

Ensembles of ecosystem service models can improve accuracy and indicate uncertainty

Willcock, Simon; Hooftman, Danny; Blanchard, Ryan; Dawson, Terence P.; Hickler, Thomas; Lindeskog, Mats; Martinez-Lopez, Javier; Reyers, Belinda; Watts, Sophie M.; Eigenbrod, Felix; Bullock, James

Science of the Total Environment

DOI:

[10.1016/j.scitotenv.2020.141006](https://doi.org/10.1016/j.scitotenv.2020.141006)

Published: 10/12/2020

Peer reviewed version

[Cyswllt i'r cyhoeddiad / Link to publication](#)

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):

Willcock, S., Hooftman, D., Blanchard, R., Dawson, T. P., Hickler, T., Lindeskog, M., Martinez-Lopez, J., Reyers, B., Watts, S. M., Eigenbrod, F., & Bullock, J. (2020). Ensembles of ecosystem service models can improve accuracy and indicate uncertainty. *Science of the Total Environment*, 747, Article 141006. <https://doi.org/10.1016/j.scitotenv.2020.141006>

Hawliau Cyffredinol / General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 **Title:** Ensembles of ecosystem service models can improve accuracy and indicate uncertainty

2 **Shortened Title:** Ecosystem service model ensembles

3 **Authors:** Simon Willcock^{†*1,2}, Danny A.P. Hooftman^{*3,4}, Ryan Blanchard⁵, Terence P. Dawson⁶, Thomas
4 Hickler^{7,8}, Mats Lindeskog⁹, Javier Martinez-Lopez^{10,11}, Belinda Reyers^{12,13}, Sophie M. Watts², Felix
5 Eigenbrod^{2,14} & James M. Bullock⁴.

6 * Joint first authors (contributed equally)

7 † Corresponding author

8 1. School of Natural Sciences, Bangor University, United Kingdom. s.willcock@bangor.ac.uk

9 2. Biological Sciences, University of Southampton, United Kingdom.
10 sophiemwatts25@gmail.com

11 3. Lactuca: Environmental Data Analyses and Modelling, The Netherlands.
12 danny.hooftman@lactuca.nl

13 4. UK Centre for Ecology and Hydrology, Wallingford, OX10 8BB, United Kingdom.
14 jmbul@ceh.ac.uk

15 5. Council for Scientific and Industrial Research, South Africa. RBlanchard@csir.co.za

16 6. Department of Geography, King's College London, United Kingdom. terry.dawson@kcl.ac.uk;

17 7. Senckenberg Biodiversity and Climate Research Centre (SBIK-F), Germany
18 thomas.hickler@senckenberg.de

19 8. Department of Physical Geography, Goethe University, Frankfurt, Germany

20 9. Department of Physical Geography and Ecosystem Science, Lund University, Sweden.
21 mats.lindeskog@nateko.lu.se

22 10. Soil Erosion and Conservation Research Group, CEBAS-CSIC, Spanish Research Council, Campus
23 de Espinardo, Murcia E-30100, PO Box 164, Spain jmartinez@cebas.csic.es

24 11. BC3 – Basque Centre for Climate Change, 48940, Leioa, Spain

25 12. Future Africa, University of Pretoria, Private bag X20, Hatfield 0028, South Africa

26 13. Stockholm Resilience Centre, Stockholm University, Stockholm SE-10691, Sweden.

27 belinda.reyers@su.se

28 14. Geography and Environment, University of Southampton, United Kingdom.

29 F.Eigenbrod@soton.ac.uk

30

31 **Contributions**

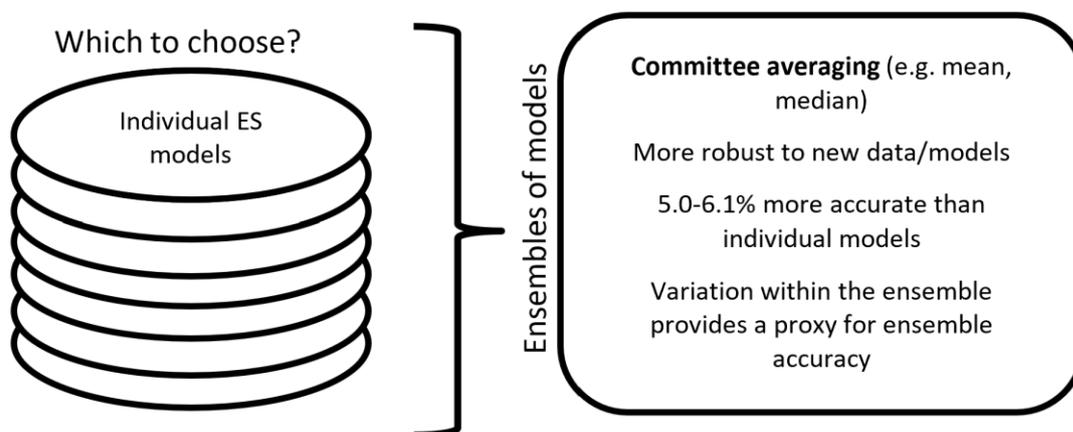
32 FE, SW, DAPH, & JMB conceived the project. DAPH & SW analysed the results. SW, DAPH, FE & JMB

33 wrote the manuscript, with comments and revisions from all other authors.

34 **Abstract**

35 Many ecosystem services (ES) models exist to support sustainable development decisions. However,
36 most ES studies use only a single modelling framework and, because of a lack of validation data, rarely
37 assess model accuracy for the study area. In line with other research themes which have high model
38 uncertainty, such as climate change, ensembles of ES models may better serve decision-makers by
39 providing more robust and accurate estimates, as well as provide indications of uncertainty when
40 validation data are not available. To illustrate the benefits of an ensemble approach, we highlight the
41 variation between alternative models, demonstrating that there are large geographic regions where
42 decisions based on individual models are not robust. We test if ensembles are more accurate by
43 comparing the ensemble accuracy of multiple models for six ES against validation data across sub-
44 Saharan Africa with the accuracy of individual models. We find that ensembles are better predictors
45 of ES, being 5.0-6.1% more accurate than individual models. We also find that the uncertainty (i.e.
46 variation among constituent models) of the model ensemble is negatively correlated with accuracy
47 and so can be used as a proxy for accuracy when validation is not possible (e.g. in data-deficient areas
48 or when developing scenarios). Since ensembles are more robust, accurate and convey uncertainty,
49 we recommend that ensemble modelling should be more widely implemented within ES science to
50 better support policy choices and implementation.

51 **Graphical Abstract**



52

53 **Key words:** Africa; carbon; charcoal; firewood; grazing; model validation; natural capital; poverty
54 alleviation; sustainable development; water.

55 **Highlights:**

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

61

62 **1. Introduction**

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

76 Despite this lack of robustness, most ES modelling applications rely on a single model for each ES
77 (Bryant et al., 2018). For example, the latest state-of-the-art ES models produced via the
78 Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) rely on
79 single model outputs with little/no validation (Chaplin-Kramer et al., 2019). Although, few studies have
80 explicitly validated ES models against independent datasets, there are notable exceptions (Bruijnzeel
81 et al., 2011; Mulligan and Burke, 2005; Redhead et al., 2018, 2016; Sharps et al., 2017; Willcock et al.,
82 2019). Willcock et al. (2019) validated multiple models for several ES, testing their accuracy against
83 empirical data across sub-Saharan Africa. While they found that more complex models (i.e. those
84 representing more processes) were sometimes more accurate (Box 1), their results suggested it would
85 be difficult to select *a priori* the most accurate of a set of models for an ES in any particular context
86 (Willcock et al., 2019).

Box 1 – Key definitions

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

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

87 One solution to inter-model variation is to utilise ensembles and apply appropriate techniques to
88 explore the resulting range of projections. Ensembles are produced by running simulations for more
89 than one set of models, initial conditions, model classes, model parameters and/or boundary
90 conditions (Araújo and New, 2007). For example, since the current state and processes of the system
91 are often uncertain, small differences in initial conditions or model parameters could result in large

92 differences in model projections (van Soesbergen and Mulligan, 2018). Similarly, different model
93 classes (e.g. statistical models vs process-based models) might be considered competing but equally
94 valid representations of a system, and hence worth exploring (Araújo and New, 2007). If only one
95 model is used, conclusions are dependent on the specific assumptions of that model. If an ensemble
96 is used, conclusions are not dependent on that one set of assumptions and parameters, hence one can
97 consider the variation (or uncertainty) in model outcomes and might obtain a better idea of what the
98 reality might be. Single model forecasts have been criticised due to their potential to result in a decision
99 that imposes rigidity, which might have serious negative consequences if there is large uncertainty and
100 inaccuracies (Araújo and New, 2007).

101 Whilst running ensembles of models is not the norm in ES studies (Bryant et al., 2018), this practice is
102 commonplace in other disciplines, most famously for climate and weather modelling (Gneiting et al.,
103 2005; Refsgaard et al., 2014). For example, in contrast to IPBES, Intergovernmental Panel on Climate
104 Change (IPCC) publications regularly use ensembles (Collins et al., 2013). These climate change
105 ensembles generate a consensus prediction by measuring the central tendency (e.g. the mean or
106 median) for the ensemble of forecasts (Araújo and New, 2007). Climate change ensemble forecasts
107 might show enhanced performance over some individual models as the averaging results in a
108 smoothing effect, reducing the impact of idiosyncratic responses of any particular model in the area
109 of space and time of interest (Marmion et al., 2009). In short, by averaging multiple models the signal
110 of interest emerges from the noise associated with individual model uncertainties (Araújo and New,
111 2007; Knutti et al., 2010). Such, so-called, committee averaging gives equal weight to all models. The
112 benefits of these techniques have been observed in multiple disciplines, ranging from agro-ecology
113 (Elias et al., 2017; Refsgaard et al., 2014) and niche modelling (Aguirre-Gutiérrez et al., 2017; Crossman
114 et al., 2012; Grenouillet et al., 2011) to market forecasting (He et al., 2012) and credit risk analysis (Lai
115 et al., 2006).

116 The level of variation within an ensemble (i.e. inconsistency among the individual models) may also be
117 informative in itself. Lower variation within an ensemble of models may indicate increased accuracy

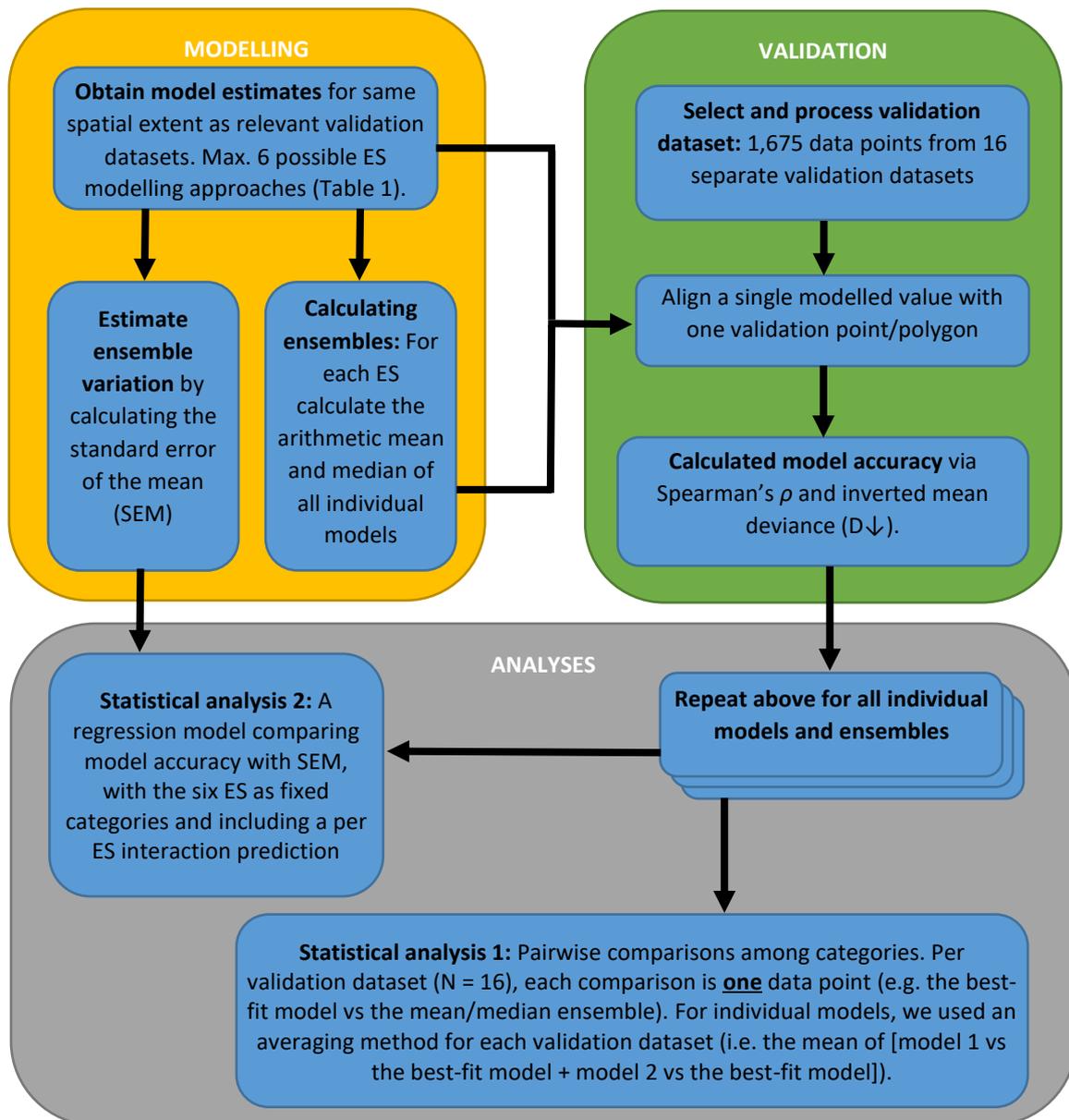
118 of the ensemble mean (Puschendorf et al., 2009). Thus, ensembles may also provide an indication of
119 uncertainty when faced with data scarcity, a potential benefit that is perhaps most pronounced in
120 many developing countries, where data collection and model assessment efforts are least advanced
121 (Suich et al., 2015) but reliance on ES for wellbeing is arguably the highest (Daw et al., 2011; Shackleton
122 and Shackleton, 2012; Suich et al., 2015).

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

131 **2. Methods**

132 Recently we validated multiple models for each of six ES in sub-Saharan Africa (stored carbon, available
133 water, water usage, firewood, charcoal, and grazing resources; Table 1) using 1,675 data points from
134 16 independent datasets (Figure SI1-1; summarised in Table SI1-2, but see Willcock et al. (2019) for
135 further information). In that paper, we used six ES modelling frameworks (InVEST (Kareiva, 2011;
136 McKenzie et al., 2012), Co\$ting Nature (Mulligan, 2015; Mulligan et al., 2010), WaterWorld (Mulligan,
137 2013), benefits transfer based on the Costanza and others (2014) values, LPJ-GUESS (Smith et al., 2014,
138 2001), and the Scholes models (comprising two grazing models and a rainfall surplus model) (Scholes,
139 1998), following Willcock et al. (2019) by using a single set of parameters for each ES per modelling
140 framework, with each framework requiring different inputs (Willcock et al., 2019). We employed two
141 performance metrics to calculate model accuracy in terms of each validation dataset: Spearman's ρ
142 and mean inverse Deviance (D^\downarrow the mean absolute distance between normalised model and validation

143 values per data-point, inversed so that a value of 1 represents a perfect fit). Both metrics have real-
144 world relevance, as decision-making can make use of both relative (e.g. rank order of sites or options)
145 and absolute (e.g. the total amount or value of service delivered) values (Willcock et al. 2016), and ρ
146 ranks locations by their relative ES values, whereas D^\downarrow reflects the degree to which models consistently
147 reflect absolute values in the validation dataset (Willcock et al. 2019). In the work reported here, we
148 use the model outcomes and calculations, and validation data and methods presented in Willcock et
149 al. (2019) (Figure 1). This includes our approach of normalising within model variation to fall within a
150 0-1 scale, following Verhagen et al. (2017), which allows comparability among the different ES studied.
151 The codes we used to do this are deposited here: https://github.com/dhooftman72/ES_Ensembles. All
152 analyses were performed in Matlab (v7.14.0.739), with ArcGIS 10.7 used only for display purposes. P
153 < 0.05 was viewed as statistically significant throughout.



154 **Figure 1** - A summary of the analytical framework, divided into modelling, validation and analysis
 155 subsets.

156

157 *2.1 Creating ensembles*

158 To depict among-model variation per service we divided the modelled areas into km² gridcells – except
 159 water, which is represented in m³ ha⁻¹ per polygon. Since all models do not cover the entire study area,
 160 we recorded the number of models with valid values per gridcell. For every gridcell where ≥ 3 modelled
 161 estimates were available, we calculated model ensembles and mapped the standard error of the mean

162 (SEM) among normalised model values.

163 As described above, ensembles are created by combining individual model outputs, resulting in a
164 smoothing effect whereby the individual model uncertainties are cancelled out and the signal of
165 interest emerges (Araújo and New, 2007; Marmion et al., 2009). However, there are multiple ways by
166 which individual models can be combined into an ensemble. For example, all models could be weighted
167 equally (i.e. committee averaging) or weighted by some measure of reliability or trust. Here, we used
168 committee averaging, but see SI3 for a further exploration of weighting. First, we created committee
169 two ensemble values for each ES by calculating the arithmetic mean and median across the i individual
170 model estimates for each modelled spatial data point (i.e. 1 km² grid cell). To evaluate ensemble
171 accuracy, we compared the ensemble estimate (E) to the validation data for that spatial location as
172 described in Willcock et al. (2019).

173 2.2 Comparing ensembles estimates

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

179 **Equation 1:** $\rho'_i = \left(\frac{\rho_i + 1}{2}\right)$

180 We analysed the proportional change in accuracy (ρ and D^\downarrow) for all possible pairs of comparisons
181 between: (i) the individual models, based on the mean accuracy statistics across the group of all
182 possible models (described below), (ii) the different ensembles (mean/median), and (iii) the best
183 performing model according to each validation dataset. We tested whether the accuracy of a first
184 category ("A", e.g., the ensemble mean) was higher – "improved" – or lower than a second category
185 ("B", e.g., the individual models). The accuracy level differed greatly across the 16 validation datasets
186 and the different ES (Willcock et al., 2019). No among ES comparison is possible as 16 validation

187 datasets across six ES provides too low a level of replication per ES, but normalising each ES allows
188 comparisons across the different ES as a whole. Normalising involved dividing the accuracy of A by the
189 accuracy of B for each validation dataset. For simplicity, we refer to the 16 resulting proportions as
190 “improvement values”, although they could indicate a loss of accuracy (values <1).

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

199 **Equation 2:** $\left(\left(\sum_i^n \frac{A}{B_i} \right) \times \frac{1}{n} \right)$, with n total of models for that validation set (i ; 4-6 models depending on
200 the service; Table 1).

201 This was done for each of the 16 validation sets. This averaging method allowed for a fully balanced
202 analysis, with a single improvement value associated with each of the 16 validation datasets.
203 Alternative analyses in which we included single comparisons for individual models per validation
204 dataset against respective ensemble scores (79 improvement values) showed similar results (Table SI-
205 1-4) as the larger variation was offset by higher degrees of freedom (78 vs 15).

206 We also tested the correlation between ensemble *uncertainty* and absolute *accuracy* using 1661 of the
207 1675 individual data-points for validation (anovan-procedure in Matlab). The large sample size meant
208 we were able to differentiate between ES in this analysis. We calculated ensembles from a minimum
209 of three models and so discarded 14 data-points since they only matched ≤ 2 modelled estimates. For
210 each data-point (X), we calculated the absolute *accuracy* of the mean ensemble ($D^{\downarrow}_{(x)}$) and calculated

211 *uncertainty* as the SEM among-modelled values (Equation 3). For statistical comparison, we used an
 212 SS type 1 mixed regression model with the six ES as fixed variables and SEM_x as the linear predictor,
 213 logit transformed, with correlation coefficient β_1 and constant β_0 , and with a per ES interaction
 214 prediction with uncertainty ($ES_x \times SEM'_x$). We identified a positive Spatial Autocorrelation (SA) for
 215 accuracy with a Moran's I of 0.073 ($P < 0.001$, based on a permutation test), using the Moran's module
 216 from <https://github.com/dhoofman72/Morans-I>. This SA has been corrected for through inclusion of
 217 a covariate within the regression model prior to estimating the model parameters of interest, with
 218 effect size β_{sa} , describing relatedness between individual samples caused by the spatial structure
 219 following Dormann et al. (2007) and Brooks et al. (2016) (Equation 4).

220 **Equation 3:** $SEM_x = \left(\frac{\sigma_x}{\sqrt{n_x}}\right)$, where X represents each 1 km² grid-cell, and n is the number of models.

221 **Equation 4:** $D_{(X)}^\downarrow \sim \beta_{sa}SA_x + ES_x + \beta_1SEM'_x + (ES_x \times SEM'_x) + \beta_0$

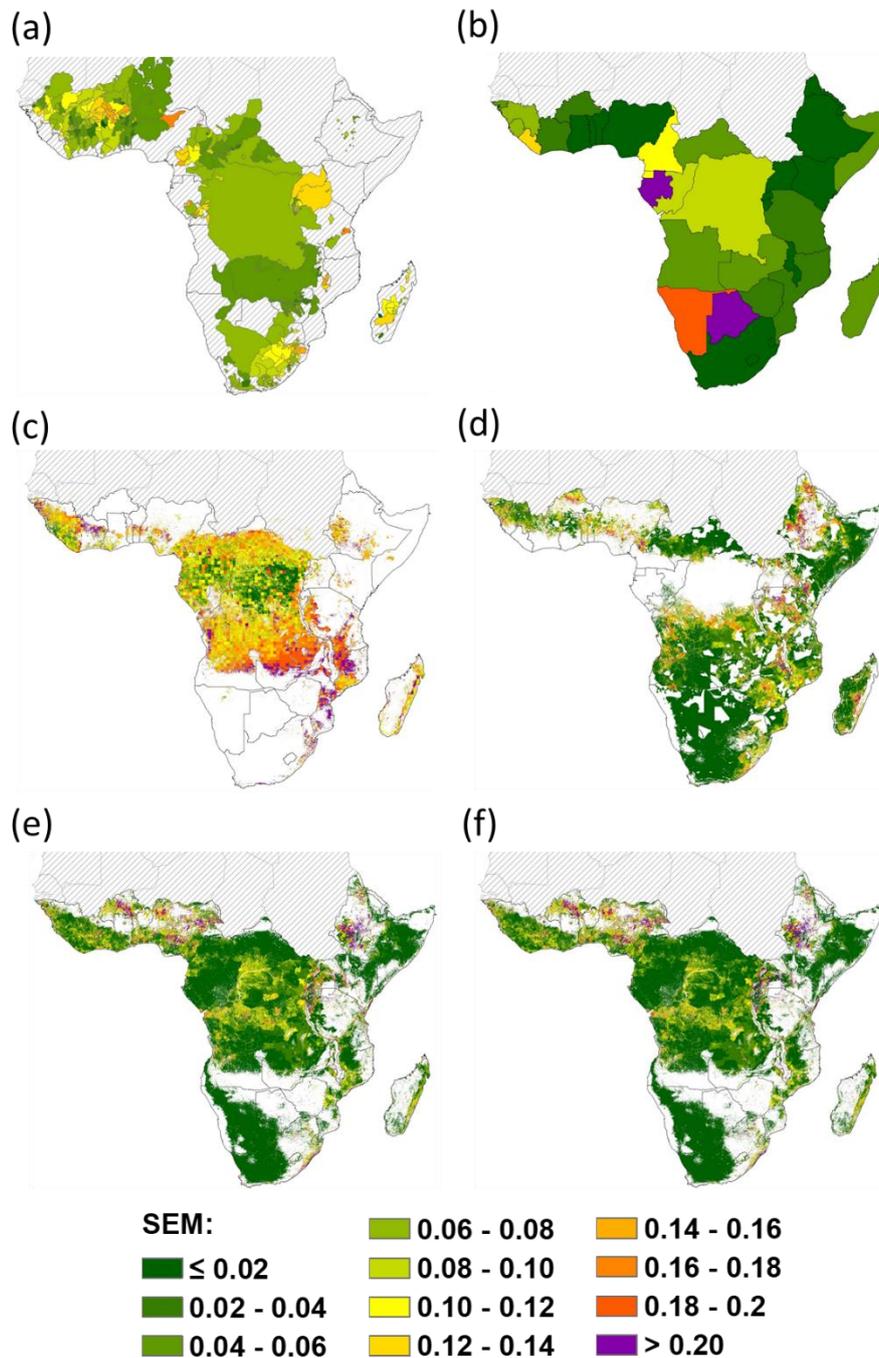
222 With $SEM'_x = \left(\log_{10}\left(\frac{SEM_x}{(1-SEM_x)} + 1\right)\right)$

223

224 **3. Results**

225 *3.1 Variation amongst models shows strong spatial patterning*

226 For sub-Saharan Africa, we found large areas for which the variation among models was relatively low
 227 (Figure 2). In these areas all models provide similar normalised predictions and so a decision based on
 228 a single model may prove robust. However, there are also notable areas of disagreement, where
 229 variation among models was higher. These appear to occur in transition zones between vegetation
 230 types (Figure 2) and, for aboveground carbon storage models, in less densely forested areas (e.g.
 231 miombo woodland; Figure 2). These maps of variation, as well as the mean and median normalised
 232 values, for sub-Saharan Africa at a 1-km-resolution are available through the Environmental
 233 Information Data Centre (EIDC; <https://eidc.ac.uk/>) repository (<https://doi.org/10.5285/11689000-f791-4fdb-8e12-08a7d87ad75f>). See SI2 and SI3 for further uses of multiple models (i.e. hotspots,
 234



236

237 **Figure 2.** Among-model variation measured as standard error of the mean (SEM) using normalised
 238 model predictions. Non-coloured areas were not modelled (i.e. are outside LCM masks or outside the
 239 catchments we analysed). a) Water supply per hectare of the catchment (6 models); b) Water usage
 240 (6 models) per hectare of the country; c) Carbon storage in forest vegetation (4 models); d) Grazing
 241 use (6 models); e) Firewood usage (5 models); f) Charcoal usage (4 models). Firewood and Charcoal

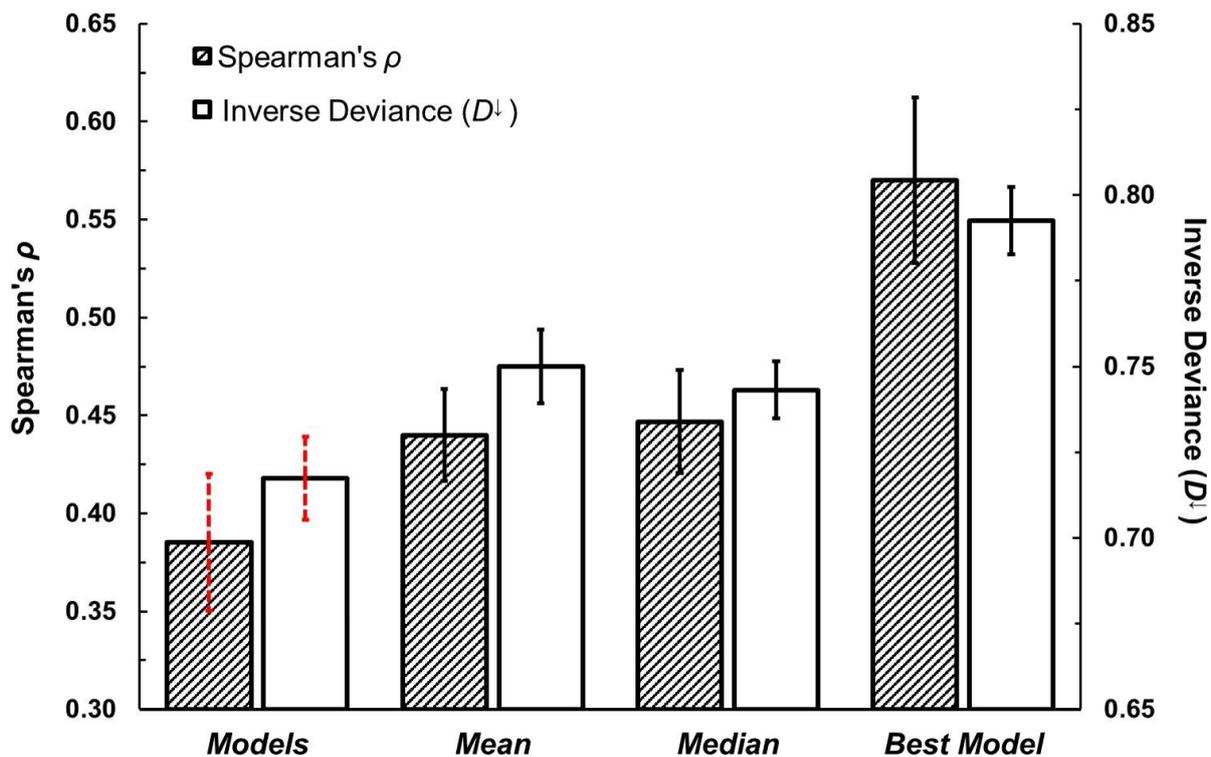
242 have four models in common that are equal once normalised. However, Firewood contains an
 243 additional bespoke Firewood model that generates more variation making (e) and (f) slightly different
 244 (see Willcock et al. (2019) for full model details).

245

246 *Ensembles perform better than individual models, on average*

247 In general, individual models as a group were inferior to the ensembles created from them: ensembles
 248 outperform individual modelling frameworks by 5% to 6% for both ρ and D^\downarrow ($P = 0.03$ and 0.008
 249 respectively; Figure 3; Table SI1-3). Ensembles were outperformed by the best model for each
 250 validation set by 13% (mean; $P = 0.04$) and 12% (median; $P = 0.05$) using ρ and 6% ($P = 0.002$) and 7%
 251 ($P < 0.001$) using D^\downarrow . Unfortunately, which model performs best for each validation dataset was hard
 252 to predict as no single model framework is consistently more accurate than others (Table SI1-1,
 253 Willcock et al. (2019)). A full matrix of statistical results and means and standard errors of these
 254 pairwise comparisons is provided in Table SI1-3.

255



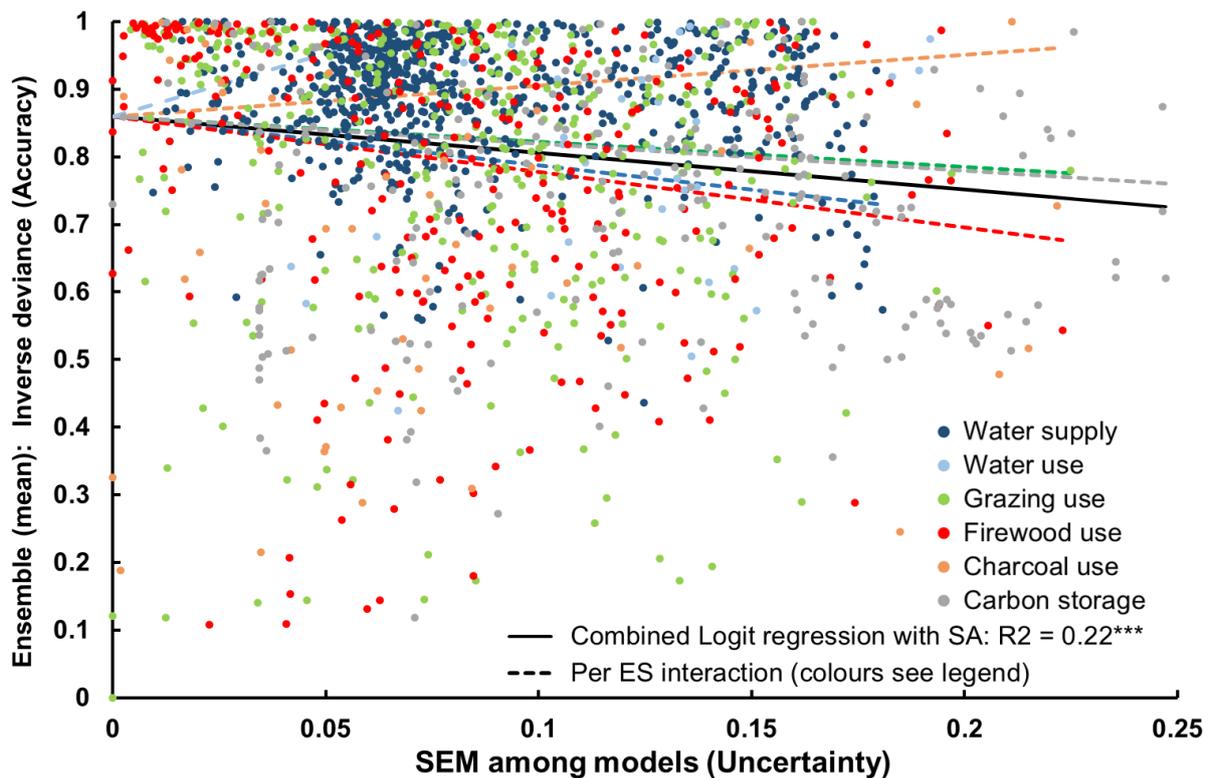
256

257 **Figure 3.** Mean ρ and D^\downarrow of the individual models (as a group), the mean and median ensembles and
258 best-fit individual model. Dark bars = Spearman's ρ ; Light bars = Inverse Deviance D^\downarrow . Black full error
259 bars indicate variation in proportional improvement against the individual models, calculated as
260 $SEM_{imp} = CV_{imp} \times \text{absolute difference}$, with CV the coefficient of variation of proportional improvement
261 based on standard error of the mean (SEM). Thus, error bars indicate the variation in improvement
262 against individual models as a group to highlight the range of improvement of ensemble techniques.
263 $N = 16$ per bar. Red dashed error bars indicate the SEM among all 79 models in this study as indication
264 of overall variation in accuracy.

265

266 *3.2 Accuracy is correlated to ensemble uncertainty*

267 The accuracy of an ensemble in relation to validation datasets could be in part inferred from the
268 variation among the models within the ensemble (Figure 4; F-value = 36.2, $P < 0.001$, $df = 1/1637$). For
269 example, for every 0.1 increase in the SEM among-modelled values, the inverse deviance decreases by
270 0.054. We found no significant interaction effects among ES and uncertainty (F-value 1.09, $df = 5/1637$)
271 suggesting results are generalisable among the tested ES in this study.



272

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

278

279 **4. Discussion**

280 We have demonstrated that there is substantial variation between ES models and the difficulty in
 281 predicting the best-fit model as no single model was consistently better than others (Table S11-1)
 282 (Willcock et al., 2019). These areas of disagreement highlight regions where decisions based on
 283 individual models are likely not robust (Figure 2). For example, all ES models agreed less in transition
 284 zones between vegetation types. The majority of the models used here (and ES models generally)
 285 require input from land cover maps, and transition zones between land cover categories are likely
 286 areas of disagreement between maps. Reasons for this might include land cover maps being produced

287 in different years and so locating the forest frontier in different places, maps/models using slightly
288 different definitions of land cover (and so drawing the boundaries between categories in different
289 places), or because land cover categories are more uncertain in transition zones (Dong et al., 2015),
290 partly due to the difficulties of accounting for degradation (Turner et al., 2016). However, even if
291 vegetation transitions are also simulated (here by a Dynamic Global Vegetation Model, LPJ-GUESS),
292 models are more likely to disagree at a transition zone compared to the central area of a vegetation
293 type. Furthermore, vegetation transitions and carbon storage in sub-Saharan Africa are strongly driven
294 by fire, which is difficult to simulate in process-based models (Hantson et al., 2016). The variation
295 between models due to different initial conditions (i.e. land cover maps) is not the focus of this paper,
296 but has been highlighted previously (van Soesbergen and Mulligan, 2018) and can lead to large error
297 propagation in downstream models (Estes et al., 2018). It is likely that such disagreement is also a key
298 factor driving variation between the ES models considered here. Similarly, aboveground carbon
299 storage models also showed disagreement in less densely forested areas (e.g. miombo woodland).
300 Thus, these differences might partly arise due to uncertainties in the carbon data used to parameterise
301 the models. Savanna and miombo ecosystems are understudied, with tree inventory plots showing a
302 bias towards closed canopy forests (Phillips et al., 2002). Added to this, less densely forested areas
303 show higher natural variation in aboveground carbon storage when compared to closed canopy forests
304 as the land cover category definitions typically cover a wider range of canopy cover (e.g. 10-80% vs 80-
305 100%) (Willcock et al., 2014; Willcock et al., 2012). Thus, further collection of primary data is needed,
306 particularly in the areas of disagreement highlighted here, to improve the next generation of ES
307 models.

308 Despite disagreement between individual models, ensemble modelling has been mostly neglected by
309 the ES community; e.g. a Web of Science search (10 February 2020) for “model ensemble” and
310 “ecosystem service” resulted in no records. This is surprising as: 1) Ensembles are commonly used for
311 model types that simulate output variables closely related to ES, but without emphasising the ES
312 concept in the publication, such as crop models (Rosenzweig et al., 2014), Dynamic Global Vegetation

313 Models simulating carbon uptake (climate mitigation, e.g. Ahlström et al., (2015)) or hydrology models
314 simulating runoff (freshwater supply).; and 2) Other disciplines have found that ensembles can show
315 enhanced robustness and performance over some individual models as the averaging minimises the
316 influence of local idiosyncratic responses of any particular model (Marmion et al., 2009). For example,
317 Inoue and Narihisa (2000) demonstrated that ensemble averaging classification problems resulted in
318 1-7% improvements in accuracy using computational experiments and similar results are widespread
319 in the literature; e.g. for species distribution models (Grenouillet et al., 2011; Marmion et al., 2009),
320 climate change models (Refsgaard et al., 2014), and economic models (He et al., 2012). These findings
321 from other disciplines mirror ours, that ensembles are around 6% more accurate than individual
322 models (Figure 2, Table S11-3). That said, if the desired model output can be validated, then accuracy
323 is increased further by identifying and using the best-fit individual model (gaining a further 12 %
324 increase in accuracy). However, using the best-fit model to support a decision does not necessarily
325 increase its robustness as inclusion of new data or models may shift which model is thought to be most
326 accurate (Table S11-1) (Willcock et al., 2019).

327 Ensembles will likely have the highest utility when validation using primary data is not possible (IPBES,
328 2016). In these situations, individual model accuracy is not known, and committee ensemble methods
329 can yield cost-effective solutions decision support tools (Araújo and New, 2007) (see S13 for a
330 discussion on weighted ensemble techniques). The sustainability agenda desperately requires
331 evidence-based policies and actions for the developing world (Clark et al., 2016). In these regions, ES
332 information is important because the rural and urban poor are often the most dependent on ES (either
333 directly or indirectly (Cumming et al., 2014)), both for their livelihoods (Daw et al., 2011; Suich et al.,
334 2015) and as a coping strategy for buffering shocks (Shackleton and Shackleton, 2012). As such, a single
335 model of unknown certainty could lack credibility, relevance and legitimacy – the major reasons for
336 the ‘implementation gap’ between ES research and its incorporation into policy- and decision-making
337 (Cash et al., 2003; Clark et al., 2016; Wong et al., 2014). Put simply, ensemble models offer a way to
338 reduce as well as acknowledge uncertainty (Bryant et al., 2018) but also potentially offer a future

339 avenue to include other sources of knowledge including local and traditional knowledge in interpreting
340 the outcomes and uncertainty of ensembles to ensure more legitimate and salient knowledge for use
341 in decision making (Díaz et al., 2018; Pascual et al., 2017). Thus, model ensembles may be useful when
342 estimating scenarios of future ES supply and use, but also for contemporary estimates in data deficient
343 areas such as sub-Saharan Africa (Willcock et al., 2016). Furthermore, we suggest that variation among
344 models can provide a first-order estimate of the quality of the prediction when no other information
345 is available (Bryant et al., 2018; Puschendorf et al., 2009). Thus, we believe the benefits of using an
346 ensemble of models in decision-making (increased robustness, increased accuracy over individual
347 models in general, and the ability to estimate uncertainty) substantially outweigh the costs (reduced
348 accuracy when compared to the best-fit model, and additional effort required).

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

359 However, whilst using ensembles of ES models is indeed possible, there are several challenges that
360 need to be overcome before it becomes standard practice within ES science. We argue that advances
361 are necessary in two key areas: accessibility and comparability. As more independent models are
362 developed, it might be hypothesised that the ease with which these models can be accessed might
363 increase. Indeed, anecdotal evidence seems to support this as, for example, InVEST historically
364 required access to expensive ArcGIS software and ARIES required extensive computational skills to run.

365 Accompanying the wider shift towards open science (Fecher and Friesike, 2014), InVEST now runs
366 independently of any commercial software, where results can be mapped using open-source GIS
367 (Bagstad et al., 2013; Peh et al., 2013) and ARIES models can be run by non-experts (Martínez-López
368 et al., 2019). Similarly, despite models becoming increasingly complex, the computational capacity
369 required to run some of these models has decreased as many modelling frameworks now make use of
370 cloud-computing resources, putting less stringent requirements on the end-user (Willcock et al., 2019).

371 Accessing multiple ES models remains a difficult undertaking. For example, whilst the software needed
372 to run InVEST is free, it still requires substantial GIS knowledge and many of the models within this
373 framework are 'data-hungry' and therefore require access to data and substantial processing power in
374 order to run (Willcock et al., 2019). By contrast, ARIES and Co\$ting Nature store the necessary data
375 and processing power on their servers, but therefore require high-speed internet access (Willcock et
376 al., 2019). Furthermore, to benefit from the full Co\$ting Nature model outputs (i.e. disaggregate
377 outputs of individual services) one either needs to enter a partnership with the model owners or pay
378 a subscription of at least 2,000 GBP yr⁻¹ (<http://www.policysupport.org/access-costs>). Thus, in order
379 to contrast or combine, for example, carbon models across these frameworks you require access to
380 the internet, adequate data and computational power, as well as the funds to support a model
381 subscription fee and the extra staff time required (i.e. when compared to running a single model). Such
382 resources are likely out of reach of many ES researchers and practitioners and so, for them, ES
383 ensembles are an unfeasible ideal. However, this can be somewhat negated if those with access to
384 these resources make the ensembles they are able to create freely available (e.g. as we have done so
385 through the EIDC repository for our committee averaged ensembles and the SEM
386 [<https://doi.org/10.5285/11689000-f791-4fdb-8e12-08a7d87ad75f>]).

387 As well as the issues surrounding the feasibility of running ensembles of models, methodological
388 limitations remain. For example, when validating any model (individual or ensembles) a reference of
389 truth is required (Box 1). Validation data have their own intrinsic inaccuracies and so it may be good
390 practice to validate models against more than one dataset per ES to ensure the accuracy assessment

391 is robust (Willcock et al., 2019). Whilst we use multiple sets of validation data here (Table S-1-2), data
392 deficiency prevented further investigations into the sources of the uncertainty we identified; e.g.
393 running simulations to vary initial conditions (e.g. spatial scale (Hou et al., 2013)), model classes, model
394 parameters and/or boundary conditions (Araújo and New, 2007). This is an exciting avenue for future
395 research, which could also compare using ensembles of models to assess uncertainty with other
396 approaches (e.g. probabilistic models (Bagstad et al., 2014; Willcock et al., 2018)). Whilst both
397 approaches are capable of estimating uncertainty, probabilistic approaches avoid the difficulties
398 associated with running multiple models (above) but provide little insight into model-structural
399 uncertainty, when compared to ensembles of models (Stritih et al., 2019). Thus, future investigations
400 should include more individual models with more varied model-structures and create ensembles using
401 a wider variety of algorithms to deepen our current understanding.

402 A further outstanding issue for enabling ensemble modelling is that any comparisons or combinations
403 of modelled outputs must involve matching like-for-like variables. This can be problematic, as, at
404 present, a selection of models for a specific ES might, to some extent, be modelling different
405 constructs. For example, Co\$ting Nature's stored carbon model includes both below- and above-
406 ground carbon while other models predict only above-ground carbon (Willcock et al., 2019). Similar
407 issues arise when linking benefit transfer models (i.e. a valuation output (Costanza et al., 2014)) with
408 both relative and quantitative estimates of available ES resource (i.e. T C ha⁻¹). To reduce these issues
409 and enable like-for-like comparisons, our statistical analyses focused on relative ranking (see Willcock
410 et al. (2019) for further details). Whilst relative rankings allow for some types of questions to be
411 answered and so are useful to support decision-making, biophysical units are required for many
412 sustainable development decisions (Willcock et al., 2019). For example, it is impossible to evaluate if
413 we are operating in the safe and just operating space (Raworth, 2012) without unit estimates
414 predicting if individuals are meeting the threshold supply of a good required to support basic needs,
415 whilst collectively not exceeding planetary thresholds (Rockström et al., 2009). Thus, concerted effort
416 is needed to standardise the outputs of ES models to increase the ease at which they can be compared.

417 Such efforts are perhaps best coordinated by large, multi-national organisations, and so the
418 Ecosystems Service Partnership (ESP) or IPBES could play a central role in defining key reporting
419 metrics, akin to the role of the IPCC in providing good practice guidance on the productions of
420 emissions estimates (Knutti et al., 2010). Due to the large quantity and diversity of ES, this is no small
421 challenge. However, the majority of ES modelling and mapping studies focus on relatively few ES
422 (Willcock et al., 2016) and so these could be prioritised. Furthermore, there is potential to use this
423 guidance to converge with other disciplines by aligning on agreed proxies/outputs required to measure
424 and monitor the attainment of the Sustainable Development Goals (SDGs;
425 <https://sustainabledevelopment.un.org/>) (Xu et al., 2020). At the very least, ES studies must validate
426 model outputs against independent data (Willcock et al., 2019) and transparently convey the identified
427 uncertainty to model users (Bryant et al., 2018; Kleemann et al., 2020). Such practices will increase
428 confidence in ES science and help to reduce the implementation gap between ES models and policy-
429 and decision-making (Cash et al., 2003; Clark et al., 2016; Voinov et al., 2014; Wong et al., 2014).

430 **5. Conclusions**

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

438 **Acknowledgements**

439 This work took place under the ‘WISER: Which Ecosystem Service Models Best Capture the Needs of
440 the Rural Poor?’ project (NE/L001322/1), funded by the UK Ecosystem Services for Poverty Alleviation
441 program (ESPA; www.espa.ac.uk) and ‘EnsemblES - Using ensemble techniques to capture the accuracy

442 and sensitivity of ecosystem service models' (NE/T00391X/1). JML acknowledges the support of the
443 Spanish Government through María de Maeztu excellence accreditation 2018-2021 (Ref. MDM-2017-
444 0714). We thank three anonymous reviewers for their insightful comments that improved this
445 manuscript.

446 **Compliance with Ethical Standards**

447 Conflict of Interest: The authors declare that they have no conflict of interest.

448 **References**

- 449 Aguirre-Gutiérrez, J., Kissling, W.D., Biesmeijer, J.C., WallisDeVries, M.F., Reemer, M., Carvalheiro,
450 L.G., 2017. Historical changes in the importance of climate and land use as determinants of
451 Dutch pollinator distributions. *J. Biogeogr.* 44, 696–707. <https://doi.org/10.1111/jbi.12937>
- 452 Ahlström, A., Raupach, M.R., Schurgers, G., Smith, B., Arneeth, A., Jung, M., Reichstein, M., Canadell,
453 J.G., Friedlingstein, P., Jain, A.K., Kato, E., Poulter, B., Sitch, S., Stocker, B.D., Viovy, N., Wang,
454 Y.P., Wiltshire, A., Zaehle, S., Zeng, N., 2015. Carbon cycle. The dominant role of semi-arid
455 ecosystems in the trend and variability of the land CO₂ sink. *Science* 348, 895–9.
456 <https://doi.org/10.1126/science.aaa1668>
- 457 Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* 22, 42–
458 47. <https://doi.org/10.1016/J.TREE.2006.09.010>
- 459 Bagstad, K.J., Semmens, D.J., Waage, S., Winthrop, R., 2013. A comparative assessment of decision-
460 support tools for ecosystem services quantification and valuation. *Ecosyst. Serv.* 5, 27–39.
461 <https://doi.org/10.1016/j.ecoser.2013.07.004>
- 462 Bagstad, K.J., Villa, F., Batker, D., Harrison-Cox, J., Voigt, B., Johnson, G.W., 2014. From theoretical to
463 actual ecosystem services: mapping beneficiaries and spatial flows in ecosystem service
464 assessments. *Ecol. Soc.* 19, art64. <https://doi.org/10.5751/ES-06523-190264>
- 465 Brooks, E.G.E., Holland, R.A., Darwall, W.R.T., Eigenbrod, F., 2016. Global evidence of positive

466 impacts of freshwater biodiversity on fishery yields. *Glob. Ecol. Biogeogr.* 25, 553–562.
467 <https://doi.org/10.1111/geb.12435>

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

470 Bryant, B.P., Borsuk, M.E., Hamel, P., Oleson, K.L.L., Schulp, C.J.E., 2018. Transparent and feasible
471 uncertainty assessment adds value to applied ecosystem services modeling. *Ecosyst. Serv.* 33,
472 103–109. <https://doi.org/10.1016/J.ECOSER.2018.09.001>

473 Cash, D.W., Clark, W.C., Alcock, F., Dickson, N.M., Eckley, N., Guston, D.H., Jäger, J., Mitchell, R.B.,
474 2003. Knowledge systems for sustainable development. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8086–
475 91. <https://doi.org/10.1073/pnas.1231332100>

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

481 Clark, W.C., Tomich, T.P., van Noordwijk, M., Guston, D., Catacutan, D., Dickson, N.M., McNie, E.,
482 2016. Boundary work for sustainable development: Natural resource management at the
483 Consultative Group on International Agricultural Research (CGIAR). *Proc. Natl. Acad. Sci. U. S. A.*
484 113, 4615–22. <https://doi.org/10.1073/pnas.0900231108>

485 Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichet, T., Friedlingstein, P., Gao, X., Gutowski,
486 W.J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A.J., Wehner, M., 2013. Long-
487 term Climate Change: Projections, Commitments and Irreversibility, in: Stocker, T.F., Qin, D.,
488 Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M.
489 (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the*
490 *Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge

491 University Press, Cambridge, United Kingdom and New York, NY, USA.

492 Costanza, R., de Groot, R., Sutton, P., van der Ploeg, S., Anderson, S.J., Kubiszewski, I., Farber, S.,
493 Turner, R.K., 2014. Changes in the global value of ecosystem services. *Glob. Environ. Chang.* 26,
494 152–158. <https://doi.org/10.1016/j.gloenvcha.2014.04.002>

495 Crossman, N.D., Bryan, B.A., Summers, D.M., 2012. Identifying priority areas for reducing species
496 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)
497 [4642.2011.00851.x](https://doi.org/10.1111/j.1472-4642.2011.00851.x)

498 Cumming, G.S., Buerkert, A., Hoffmann, E.M., Schlecht, E., von Cramon-Taubadel, S., Tschardtke, T.,
499 2014. Implications of agricultural transitions and urbanization for ecosystem services. *Nature*
500 515, 50–57.

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

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

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

513 Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A.,
514 Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder,
515 B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis

516 of species distributional data: a review. *Ecography (Cop.)*. 30, 609–628.
517 <https://doi.org/10.1111/j.2007.0906-7590.05171.x>

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

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

525 Fecher, B., Friesike, S., 2014. Open Science: One Term, Five Schools of Thought, in: *Opening Science*.
526 Springer International Publishing, Cham, pp. 17–47. [https://doi.org/10.1007/978-3-319-00026-](https://doi.org/10.1007/978-3-319-00026-8_2)
527 [8_2](https://doi.org/10.1007/978-3-319-00026-8_2)

528 Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., Gneiting, T., Raftery, A.E., III, A.H.W.,
529 Goldman, T., 2005. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics
530 and Minimum CRPS Estimation. *Mon. Weather Rev.* 133, 1098–1118.
531 <https://doi.org/10.1175/MWR2904.1>

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

535 Guerry, A.D., Polasky, S., Lubchenco, J., Chaplin-Kramer, R., Daily, G.C., Griffin, R., Ruckelshaus, M.,
536 Bateman, I.J., Duraiappah, A., Elmqvist, T., Feldman, M.W., Folke, C., Hoekstra, J., Kareiva, P.M.,
537 Keeler, B.L., Li, S., McKenzie, E., Ouyang, Z., Reyers, B., Ricketts, T.H., Rockström, J., Tallis, H.,
538 Vira, B., 2015. Natural capital and ecosystem services informing decisions: From promise to
539 practice. *Proc. Natl. Acad. Sci. U. S. A.* 112, 7348–55. <https://doi.org/10.1073/pnas.1503751112>

540 Hantson, S., Arneeth, A., Harrison, S.P., Kelley, D.I., Prentice, I.C., Rabin, S.S., Archibald, S., Mouillot, F.,

541 Arnold, S.R., Artaxo, P., Bachelet, D., Ciais, P., Forrest, M., Friedlingstein, P., Hickler, T., Kaplan,
542 J.O., Kloster, S., Knorr, W., Lasslop, G., Li, F., Mangeon, S., Melton, J.R., Meyn, A., Sitch, S.,
543 Spessa, A., van der Werf, G.R., Voulgarakis, A., Yue, C., 2016. The status and challenge of global
544 fire modelling. *Biogeosciences* 13, 3359–3375. <https://doi.org/10.5194/bg-13-3359-2016>

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

547 Hou, Y., Burkhard, B., Müller, F., 2013. Uncertainties in landscape analysis and ecosystem service
548 assessment. *J. Environ. Manage.* 127 Suppl, S117-31.
549 <https://doi.org/10.1016/j.jenvman.2012.12.002>

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

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

559 Kareiva, P.M., 2011. *Natural capital : theory & practice of mapping ecosystem services*. Oxford
560 University Press.

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

566 Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P.J., Hewitson, B., Mearns, L., 2010. Good
567 Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections, in:
568 Stocker, T., Dahe, Q., Plattner, G.-K., Tignor, M., Midgley, P. (Eds.), Meeting Report of the
569 Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi
570 Model Climate Projections. IPCC Working Group I Technical Support Unit, University of Bern,
571 Bern, Switzerland, p. 13.

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

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

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

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

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

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

588 Mulligan, M., Burke, S., 2005. Global cloud forests and environmental change in a hydrological
589 context. DFID FRP Project ZF0216 Final Technical Report. pp74.

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

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

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

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

612 Puschendorf, R., Carnaval, A.C., VanDerWal, J., Zumbado-Ulate, H., Chaves, G., Bolaños, F., Alford,
613 R.A., 2009. Distribution models for the amphibian chytrid *Batrachochytrium dendrobatidis* in
614 Costa Rica: proposing climatic refuges as a conservation tool. *Divers. Distrib.* 15, 401–408.

615 <https://doi.org/10.1111/j.1472-4642.2008.00548.x>

616 Raworth, K., 2012. A safe and just space for humanity: can we live within the doughnut?, Oxfam
617 Discussion Paper. Oxfam, Oxford, UK.

618 Redhead, J.W., May, L., Oliver, T.H., Hamel, P., Sharp, R., Bullock, J.M., 2018. National scale
619 evaluation of the InVEST nutrient retention model in the United Kingdom. *Sci. Total Environ.*
620 610–611, 666–677. <https://doi.org/10.1016/J.SCITOTENV.2017.08.092>

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

624 Refsgaard, J.C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T.A., Drews, M.,
625 Hamilton, D.P., Jeppesen, E., Kjellström, E., Olesen, J.E., Sonnenborg, T.O., Trolle, D., Willems,
626 P., Christensen, J.H., 2014. A framework for testing the ability of models to project climate
627 change and its impacts. *Clim. Change* 122, 271–282. [https://doi.org/10.1007/s10584-013-0990-](https://doi.org/10.1007/s10584-013-0990-2)
628 2

629 Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the
630 environmental modelling process – A framework and guidance. *Environ. Model. Softw.* 22,
631 1543–1556. <https://doi.org/10.1016/J.ENVSOFT.2007.02.004>

632 Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin III, F.S., Lambin, E., Lenton, T.M., Scheffer,
633 M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe,
634 H., Sörlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W.,
635 Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J., 2009.
636 Planetary Boundaries: Exploring the Safe Operating Space for Humanity. *Ecol. Soc.* 14.

637 Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C.,
638 Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T.A.M., Schmid, E., Stehfest, E., Yang,
639 H., Jones, J.W., 2014. Assessing agricultural risks of climate change in the 21st century in a

640 global gridded crop model intercomparison. *Proc. Natl. Acad. Sci. U. S. A.* 111, 3268–73.
641 <https://doi.org/10.1073/pnas.1222463110>

642 Scholes, R.J., 1998. The South African 1: 250 000 maps of areas of homogeneous grazing potential.

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

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

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

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

657 Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J., Jarvis, A., Hashimoto, H., 2015. Disaggregating
658 Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary
659 Data. *PLoS One* 10, e0107042. <https://doi.org/10.1371/journal.pone.0107042>

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

663 Suich, H., Howe, C., Mace, G., 2015. Ecosystem services and poverty alleviation: A review of the

664 empirical links. *Ecosyst. Serv.* 12, 137–147. <https://doi.org/10.1016/j.ecoser.2015.02.005>

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

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

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

676 Voinov, A., Seppelt, R., Reis, S., Nabel, J.E.M.S., Shokravi, S., 2014. Values in socio-environmental
677 modelling: Persuasion for action or excuse for inaction. *Environ. Model. Softw.* 53, 207–212.
678 <https://doi.org/10.1016/J.ENVSOF.2013.12.005>

679 Walker, W.E., Harremoës, P., Rotmans, J., van der Sluijs, J.P., van Asselt, M.B.A., Janssen, P., Krayen
680 von Krauss, M.P., 2003. Defining Uncertainty: A Conceptual Basis for Uncertainty Management
681 in Model-Based Decision Support. *Integr. Assess.* 4, 5–17.
682 <https://doi.org/10.1076/iaij.4.1.5.16466>

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

687 Willcock, S., Hooftman, D.A.P., Balbi, S., Blanchard, R., Dawson, T.P., O'Farrell, P.J., Hickler, T.,
688 Hudson, M.D., Lindeskog, M., Martinez-Lopez, J., Mulligan, M., Reyers, B., Shackleton, C., Sitas,

689 N., Villa, F., Watts, S.M., Eigenbrod, F., Bullock, J.M., 2019. A Continental-Scale Validation of
690 Ecosystem Service Models. *Ecosystems* 22, 1902–1917. [https://doi.org/10.1007/s10021-019-](https://doi.org/10.1007/s10021-019-00380-y)
691 00380-y

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

695 Willcock, S., Phillips, O.L., Platts, P.J., Balmford, A., Burgess, N.D., Lovett, J.C., Ahrends, A., Bayliss, J.,
696 Doggart, N., Doody, K., Fanning, E., Green, J., Hall, J., Howell, K.L., Marchant, R., Marshall, A.R.,
697 Mbilinyi, B., Munishi, P.K.T., Owen, N., Swetnam, R.D., Topp-Jorgensen, E.J., Lewis, S.L., 2012.
698 Towards Regional, Error-Bounded Landscape Carbon Storage Estimates for Data-Deficient Areas
699 of the World. *PLoS One* 7, e44795. <https://doi.org/10.1371/journal.pone.0044795>

700 Willcock, S., Phillips, O.L., Platts, P.J., Balmford, A., Burgess, N.D., Lovett, J.C., Ahrends, A., Bayliss, J.,
701 Doggart, N., Doody, K., Fanning, E., Green, J.M.H., Hall, J., Howell, K.L., Marchant, R., Marshall,
702 A.R., Mbilinyi, B., Munishi, P.K.T., Owen, N., Swetnam, R.D., Topp-Jorgensen, E.J., Lewis, S.L.,
703 2014. Quantifying and understanding carbon storage and sequestration within the Eastern Arc
704 Mountains of Tanzania, a tropical biodiversity hotspot. *Carbon Balance Manag.* 9.

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

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

712

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

Model framework	Description*	Ecosystem services currently available	Spatial grain	Ecosystem service modelled in this study
WaterWorld	An internally parameterised model of accumulated water run-off. This web-based model incorporates all data required for application.	<ul style="list-style-type: none"> • Water Supply 	1 km ² gridcells for continental scale calculations	Water supply
Co\$ting Nature	A web-based series of interactive maps that defines the contribution of ecosystems to the global reservoir of a particular ES and its realisable value (based on flows to beneficiaries of that service).	<ul style="list-style-type: none"> • Biodiversity Resources • Carbon Storage & Sequestration • Recreation value • Hazard Mitigation • Water Quality • Water Supply 	1 km ² gridcells for continental scale calculations	Water supply ≈ Clean water run-off Stored Carbon ≈ above and below ground carbon
LPJ-GUESS	The Lund–Potsdam–Jena General Ecosystem Simulator model (Smith et al., 2014, 2001). LPJ-GUESS is a dynamic vegetation/ecosystem model designed for regional to global applications. The model combines process-based representations of terrestrial vegetation dynamics and land–atmosphere carbon and water exchanges in a modular framework.	<ul style="list-style-type: none"> • Carbon Storage & Sequestration • Nitrogen Storage & Sequestration • Water run-off 	0.5 degree ≈ 55.6 x 55.6 km gridcells	Water supply Woody species carbon Grazing = C3/C4 carbon
InVEST	A suite of free, open-source software models from the Natural Capital Project, used to map and value the goods and services from nature. InVEST returns results in either biophysical or economic terms.	<ul style="list-style-type: none"> • Carbon: Terrestrial & Coastal Storage & Sequestration • Crops: Pollination & Production • Scenic Quality, Recreation & Tourism • Fisheries: Marine & Aquaculture • Habitat: Quality & Risk • Marine Water Quality • Water Quality: Nutrients and Sediment • Water Supply • Wind & Wave Energy 	Any, land-use map input data depending	Water supply Carbon (above ground only)
Benefit transfer	Bespoke adaptations of Costanza and others (2014) for the study region in \$ per hectare. Benefit transfer assumes a	<ul style="list-style-type: none"> • Gas regulation 	Any, land-use map input data	Water yield ≈ Water supply

	constant unit value per hectare of ecosystem type and multiplies that value by the area of each type to arrive at aggregate totals.	<ul style="list-style-type: none"> • Climate regulation • Disturbance regulation • Water regulation • Water supply • Erosion control • Soil formation • Nutrient cycling • Waste treatment • Pollination • Biological control • Habitat/Refugia • Food production • Raw materials • Genetic resources • Recreation • Cultural 	depending	Carbon ≈ Climate regulation value Charcoal use ≈ Raw materials value Firewood use ≈ Raw materials value	
Scholes models	Interpretation of Scholes (1998).	<ul style="list-style-type: none"> • Grazing • Firewood • Water supply** 	Any, input data depending	Water surplus ** ≈ Water supply Grazing use ^{††} Firewood use ^{††}	
New models [§]	Bespoke calculation of Water use per country, calculated as the sum of all run-off per country [#] divided by the full population per country as calculated from Afripop 2010 (Stevens et al. 2015)	Bespoke models made in this study from Willcock et al. (2019)	All models with Water Supply above	Depending on water supply source data	Water use
	Bespoke models for carbon based services grazing, charcoal and firewood using as input the carbon stock output of the existing carbon models and adapted using multiplication factors and spatial masks (see Willcock et al. (2019) for full details).		Co\$ting Nature carbon	Depending on carbon source data	Grazing use
			InVEST carbon		Charcoal use
			LPJ-GUESS woody species carbon		Firewood use
			Benefit transfer carbon		Grazing use

715 * All 1x1 km in this study, unless otherwise noted. Willcock et al. (2019) investigated the impact of spatial scale on ecosystem service models and found no significant impact
 716 (unpublished results). Thus, spatial scales are unlikely to affect results here. § These services were not modelled in these model frameworks when we conducted our model

717 runs (in 2016). We developed new models using carbon stock outputs from existing models as input (see Willcock et al. (2019) for full details). The original models and their
718 developers should not be held responsible for the results from these new models. # except for accumulated flow from WaterWorld which is the sum over all watersheds
719 within countries of the maximum flow per watershed. **Estimated as number of days that precipitation exceeds evapotranspiration, this service was added by the current
720 study to the available Scholes models (Scholes, 1998). †† We have two Scholes grazing models in our study, a generic international model using freely available global data
721 and a locally parameterised South African model (see Willcock et al. (2019) for full details). ‡‡ Modelled at a 5x5 km resolution.