et al.*
240 (2019).
241

2.2. Validation datasets (step 2)

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/; density shown in Figure 3; locations in Figure SI-1-2). In 201,143 forest compartments of varying size (mean: 4.4 hectares, median 1.6 hectares, ± 22.1), tree species, stand age and thinning regime were recorded for three vegetation layers. For each compartment and layer therein, the unique combination of stand age, thinning regime and tree species of the inventory data was searched in the UK Carbon Code tables (woodlandcarboncod.org.uk) and life-time accumulated biomass was converted to total standing carbon per hectare estimates per compartment, with the layers summed per compartment (SI-1-2). Subsequently, compartments were spatially joined into 2078 polygons of ‘forest’ that were separated if more than 25 meters distance from each other.

Our water supply validation dataset comprised 519 hydrometric gauging stations from the National River Flow Archive of the UK (NRFA; nrfa.ceh.ac.uk), with associated catchments representing a variety of sizes distributed across the whole of the UK (Figure 3). From the 1598 potential catchments in NRFA, we selected those that were $>100 \text{ km}^2$ 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 $>25 \text{ km}^2$ to avoid this part of the UK being underrepresented. The data were polygons encompassing these catchments. Details are provided in SI-1-2.

2.3. Model predictions, normalisation (step 3) and validation of model accuracy (step 4)

For each individual model, predictions were obtained for each polygon in the validation dataset using the ArcGIS spatial analyst Zonal tool with a forced 2.5 m grid size environmental setting to minimise edge effects; *i.e.* all predicted values were obtained by resampling into $2.5 \times 2.5 \text{ m}$ grid cells. In most cases the modelled value per polygon was obtained by taking the sum of all constituent grid cell values, corrected for both actual grid size and the resampling to 2.5 m. 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 (with accumulated run-off estimates post-processed to give net run-off per cell; SI-1-1). 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 such data-points, we employed a double-sided Winsorising protocol for normalisation (Willcock *et al.* 2019; Verhagen *et al.* 2017), 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). This winsorising normalisation protocol assumes outlier data are valid, but skewed values, in our case mainly by per area averaging, and corrects for this by compressing the variance tails rather than trimming them (Keselman *et al.* 2008; Erceg & Miroseovich 2008). Hence, we trade-off an even data distribution over the full 0-1 normalised range against the chance of having a true far outlier maximum (see SI-5 for a full investigation into the impact of the Winsorising protocol over standard normalisation for the validation data distribution). For each model, normalisation was done prior to creating ensembles.

For validation, we employed two accuracy measures (Willcock *et al.* 2019; Willcock *et al.* 2020), which are related to different aims in modelling ES (Table 1):

- 293 1) Comparing the rank order of predicted and validation data using Spearman ρ . This is relevant where
 294 modelling is used to discover, for example, the most important locations for delivering an ES, or
 295 conversely, those areas whose development may have least impact on ES delivery.
 296 2) Ascertaining the absolute difference of each modelled value from its validation value using the inverse
 297 of the deviance (D^\downarrow). This is relevant where modelled values are important, *e.g.* when testing where ES
 298 levels exceed a minimum threshold. We used the inverse of the deviance so that, like ρ , a higher value
 299 indicated greater accuracy.

300
 301 *2.4. Generate ensembles (step 5) and compare accuracy among ensemble types (step 6)*
 302 We tested whether model ensembles were more accurate than the individual constituent models and which
 303 approaches for creating ensembles were the most accurate in terms of fit to validation data. We created
 304 ensembles using a range of methods, from the simplest calculation of an average value of the models at each
 305 location ('unweighted averaged ensembles', *e.g.* Marmion *et al.* 2009, Grenouillet *et al.* 2011) to ensembles
 306 with the contributions from different models weighted unequally ('weighted ensembles'), following
 307 Dormann *et al.* (2018) (Table 1; further explanation and a model flow are provided in SI-1-3). We used
 308 relatively straightforward approaches that would be feasible for a wide community of scientists and
 309 decision-makers, and avoided more complex mathematical and/or statistical techniques such as Bayesian
 310 networks (Bryant *et al.* 2018), which would require detailed specialist knowledge. Weights over all models
 311 were normalised to sum to 1. Together with normalisation of the ensemble outputs (see above), this assured
 312 equal scaling among all models and ensembles.

313
 314 For unweighted average ensembles, we calculated both the mean and the median of modelled values at each
 315 location as alternative measures of the central tendency which are differently affected by skew in the data
 316 (Table 1, En-1 & En-2).

317
 318 For weighted ensembles we calculated:
 319
$$E_{(x)} = \sum_i^n \left(\frac{\omega_i}{\sum_i^n \omega_i} \times Y_i \right)_{(x)}$$
 with positive weights ω_i for model i of validation polygon x , weights ω_i are
 320 normalised to sum to 1, Y the modelled values for i per polygon (step 3), and n the total number
 321 of models per service.
 322

323 To determine ω_i , the weighting value for each model i , we employed a range of methods that can be broadly
 324 categorised as two main types of ensemble approach (untrained and trained), with further subdivision as:
 325 deterministic consensus, iterated consensus, and attribute-based. The ensembles are listed as equations in
 326 Table 1 (see SI-1-3 for further details).

- 327 1) Untrained ensembles (En-3 to En-8) represent a situation in which there is no validation data. To generate
 328 uncertainty estimates allowing statistical comparison with the models and among ensembles we jack-
 329 knifed (Araújo & New 2007; Refsgaard *et al.* 2014) with 50% of the spatial data polygons for 250 runs,
 330 *i.e.* every run contained a new selection of half the dataset. We tested three approaches to produce the
 331 ensembles:
 332 - *Deterministic consensus* among models can be calculated using several approaches, including the fit
 333 to a common consensus axis such as from a Principal Components Analysis (Marmion *et al.* 2009;
 334 Grenouillet *et al.* 2011) or weighting by correlation coefficients (En-3 & En-4; ensemble numbering
 335 follows Table 1).
 336 - *Iterative approaches* might more accurately quantify consensus among models through using
 337 structured trial-and-error (Dormann *et al.* 2018; Tebaldi & Knutti 2007). We use two regression
 338 techniques: between the individual models and the median (En-5) and leave-one-out cross-validation
 339 (En-6) following the suggestion in Dormann *et al.* (2018).
 340 - One might *a priori* place value on a particular model attribute and use this to create weights (Englund
 341 *et al.* 2017; Willcock *et al.* 2019; Brun *et al.* 2020; En-7, En-8 & En-9). For example, one could up-
 342 or down-weight more distinct model types through a binary matrix of differences (En-8 & En-9; SI-

343 1-4) in land cover map used, grid-size, measured or modelled climate, model extent, presence of
344 time-series, time step-size and model type (*i.e.* look-up table, deterministic or process based).
345 Alternatively models that run at coarser spatial resolutions are penalised (En-7): smaller grid sizes
346 are deemed more useful for decision-making (Willcock *et al.* 2016).

347 2) Trained ensembles (En-9 & En-10), as often used for species distribution models (*e.g.* Refsgaard *et al.*
348 2014; Elith *et al.* 2011), represent a situation in which validation data are available from a similar region
349 or part of the study area and so cannot be used to directly validate or substitute for the models in the
350 study area, but can be used to weight these models. Here, ω_i was trained with the validation data on a
351 jack-knifed 50% of the dataset to achieve maximum accuracy (En-10) and subsequently ω_i was
352 transferred to the other half of the dataset. We used 250 such jack-knife runs (see above), with the same
353 selections as above. Moreover, we included weighting by individual model accuracy (Marmion *et al.*
354 2009; Liu *et al.* 2020) using the same jack-knife approach (En-9).

355
356 After creating the ensembles, their accuracy was assessed following step 4 using the two measures (see 2.3):
357 Spearman ρ and the inverse of the deviance (D^\dagger). We assessed any improvement over the unweighted mean-
358 averaged ensemble as the reference with pairwise t-tests against the null hypothesis of equal accuracy
359 (Matlab *ttest*-tool). A similar analysis against the median-averaged ensemble as reference can be found in
360 SI-2. To avoid spurious findings of significance through having a large number of replicates, we assessed
361 improvement using bootstrapped tranches of 50 runs each with 250 replicates, and averaging the P-values.
362 Since we used the same statistical test 12-times per service per accuracy estimate, we employed a full
363 conservative Bonferroni correction; ($\alpha = 0.05/12$) on the resulting average P-values. To compare the
364 ensembles with the individual models we calculated per replicate the mean difference in accuracy among

365 all models (A_i) against accuracy of an ensemble (A_E) following: $\left(\left(\sum_i^n \left(\frac{A_E}{A_i} - 1 \right) \right) \times \frac{1}{n} \right)$, with n the number
366 models and i an individual model.

367
368 Steps 5 and 6 were repeated using independent data and models from a different study area (sub-Saharan
369 Africa; Willcock *et al.* 2019) to investigate the transferability of the results presented here (Figure SI-2-2).

371 2.5. Spatial representation of ensembles and uncertainty (step 7)

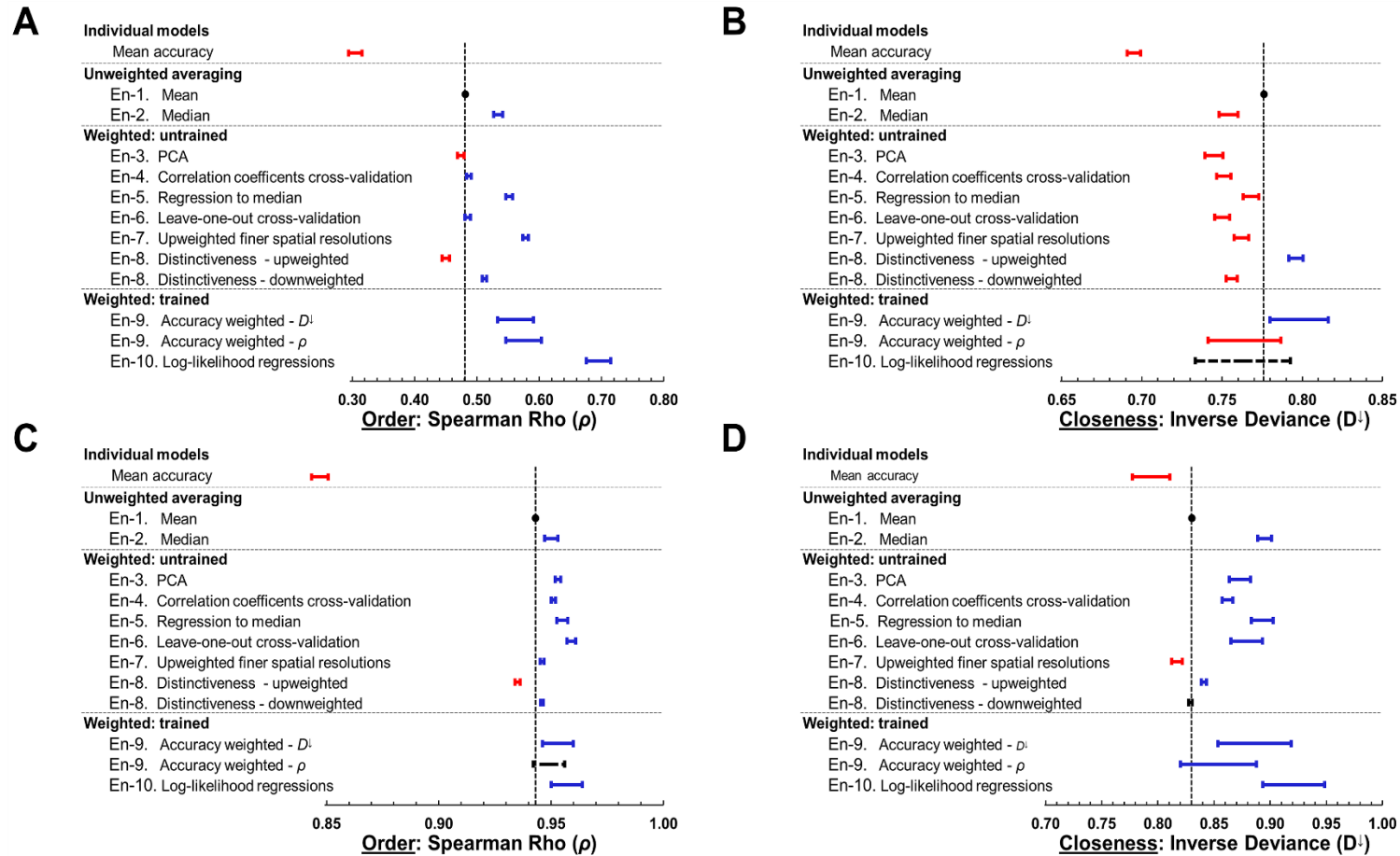
372 To better support decision-making, we mapped our ES ensembles for the UK. For all the water ensembles,
373 the mean normalised value across jack-knifed ensemble predictions per ensemble method were mapped as
374 catchment polygons (step 5, $N = 519$). For all carbon ensembles we mapped as 1 km² grid cells. Here, for
375 each ensemble approach, the estimated weights as calculated for the validation polygons – mean averaged
376 among jack-knife runs – were transferred to the full area, with the result aggregated to a 1 km² resolution
377 based on the mean value among 1 hectare grid cells. In total, this carbon dataset has 253,802 cells that
378 (partially) contain non-sea land cover. We transferred the weights calculated for the forests since running
379 cross-validation approaches on over 250K data points would extremely time consuming to compute.
380 However, since our validation data are only from forests/woodlands, we are aware of introducing a potential
381 bias that could skew non-forested areas to lower values. Furthermore, we generated UK-scale maps of
382 spatial variation in the differences among the untrained ensemble approaches, by calculating the standard
383 error of the mean (SEM) among these spatial outputs. These maps are freely available online
384 (<https://doi.org/10.5285/a9ae773d-b742-4d42-ae42-2b594bae5d38>), and spatial patterns of uncertainty are
385 discussed in SI-4.

387 3. Results

388 3.1. Ensembles are more accurate than individual models

389 The average accuracy of individual models, represented by the mean of accuracy values taken across all
390 models, was lower than that for any of the ensembles we created. The accuracy of the unweighted averaged
391 ensembles (of modelled values at each location, *e.g.* ‘mean ensemble’) was appreciably higher than the
392

393 mean value for accuracy of the individual models for both carbon and water: 19% \pm 1.1% [sd] for ρ and
394 12.1% \pm 0.5% for D^\downarrow improvement in fit to the validation data for carbon and 5.7% \pm 0.4% for ρ and 9.5%
395 \pm 1.7% for D^\downarrow for water (Figure 2). Untrained weighted ensembles showed large improvements – for most,
396 larger than the unweighted ensembles – over the mean accuracy of the individual models of 17% to 27%
397 (ρ) and 7.6% to 15% (D^\downarrow) for carbon (Figure 2A and B), and 5.3% to 6.5% (ρ) and 7.7% to 18% (D^\downarrow) for
398 water (Figure 2C and D). In all cases, pairwise t-tests indicated highly significant differences between each
399 ensemble and the mean value of accuracy of individual models (all $P < 1E^{-10}$). Thus, creating an ensemble
400 improves prediction accuracy against a randomly chosen individual model irrespective of the ensemble
401 approach chosen.



402
 403
 404
 405
 406
 407
 408
 409
 410

Figure 2. Accuracy of above ground carbon stock ensembles (10 models; A and B), and of water supply ensembles (9 models; C and D) against validation data. The mean of accuracy values across the containing models – *i.e.* a randomly chosen model – is provided for comparison. For detail on the different ensemble types see Table 1 and SI-1-3. We show the average accuracy of 250 bootstrap runs with 50% of the dataset. The vertical dashed line indicates the reference unweighted mean-averaged ensemble (black dot, ‘mean ensemble’). Error bars indicate the standard deviation among runs in terms of proportional difference to the mean ensemble, calculated per bootstrap run as the difference in accuracy to the mean ensemble divided by the accuracy of the mean ensemble. The coefficient of variation among bootstraps for the mean carbon ensemble was 4% and 1%, for ρ and D^{\downarrow} respectively, and 1% and 2% for water (not shown). **Blue** coloured ensemble accuracies are significantly higher than the unweighted mean ensemble (Bonferroni corrected $\alpha = (0.05/12)$); **Red** coloured bars are significantly lower; **Black** dashed bars are not significantly different to the mean ensemble.

411 3.2. *Weighted ensembles are more accurate than unweighted ensembles*

412 All weighted ensembles, whether trained or untrained, significantly outperformed the reference unweighted
413 mean ensemble (Figure 2), with the exception of D^\downarrow for carbon. In all cases, pairwise t-tests indicated these
414 differences were highly significant ($P < 1E^{-10}$; see Figure SI-2-1 for similar analyses against the median-
415 averaged ensemble).

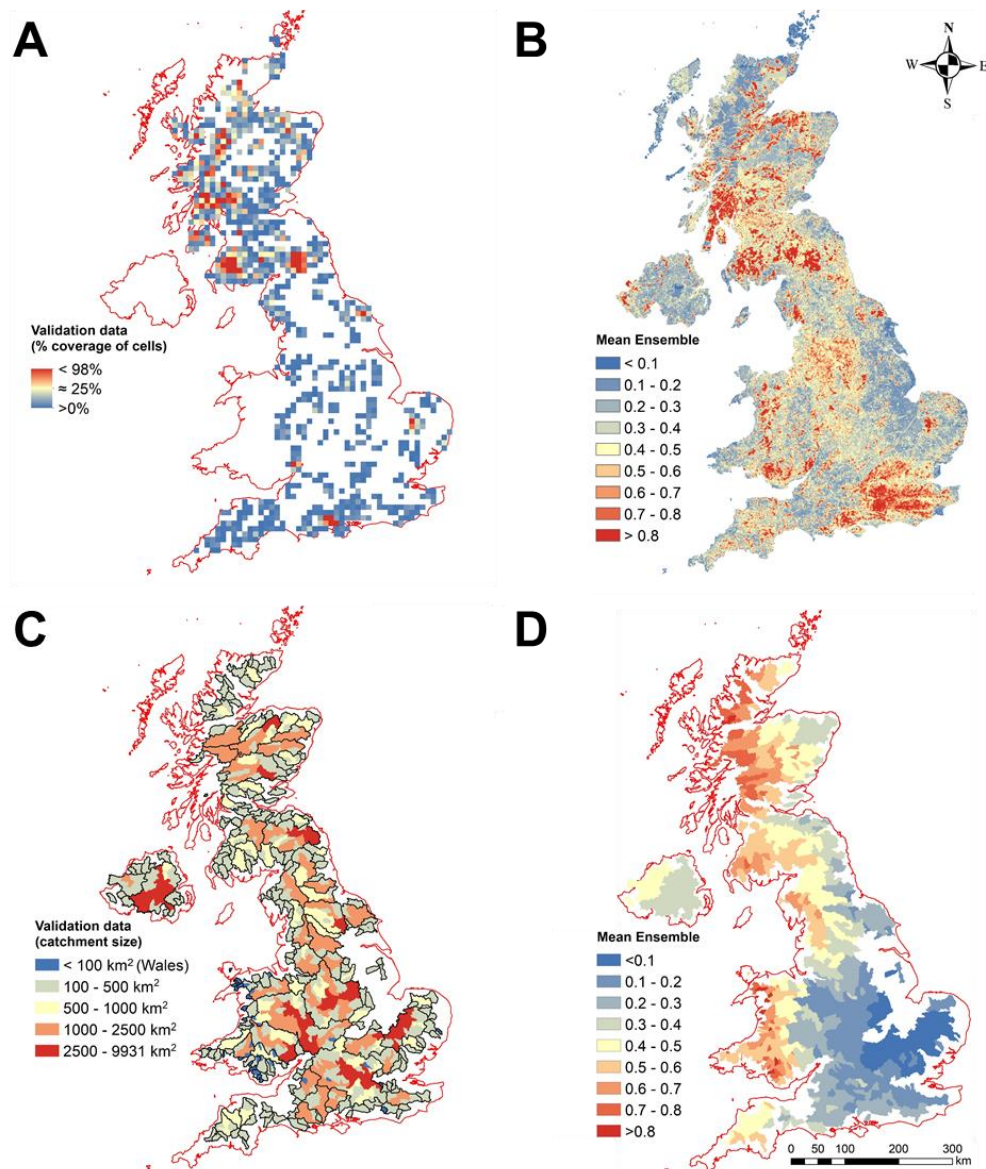
416
417 For untrained weighted ensembles, prediction accuracy was elevated by up to $4.8\% \pm 0.6\%$ for carbon ρ
418 (best: regression to median; Figure 2), with no improvement for carbon D^\downarrow , and $0.8\% \pm 0.3\%$ and 7.5%
419 $\pm 1.1\%$ for water supply ρ and D^\downarrow respectively (regression to median; Figure 2). Conclusions as to the best
420 model attributes to use for untrained weighting were dependent on the accuracy metric used (ρ or D^\downarrow). By
421 comparison to the unweighted mean ensembles, upweighting model outputs with finer spatial resolution
422 improved ρ by up to $6.6\% \pm 0.5\%$ and $0.2\% \pm 0.1\%$ for carbon and water respectively but contrastingly
423 decreased D^\downarrow . Upweighting more distinctive models was positive for D^\downarrow with $2.5\% \pm 0.4\%$ and $1.3\% \pm 0.3\%$
424 greater accuracy compared to the unweighted mean ensemble for carbon and water supply respectively, but
425 was negative for ρ . In summary, creating untrained weighted ensembles through iterative approaches was
426 overall the most robust – particularly regression to the median (Table 1: En-5), showing greater accuracy
427 than the unweighted mean-averaged ensembles in 3 out of 4 of our tests, and lower accuracy in 1 (Figure
428 2).

429
430 For trained weighting ensembles, using an iterative log-likelihood regression approach (Table 1: En-10) to
431 establish weights elevated prediction accuracy compared to the unweighted mean ensemble by up to 14.5%
432 $\pm 2.6\%$ for carbon ρ (no improvement for carbon D^\downarrow) and $0.8\% \pm 0.7\%$ and $11.1\% \pm 3.4\%$ for water supply ρ
433 and D^\downarrow respectively (Figure 2). Compared to such regressions, upweighting models with higher accuracy in
434 the training set (accuracy-weighted ensembles; En-9; Figure 2) gave less improvement over the unweighted
435 mean ensemble. Iteratively creating trained weighted ensembles using a log-likelihood regression approach
436 (Table 1: En-10) was most robust – showing greater accuracy than the unweighted mean-averaged
437 ensembles in 3 out of 4 of our tests, and is no worse in 1 (Figure 2).

438
439 The reference unweighted mean ensembles for carbon and water are mapped for the UK in Figure 3. Maps
440 for all other ensembles can be found in SI-3 and uncertainty among models and ensembles in SI-4. In
441 accordance with *a priori* predictions, the uncertainty associated with selecting a single model was several
442 times greater than that associated with selecting any single ensemble method for both ES. For carbon, the
443 standard error of the means (SEM) among individual models per 1 km² grid cell (SEM = $9.0\% \pm 2.8\%$, SI-
444 4) was ca. 3.5-times larger than among ensembles (SEM = $2.5\% \pm 1.1\%$). Similarly, the SEM among
445 individual water models per watershed (SEM = $7.8\% \pm 3.4\%$, SI-4) was substantially greater than among
446 ensembles (SEM = $1.3\% \pm 0.7\%$). In SI-4 we investigate spatial drivers for this uncertainty, discussing these
447 patterns at length.

448
449 We validated the robustness of our results using independent data and models from a different area (Sub-
450 Saharan Africa; Willcock *et al.* 2019), which gave similar results of weighted ensembles outperforming the
451 reference mean ensemble (Figure SI-2-2).

452



453
 454 **Figure 3. Spatial distribution of validation points and the reference mean ecosystem service value. A**
 455 **the Distribution of 2078 carbon validation forests as coverage of 10 × 10 km cells – many individual forest**
 456 **fragments would be too small to be clear at this scale, see SI SI-1-2 –, white cells are empty. B** the reference
 457 **unweighted mean ensemble of carbon across 10 models, normalised on scale 0-1. C** the 519 catchments
 458 **used for water validation and ensemble calculations coloured by their size – smaller watersheds that overlap**
 459 **larger ones are displayed on top; lines show underlying largest catchment level. D** the reference unweighted
 460 **mean ensemble of water supply across 9 models, normalised on scale 0-1. All maps here, in SI-3 (all**
 461 **ensembles) and SI-4 (uncertainty) could support landscape decisions in the UK and are available via**
 462 <https://doi.org/10.5285/a9ae773d-b742-4d42-ae42-2b594bae5d38>.
 463

464 4. Discussion

465 We have shown that predictions from ensembles of models have substantially higher accuracy than a
 466 randomly selected single ES model, and especially that weighting approaches increase ensemble accuracy.
 467 Finding increased performance through use of ensemble approaches is common in other fields. For example,
 468 the increased accuracy of ensemble species distribution models ranges from 1-2% (Crossman *et al.* 2012;
 469 Abrahms *et al.* 2019) to 12% (Grenouillet *et al.* 2011), although an increase is not universal (Hao *et al.*
 470 2020). Similarly, 2% accuracy increases were found for market forecasting ensembles (He *et al.* 2012), and
 471 neural network ensemble averaging resulted in up to 7% improvements in accuracy (Inoue & Narisha 2000).

472
473 Specific to ES, unweighted averaged ensembles have been shown to be 5.0–6.1% more accurate than
474 individual models (Willcock *et al.* 2020). Our improvements with ES ensembles are at minimum 5%-17%,
475 suggesting substantial differences among models in their adequacy (Dormann *et al.* 2018), but also that
476 ensemble approaches that use more information offer greater increases in accuracy. We found that taking
477 the median generally outperforms a mean ensemble, probably because the latter is more influenced by
478 outliers. Our results provide evidence that weighted ES ensembles created using consensus techniques
479 produce more accurate outputs than unweighted ensembles. This finding is supported by our additional
480 analysis using independent models and data from Sub-Saharan Africa (in a biome with very different
481 climatic and soil characteristics; SI-2), suggesting our findings may be generalisable, although investigating
482 this specifically (e.g., for different ES, regions and validation datasets) is an important avenue for future
483 research.

484
485 Predictions from models, including those from ES models, are all potentially biased in direction and amount
486 because of their underlying assumptions. These biases could differ among models due to their specific
487 construction. Therefore, models are likely to differ in their accuracy when compared to reality (Dormann *et al.*
488 *et al.* 2018). The improvement in accuracy when using ensembles, as we have shown here, is referred to as a
489 ‘portfolio effect’ by which a (weighted) combination of replications of possible states of a system suppresses
490 idiosyncratic differences and provides a more reliable average estimate (Thibaut & Connolly 2013;
491 Dormann *et al.* 2018; Lewis *et al.* 2021). However, this effect is lessened if models share similar
492 assumptions and, therefore, concomitant biases – highlighting the importance of including multiple model
493 outputs (Ding & Bullock 2018) and, where data are available, model validation (Willcock *et al.* 2019). In
494 particular, the use of models not usually packaged as ES models – such as LPJ-GUESS – might help with
495 increasing the variety of inputs for ensembles. If some models systematically overestimate and other models
496 underestimate, averaging delivers smaller prediction errors when models are weighted (Dormann *et al.*
497 2018). Hence, the resulting weighted ensemble is more accurate than most individual models and
498 unweighted approaches (Marmion *et al.* 2009, Grenouillet *et al.* 2011); see Dormann *et al.* (2018) for
499 theoretical explorations.

500
501 We have shown the general potential of weighting to re-balance the contribution of different ES models,
502 but also find that some weighting approaches seem more suitable. Specifically, structured trial-and-error
503 iterative approaches may more accurately maximise consensus among models than deterministic approaches
504 (Dormann *et al.* 2018; Gobeyn *et al.* 2019). The PCA and correlation coefficient approaches (Table 1: En-
505 3 & En-4) deterministically assess consensus among individual models. By contrast, regression to the
506 median, leave-one-out cross validation, and log-likelihood approaches (Table 1: En-5, En-6, En-10) are
507 examples of iterative processes that optimise for the highest level of consensus in full parameter space
508 (Dormann *et al.* 2018). Attribute-based approaches as used by Masson & Knutti (2011) and Willcock *et al.*
509 (2019) (e.g. weighting by model distinctiveness or grid size; Table 1: En-7 and En-8) produce conflicting
510 results. Model attributes such as these may not correctly describe why model outputs vary, or capture their
511 complexity (Willcock *et al.* 2019; Brun *et al.* 2020) and so weighting by among-model agreement produces
512 more accurate ensemble outputs. One might expect accuracy-weighted ensembles (Table 1: En-9) to
513 perform best. However, model accuracy can be location specific and poorly transferable elsewhere – even
514 with similar model accuracy, some grid cells may be well represented by some models and less by others
515 (Graham *et al.* 2008; Marmion *et al.* 2009; Zulian *et al.* 2018). As a result accuracy-derived weights show
516 high uncertainty in areas where training data were not available (i.e. non-forested areas; SI-4), likely because
517 of over-fitting to areas with available data (i.e. forests/woodlands) producing correlative patterns that
518 explain other areas less well. In SI-4, we investigated environmental and spatial drivers of uncertainty
519 among predictions. Broadly, these supplementary results show that carbon models and ES ensembles are
520 less accurate in urban areas. We also find that ensembles for water are less accurate in areas of high rainfall,
521 seasonality and rugosity (see SI-4 for full details). That said, as uncertainty among ES ensembles is almost
522 4-times lower than among individual models, this suggests less need to make the ‘right choice’ of method

523 when selecting an ensemble approach. Thus, although there is some chance of picking a superior individual
524 model (Willcock *et al.* 2018), the risk of a sub-optimal prediction is substantially lowered by applying any
525 ensemble method and this risk is further reduced when a weighted ensemble is used.

526
527 Our results should serve as a ‘call to arms’ for ES researchers and practitioners to increasingly use ensembles
528 of models to support decision-making for sustainability. Using an individual ES model is fraught with
529 concerns as *a priori* it is not known which is the most accurate and choosing only one model can, at worst,
530 result in perverse decisions (Willcock *et al.* 2019). Deriving decisions from an ensemble of ES models
531 provides an improvement over using one model for any location (which may be large or small, depending
532 on the local context and the models used), but also more consistency over space, as model accuracy varies
533 spatially (see results in SI-4). Therefore, using ensemble approaches, and especially weighted ensembles,
534 would increase credibility and so help reduce the implementation gap between research and policy- and
535 decision-making (Wong *et al.* 2014; Willcock *et al.* 2016). We acknowledge the lack of standardised metrics
536 across models and limited computational and financial resources that could restrict the uptake of ensembles
537 – indeed, many practitioners only run a single model. However, given the errors associated with single
538 models (this paper; Willcock *et al.* 2020; Eigenbrod *et al.* 2010), we argue that a single model is inadequate,
539 although more complex models are sometimes more accurate (Willcock *et al.* 2019). The most complex (a
540 priori best) ES models require substantial inputs (i.e. data, computational power, subscription fees, and staff
541 time), and so running multiple models – whilst requiring additional resources – results in a large gain per
542 extra unit resource. For example, as even untrained weighted ensembles developed using iterative
543 approaches (e.g. regression to the median, leave-one-out cross validation) enable a 3-fold reduction in
544 variation, such an ensemble approach seems a reasonable minimum standard for ES modelling – striking
545 the right balance between feasibility and robustness (Willcock *et al.* 2016). Whilst such ensembles will be
546 outperformed by the best-performing individual models, these cannot be identified without running multiple
547 models – a ‘Catch-22’ (Willcock *et al.* 2019). Thus, we recommend that multiple models be developed for
548 ES where they are lacking (e.g. cultural services; Martínez-Harms and Balvanera, 2012; Wong *et al.* 2014),
549 and that those with access to sufficient resources to run multiple models ensure the ensemble outputs are
550 freely available, making the use of these ensembles more feasible and accessible for all (Willcock *et al.*
551 2020).

552

553 **5. Conclusion**

554 We show that in situations with no *a priori* validation evidence guiding model selection, predictions from
555 ensembles of models have a higher accuracy than selecting an individual model by chance. Weighted
556 averaging further improves accuracy, suppressing idiosyncratic differences through producing consensus
557 (Araújo & New 2007; Dormann *et al.* 2018). Doing so not only elevates accuracy but substantially decreases
558 uncertainty among ensemble approaches compared to uncertainty among models, a further indication of
559 increased fit to reality (Chaplin-Kramer *et al.* 2019; Willcock *et al.* 2020). In summary, even if a less
560 accurate ensemble weighting approach is used, one would on average have lower uncertainty than selecting
561 an individual model by chance. Thus, particularly when validation data are not available, we recommend
562 the use of weighted ensembles in ES research to substantially reduce uncertainty and to support robust
563 decision-making for sustainable development.

564

565 **References**

- 566 Abrahms, B. *et al.* (2019). Dynamic ensemble models to predict distributions and anthropogenic risk
567 exposure for highly mobile species. *Divers. Distrib.* **25**, 11821193.
568 <https://doi.org/10.1111/ddi.12940>
- 569 Ahlström, A. *et al.* (2015). Carbon cycle. The dominant role of semi-arid ecosystems in the trend and
570 variability of the land CO₂ sink. *Science* **348**, 895–899. <https://doi.org/10.1126/science.aaa1668>
- 571 Araújo, M.B. & New, M. (2007). Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**,
572 42–47. <https://doi.org/10.1016/j.tree.2006.09.010>.

573 Bagstad, K.J. *et al.* (2013). A comparative assessment of decision-support tools for ecosystem services
574 quantification and valuation. *Ecosyst. Serv.* **5**, 27–39. <https://doi.org/10.1016/j.ecoser.2013.07.004>
575 Barredo, J.I. *et al.* (2012). *A European map of living forest biomass and carbon stock*. (European
576 Commission, Joint Research Centre). [https://op.europa.eu/en/publication-detail/-](https://op.europa.eu/en/publication-detail/-/publication/b9345574-a96f-4417-87ed-1a85d2252834/language-en)
577 [/publication/b9345574-a96f-4417-87ed-1a85d2252834/language-en](https://op.europa.eu/en/publication-detail/-/publication/b9345574-a96f-4417-87ed-1a85d2252834/language-en)
578 Bell, V.A. *et al.* (2018a). The MaRIUS-G2G datasets: Grid-to-Grid model estimates of flow and soil
579 moisture for Great Britain using observed and climate model driving data. *Geosci. Data J.* **5**, 63–72.
580 <https://doi.org/10.1002/gdj3.55>
581 Bell, V.A. *et al.* (2018b). *Grid-to-Grid model estimates of monthly mean flow and soil moisture for*
582 *Great Britain (1891 to 2015): observed driving data [MaRIUS-G2G-Oudin-monthly]*. [Data Set]
583 (NERC Environmental Information Data Centre). [https://doi.org/10.5285/f52f012d-9f2e-42cc-](https://doi.org/10.5285/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0)
584 [b628-9cdea4fa3ba0](https://doi.org/10.5285/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0)
585 Brun, P. *et al.* (2020). Model complexity affects species distribution projections under climate
586 change. *J. Biogeogr.* **47**, 130–142. <https://doi.org/10.1111/jbi.13734>
587 Bryant, B.P. *et al.* (2018). Transparent and feasible uncertainty assessment adds value to applied
588 ecosystem services modeling. *Ecosyst.Serv.* **33**, 103–109.
589 <https://doi.org/10.1016/j.ecoser.2018.09.001>
590 Chaplin-Kramer, R. *et al.* (2019). Global modeling of nature’s contributions to people. *Science* **366**,
591 255–258. <https://science.sciencemag.org/content/366/6462/255.abstract>
592 Costanza, R. *et al.* (2014). Changes in the global value of ecosystem services. *Glob. Environ.*
593 *Change* **26**, 152–158. <https://doi.org/10.1016/j.gloenvcha.2014.04.002>
594 Costanza, R. *et al.* (2017). Twenty years of ecosystem services: how far have we come and how far do
595 we still need to go? *Ecosyst. Serv.* **28**, 1–16. <https://doi.org/10.1016/j.ecoser.2017.09.008>
596 Coxon, G. *et al.* (2019a). DECIPHeR v1: Dynamic fluxEs and ConnectIvity for Predictions of
597 HydRology. *Geosci. Model Dev.* **12**, 2285–2306. <https://doi.org/10.5194/gmd-12-2285-2019>
598 Coxon, G. *et al.* (2019b). *DECIPHeR model estimates of daily flow for 1366 gauged catchments in*
599 *Great Britain (1962-2015) using observed driving data*. [Data Set] (NERC Environmental
600 Information Data Centre). <https://doi.org/10.5285/d770b12a-3824-4e40-8da1-930cf9470858>
601 Crossman, N.D., Bryan, B.A. & Summers, D.M. (2012). Identifying priority areas for reducing species
602 vulnerability to climate change. *Divers. Distrib.* **18**, 60–72. [https://doi.org/10.1111/j.1472-](https://doi.org/10.1111/j.1472-4642.2011.00851.x)
603 [4642.2011.00851.x](https://doi.org/10.1111/j.1472-4642.2011.00851.x)
604 Diengdoh, V.L. *et al.* (2020). A validated ensemble method for multinomial land-cover
605 classification. *Ecol. Inform.* **56**, 101065. <https://doi.org/10.1016/j.ecoinf.2020.101065>
606 Ding, H. & Bullock, J.M. (2018). *A Guide to Selecting Ecosystem Service Models for Decision-*
607 *Making: Lessons from Sub-Saharan Africa*. (World Resources Institute). [wri.org/publication/guide-](http://wri.org/publication/guide-selecting-ecosystem-service)
608 [selecting-ecosystem-service](http://wri.org/publication/guide-selecting-ecosystem-service)
609 Dormann, C.F. *et al.* (2018). Model averaging in ecology: a review of Bayesian, information-theoretic,
610 and tactical approaches for predictive inference. *Ecol. Monogr.* **88**, 485–504.
611 <https://doi.org/10.1002/ecm.1309>
612 Eigenbrod, F. *et al.* (2010) The impact of proxy-based methods on mapping the distribution of
613 ecosystem services. *J. Appl. Ecol.* **47.2**, 377-385.
614 Elith, J. *et al.* (2011). A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57.
615 <https://doi.org/10.1111/j.1472-4642.2010.00725.x>
616 Englund, O., Berndes, G. & Cederberg, C. (2017). How to analyse ecosystem services in landscapes—A
617 systematic review. *Ecol. Indic.* **73**, 492–504. <https://doi.org/10.1016/j.ecolind.2016.10.009>
618 Erceg-Hurn, D.M. & Mirosevich, V.M. (2008). Modern robust statistical methods: an easy way to
619 maximize the accuracy and power of your research. *Am. Psychol.* **63**, 591–601.
620 <http://dx.doi.org/10.1037/0003-066X.63.7.591>
621 Forestry Commission, United Kingdom. (2018). *National Forest Inventory Woodland GB 2018*. [Data
622 Set] (Forestry Commission Open Data). [http://data-](http://data-forestry.opendata.arcgis.com/datasets/d3d7bfba1cba4a3b83a948f33c5777c0_0)
623 [forestry.opendata.arcgis.com/datasets/d3d7bfba1cba4a3b83a948f33c5777c0_0](http://data-forestry.opendata.arcgis.com/datasets/d3d7bfba1cba4a3b83a948f33c5777c0_0)

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

626 Gobeyn, S. *et al.* (2019). Evolutionary algorithms for species distribution modelling: A review in the
627 context of machine learning. *Ecol. Modell.* **392**, 179–195.
628 <https://doi.org/10.1016/j.ecolmodel.2018.11.013>

629 Graham, C.H. *et al.* (2008). The influence of spatial errors in species occurrence data used in
630 distribution models. *J Appl. Ecol.* **45**, 239–247. <https://doi.org/10.1111/j.1365-2664.2007.01408.x>

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

633 Griggs, D. *et al.* (2013). Sustainable development goals for people and planet. *Nature* **495**, 305–307.
634 <https://doi.org/10.1038/495305a>

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

637 Hao, T. *et al.* (2020). Testing whether ensemble modelling is advantageous for maximising predictive
638 performance of species distribution models. *Ecography* **43**, 549–558.
639 <https://doi.org/10.1111/ecog.04890>

640 He, X. . *et al.* (2021). Climate-informed hydrologic modeling and policy typology to guide managed
641 aquifer recharge. *Science Advances* **7**, p.eabe6025. <https://doi.org/10.1126/sciadv.abe6025>

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

644 Henrys, P.A., Keith, A. & Wood, C.M. (2016). *Model estimates of aboveground carbon for Great*
645 *Britain*. [Data Set] (NERC Environmental Information Data
646 Centre). <https://doi.org/10.5285/9be652e7-d5ce-44c1-a5fc-8349f76f5f5c>

647 Inoue, H. & Narihisa, H. (2000) in *Knowledge Discovery and Data Mining. Current Issues and New*
648 *Applications* (eds Terano, T, Liu, H. & Chen, A.L.P.) 177-180 (Springer).
649 <https://link.springer.com/book/10.1007/3-540-45571-X>

650 Kareiva, P. *et al.* (2011). *Natural Capital: Theory and Practice of Mapping Ecosystem Services*.
651 (Oxford University Press).
652 [https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199588992.001.0001/a](https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199588992.001.0001/acprof-9780199588992)
653 [cprof-9780199588992](https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199588992.001.0001/acprof-9780199588992)

654 Keselman, H. J. *et al.* (2008). A generally robust approach for testing hypotheses and setting
655 confidence intervals for effect sizes. *Psychol. Methods* **13**, 110–129.
656 <https://doi.apa.org/doi/10.1037/1082-989X.13.2.110>

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

659 Knutti, R., Masson, D. & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and
660 how we got there. *Geophys. Res. Lett.* **40**, 1194–1199. <https://doi.org/10.1002/grl.50256>

661 Lewis, K.A. . *et al.* (2021). Using multiple ecological models to inform environmental decision-
662 making. *Front. Mar. Sci.* **8**, 283. <https://doi.org/10.3389/fmars.2021.625790>

663 Liu, D., Li, T. & Liang, D. (2020). An integrated approach towards modeling ranked weights. *Comput.*
664 *Ind. Eng.* **147**, 106629. <https://doi.org/10.1016/j.cie.2020.106629>

665 Malinga, R. *et al.* (2015). Mapping ecosystem services across scales and continents—A review. *Ecosyst.*
666 *Serv.* **13**, 57–63. <https://doi.org/10.1016/j.ecoser.2015.01.006>

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

669 Martínez-Harms, M.J. & Balvanera, P. (2012). Methods for mapping ecosystem service supply: a
670 review. *Int. J. Biodivers. Sci. Ecosyst. Serv. Manag.* **8**, 17-25.

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

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

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

678 Ochoa, V. & Urbina-Cardona, N. (2017). Tools for spatially modeling ecosystem services: Publication
679 trends, conceptual reflections and future challenges. *Ecosyst.Serv.* **26**, 155–169.
680 <https://doi.org/10.1016/j.ecoser.2017.06.011>

681 Pascual, U. *et al.* (2017). Valuing nature’s contributions to people: the IPBES approach. *Curr. Opin.*
682 *Environ. Sustain.* **26–27**, 7–16. <https://doi.org/10.1016/j.cosust.2016.12.006>

683 Redhead, J.W. *et al.* (2016). Empirical validation of the InVEST water yield ecosystem service model
684 at a national scale. *Sci. Total Environ.* **569**, 1418–1426 (2016).
685 <https://doi.org/10.1016/j.scitotenv.2016.06.227>

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

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

690 Sharps, K. *et al.* (2017). Comparing strengths and weaknesses of three ecosystem services modelling
691 tools in a diverse UK river catchment. *Sci. Total Environ.* **584**, 118–130.
692 <https://doi.org/10.1016/j.scitotenv.2016.12.160>

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

696 van Soesbergen, A. & Mulligan, M. (2018). Uncertainty in data for hydrological ecosystem services
697 modelling: Potential implications for estimating services and beneficiaries for the CAZ Madagascar.
698 *Ecosyst. Serv.* **33**, 175–186. <https://doi.org/10.1016/j.ecoser.2018.08.005>

699 Suich, H., Howe, C. & Mace, G. (2015). Ecosystem services and poverty alleviation: A review of the
700 empirical links. *Ecosyst. Serv.* **12**, 137–147. <https://doi.org/10.1016/j.ecoser.2015.02.005>

701 Tebaldi, C. & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate
702 projections. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **365**, 2053–2075.
703 <https://doi.org/10.1098/rsta.2007.2076>

704 Thibaut, L.M. & Connolly, S.R. (2013). Understanding diversity–stability relationships: towards a
705 unified model of portfolio effects. *Ecol. Lett.* **16**, 140–150. <https://doi.org/10.1111/ele.12019>

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

708 UKNEA. (2011). *The UK National Ecosystem Assessment: Synthesis of the Key Findings.* (UNEP-
709 WCMC, Cambridge). [https://www.unep-wcmc.org/resources-and-data/UK-national-ecosystem-](https://www.unep-wcmc.org/resources-and-data/UK-national-ecosystem-assessment)
710 [assessment](https://www.unep-wcmc.org/resources-and-data/UK-national-ecosystem-assessment)

711 Verhagen, W. *et al.* (2017). Use of demand for and spatial flow of ecosystem services to identify
712 priority areas. *Conserv. Biol.* **31**, 860–871. <https://doi.org/10.1111/cobi.12872>

713 Wang, H.M. *et al.* (2019). Does the weighting of climate simulations result in a better quantification of
714 hydrological impacts? *Hydrol. Earth Syst. Sci.* **23**, 4033–4050. [https://doi.org/10.5194/hess-23-](https://doi.org/10.5194/hess-23-4033-2019)
715 [4033-2019](https://doi.org/10.5194/hess-23-4033-2019)

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

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

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

723 Wong, C.P. *et al.* (2014). Linking ecosystem characteristics to final ecosystem services for public
724 policy. *Ecol. Lett.* **18**, 108–118. <https://doi.org/10.1111/ele.12389>

725 Zulian, G. *et al.* (2018). Practical application of spatial ecosystem service models to aid decision
726 support. *Ecosyst. Serv.* **29**, 465–480. <https://doi.org/10.1016/j.ecoser.2017.11.005>
727