

1 **Title:** A continental scale validation of ecosystem service models
2 **Shortened Title:** Validation of ecosystem service models
3 **Authors:** Simon Willcock^{†*1,2}, Danny A.P. Hooftman^{*3,4}, Stefano Balbi⁵, Ryan Blanchard⁶, Terence P.
4 Dawson⁷, Patrick J. O'Farrell^{6,8}, Thomas Hickler^{9,10}, Malcolm D. Hudson¹¹, Mats Lindeskog¹², Javier
5 Martinez-Lopez⁵, Mark Mulligan⁷, Belinda Reyers¹³, Charlie Shackleton¹⁴, Nadia Sitas⁶, Ferdinando
6 Villa^{5,15}, Sophie M. Watts², Felix Eigenbrod^{2,16} & James M. Bullock⁴.

7 * Joint first authors (contributed equally)

8 † Corresponding author

- 9 1. School of Natural Sciences, Bangor University, United Kingdom. s.willcock@bangor.ac.uk
10 2. Biological Sciences, University of Southampton, United Kingdom.
11 sophiemwatts25@gmail.com
12 3. Lactuca: Environmental Data Analyses and Modelling, The Netherlands.
13 danny.hooftman@lactuca.nl
14 4. Centre for Ecology and Hydrology, Wallingford, OX10 8BB, United Kingdom. jmbul@ceh.ac.uk
15 5. Basque Centre of Climate Change, Spain. stefano.balbi@bc3research.org;
16 javier.martinez@bc3research.org; ferdinando.villa@bc3research.org
17 6. Council for Scientific and Industrial Research, South Africa. POFarrell@csir.co.za;
18 NSitas@csir.co.za; RBlanchard@csir.co.za
19 7. Department of Geography, King's College London, United Kingdom. terry.dawson@kcl.ac.uk;
20 mark.mulligan@kcl.ac.uk;
21 8. Percy FitzPatrick Institute of African Ornithology, University of Cape Town, South Africa.
22 9. Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Germany
23 thomas.hickler@senckenberg.de
24 10. Department of Physical Geography, Goethe University, Frankfurt, Germany
25 11. Engineering and the Environment, University of Southampton, United Kingdom.
26 mdh@soton.ac.uk
27 12. Department of Physical Geography and Ecosystem Science, Lund University, Sweden.
28 mats.lindeskog@nateko.lu.se
29 13. Department of Conservation Ecology and Entomology, Stellenbosch University, South Africa.
30 belinda.reyers@su.se
31 14. Department of Environmental Science, Rhodes University, South Africa.
32 c.shackleton@ru.ac.za
33 15. IKERBASQUE, Basque Foundation for Science. Bilbao, Bizkaia, Spain.
34 16. Geography and Environment, University of Southampton, United Kingdom.
35 F.Eigenbrod@soton.ac.uk
36

37 **Contributions**

38 FE, SW, DAPH, TPD, POF, MDH, BR, CS, FV, & JMB conceived the project. SW, DAPH, RB, TPD, POF,
39 MDH, MM, BR, CS, NS, FV, FE & JMB developed the investigation at a workshop on model complexity.
40 SW & SMW collated and prepared the validation dataset. SW & MM ran the CoSting Nature and
41 WaterWorld models. DAPH ran the InVEST models and Scholes international & water supply models.
42 SW, SB, JML ran LPJ-GUESS, ML provided the LPJ-GUESS code and, with TH, provided advice on the
43 use of LPJ-GUESS. RB, POF & BR ran the local South African grazing and firewood models. DAPH, JMB,
44 SW & FE analysed the results. SW, DAPH, FE & JMB wrote the manuscript, with comments and
45 revisions from all other authors. DAPH collated the SI, with comments and revisions from all other
46 authors.

47 **Abstract**

48 Faced with environmental degradation, governments world-wide are developing policies to safeguard
49 ecosystem services (ES). Many ES models exist to support these policies, but they are generally poorly
50 validated, especially at large scales, which undermines their credibility. To address this gap, we
51 describe a study of multiple models of five ES, which we validate at an unprecedented scale against
52 1,675 data points across sub-Saharan Africa. We find that potential ES (biophysical supply of carbon
53 and water) are reasonably well predicted by existing models. These potential ES models can also be
54 used as inputs to new models for realised ES (use of charcoal, firewood, grazing resources and water),
55 by adding information on human population density. We find that increasing model complexity can
56 improve estimates of both potential and realised ES, suggesting that developing more detailed models
57 of ES will be beneficial. Furthermore, in 85% of cases, human population density alone was as good or
58 a better predictor of realised ES than ES models, suggesting that it is demand, rather than supply that
59 is predominantly determining current patterns of ES use. Our study demonstrates the feasibility of ES
60 model validation, even in data-deficient locations such as sub-Saharan Africa. Our work also shows
61 the clear need for more work on the supply-side of ES models, and the importance of model validation
62 in providing a stronger base to support policies which seek to achieve sustainable development in
63 support of human well-being.

64 **Key words:** Africa; beneficiary; carbon; charcoal; complexity; firewood; grazing; natural capital; water.

65 **Highlights:**

- 66 • We validate multiple ecosystem services (ES) models across sub-Saharan Africa (SSA)
- 67 • We find that more complex ES models sometimes provide more accurate estimates
- 68 • Realised use of ES is closely aligned with human population density (demand) in SSA

69 **Introduction**

70 Ecosystem services (ES) – nature’s contributions to people (Pascual and others 2017) – are of global

71 importance to human well-being, but are increasingly threatened by human activities (Steffen and
72 others 2015). As a result, many governments are now moving to ES-based management of natural
73 resources (Wong and others 2014) and 132 United Nation member states have signed up to the
74 Intergovernmental Science-Policy Platform for Biodiversity and Ecosystem Services (IPBES;
75 www.ipbes.net). This shift in policy requires accurate spatial modelling of ES (Malinga and others
76 2015), as managing ES requires an understanding of their spatial distribution and heterogeneity
77 (Swetnam and others 2011; Spake and others 2017) and the ability to project and compare the
78 outcomes of management scenarios (Willcock and others 2016). Models can provide credible
79 information where empirical ES data are sparse, which is especially the case in many developing
80 countries (Suich and others 2015).

81 To meet demand for an enhanced understanding of ES flows, many spatial modelling methods and
82 tools for mapping ES have been developed, ranging from very simple land-cover based proxies to
83 sophisticated process-based models (IPBES 2016). Whilst a growing literature is comparing the
84 outputs and features of the different tools (Bagstad and others 2013; Turner and others 2016),
85 validation of these models is challenging and thus rare in the literature (Bennett and others 2013).
86 Few studies have validated single ES models against independent datasets, and then only rarely at a
87 larger, country scales (Mulligan and Burke 2005; Bruijnzeel and others 2011; Redhead and others
88 2016, 2018). Even more rare are studies that explicitly validate multiple ES models simultaneously,
89 and these generally involve small areas at catchment scale (Sharps and others 2017). As a
90 consequence, the uncertainties associated with most ES models and the datasets that underpin them
91 remain largely unknown (Bryant and others 2018; van Soesbergen and Mulligan 2018). This is a
92 particular issue as the results of local-scale validation are likely not to be transferable to new locations
93 (Redhead and others 2016) or to the regional and national scales at which ES model outputs are most
94 widely used (Willcock and others 2016). As a result, attempts at validation by those applying models
95 in new settings are all the more important (Bryant and others 2018). Indeed, rescaling social-ecological

96 patterns and processes to different spatial resolutions and extents can induce substantial systematic
97 bias (Grêt-Regamey and others 2014), providing challenges to decision making in situations where
98 model results are the only source of information. Lack of proven credibility, salience and legitimacy
99 are the major reasons for the 'implementation gap' between all ES research (not just ES models) and
100 its incorporation into policy- and decision-making (Cash and others 2003; Voinov and others 2014;
101 Wong and others 2014; Clark and others 2016).

102 Approaches to improve the reliability of model predictions in general include increasing model
103 complexity (defined here as model structural complexity (Kolmogorov 1998), sometimes also referred
104 to as model complicatedness (Sun and others 2016)). Computational capacity has rapidly increased
105 over time, enabling ES models to become more complex and multiple models to be run at higher
106 resolutions across larger spatial ranges (Levin and others 2013). However, increasing the complexity
107 of ecological models typically also increases the amount of data and expertise required for
108 implementation and interpretation, with unclear consequences for the results (Merow and others
109 2014). In short, it is unclear if an investment in increasing model complexity lead to more accurate
110 information for policy- and decision-making on local and regional scales.

111 The unknown credibility of ES models (Voinov and others 2014) is most pronounced where they are
112 arguably most needed – in many developing countries, where data collection and model development
113 efforts are least advanced (Suich and others 2015). Such ES information is important because the rural
114 and urban poor are often the most dependent on ES (either directly or indirectly (Cumming and others
115 2014)), both for their livelihoods (Daw and others 2011; Suich and others 2015) and as a coping
116 strategy for buffering shocks (Shackleton and Shackleton 2012). A major barrier to the understanding
117 and management of these benefit flows to the poor is a lack of information on the potential supply
118 and realised use of ES, particularly in the developing world (Wong and others 2014; Willcock and
119 others 2016; Cruz-Garcia and others 2017). Indeed, the comparisons of ES models to primary data that
120 do exist are all focused on potential and not realised ES (i.e. biophysical supplies only and not actual

121 use by beneficiaries) (Bagstad and others 2014). Analyses need to be disaggregated to focus on how
122 people use ES, from which ecosystems, and how such benefits contribute to the people's well-being
123 (Daw and others 2011; Bagstad and others 2014; Cruz-Garcia and others 2017).

124 In this paper, we validate ES models against measured ES data extending over 36 countries in sub-
125 Saharan Africa, covering 16.7 million km² – over half of the land area of Africa – and including some of
126 the world's poorest regions (Handley and others 2009). We focus on five ES of high policy relevance
127 in sub-Saharan Africa (Willcock and others 2016), and for which validation data exist in multiple
128 locations. The potential supply of two ES (stored carbon and available water) are modelled using
129 existing models and a further three ES (firewood, charcoal, and grazing resources) predominantly
130 using new models generated from stored carbon outputs of existing models. To assess ES use, we
131 developed new standardised models for realised ES (i.e. actual use by people) by weighting models of
132 potential ES (biophysical supply) by human population density for the four measured ES where the
133 location of beneficiaries is important (use of charcoal, firewood, grazing resources and water). We
134 hypothesised that these new realised ES models have higher predictive power than potential ES
135 models for these ES. We also assessed the performance of human population density alone as a
136 predictor of ES use, as this represents the simplest possible globally available ES use model. Our
137 rationale for doing so is that local population density is a straightforward indicator of the number of
138 people making use of the ES, and such a simple approach for modelling realised services would be
139 very useful if it proved to be accurate. We do not focus on comparing specific modelling platforms, as
140 the identification of the best specific model for a particular use may shift as new models are developed
141 and is likely be location specific: such site specific comparisons have been done elsewhere (Bagstad
142 and others 2013; Ochoa and Urbina-Cardona 2017). As such, our aims in this study are two-fold: 1) to
143 compare the general performance of models predicting ES supply (for stored carbon and available
144 water) to realised ES (charcoal, firewood, grazing and water use); and 2) whether more complex ES
145 models make better predictions.

146 **Methods**

147 Our approach to modelling and validation is summarised in Figure 1. We validated existing and new –
148 developed using outputs from existing models (see below) – ES models against ES data, using 1,675
149 data points from 16 independent datasets extending over sub-Saharan Africa (carbon: 214, water:
150 736, firewood: 285, charcoal: 59, grazing: 401; Table 1, Figure 2). We compared approaches for
151 modelling ES ranging in complexity from simple land cover-driven production functions to process-
152 based models (IPBES 2016). As our validation datasets vary in spatial extent and location, we
153 accounted for the effects of spatial extent and context (Figure 1). We tested the hypotheses that ES
154 models incorporating a more complex causal structure have higher predictive power. Since decision-
155 makers in sub-Saharan Africa consider model complexity to mean more inputs being used to model
156 more processes (Willcock and others 2016), we assessed model complexity in terms of the number of
157 input variables, defining inputs as a coherent set of values covering the research area for a single
158 feature, be it categorical or numerical (Merow and others 2014).

159 **Description of ecosystem service models.** We selected ES models to test, focussing on: 1) ES models
160 capable of estimating some of our selected potential ES (stored carbon, available water) and providing
161 inputs to our new models of firewood, charcoal, and grazing resources within our study area; 2) the
162 subset of these models for which adequate validation data could be identified, allowing like-for-like
163 comparisons between modelled outputs and validation data; and 3) models representing a range of
164 complexities from simple production functions to process-based models. As such, we used six existing
165 ES modelling frameworks that contain one or more models meeting these criteria (Table SI-1-1, SI-1-
166 1). We selected InVEST (Kareiva 2011; McKenzie and others 2012), Co\$ting Nature (Mulligan and
167 others 2010; Mulligan 2015), WaterWorld (Mulligan 2013) and benefits transfer (based on coupling
168 the Costanza and others (2014) values with GlobCover 2009 landcover categories; SI-1-2) due to their
169 widespread use and global applicability (Bagstad and others 2013). We also included the well-known
170 and partially validated (Pachzelt and others 2015) dynamic vegetation model LPJ-GUESS (Smith and

171 others 2001, 2014). Although LPJ-GUESS is not traditionally considered an ES model and has a
172 relatively coarse native resolution (0.5 x 0.5 degrees, but constrained mainly by the resolution of
173 environmental input variables), it is increasingly used for ES modelling applications (including
174 implementation within the ARIES technology (Villa and others 2014)) and it is a process-based model
175 that gives outputs that effectively track the biophysical supply of many potential ES (Bagstad and
176 others 2014; Lee and Lautenbach 2016). Furthermore, we included the Scholes models (comprising
177 two grazing models and a rainfall surplus model) as it is the only large-scale ES models designed
178 specifically for use in sub-Saharan Africa (Scholes 1998) (SI-1-3). Ideally, we would also have compared
179 bespoke local models with local data. However, such models simply do not exist in sub-Saharan Africa
180 in most places. Moreover, as the global models we compare run at fine spatial resolutions (except LPJ-
181 GUESS), it is reasonable to investigate how well they perform in terms of accuracy against local data
182 collected in many locations in many different ways (as is the case here).

183 At time of analysis (March 2017), InVEST, Co\$ting Nature, and LPJ-Guess did not have models that
184 focus on firewood, charcoal or grazing resources, but they did explicitly output stored vegetation
185 carbon. As the supply of these three ES is directly dependent on the amount of biomass present, which
186 is what underpins estimates of stored vegetation carbon in all three models, we built eight new
187 predictors using the outputs from these three existing carbon modules (to estimate the potential
188 supply of these three additional ES (SI-1-4). These new models used spatial masks to estimate the
189 biomass available on relevant land uses (SI-1). For example, we applied a “grazing” spatial mask to
190 derive grassland carbon from InVEST and Co\$ting Nature standing carbon outputs. We excluded areas
191 in which little to no grazing activity was expected (e.g. protected areas) and included areas in which
192 most of the above-ground carbon is assumed to be available for livestock grazing (Figure SI-1-1A; Table
193 SI-1-5). For LPJ-Guess we used C3/C4 carbon outputs as estimate for grazing resources. Thereafter, we
194 converted grazing biomass to FAO livestock units for sub-Saharan Africa using the conversion factors
195 from Houerou & Hoste (1977). Henceforth, we refer to these carbon-based predictors as ES models.

196 Finally, we created models of realised ES by weighting models of potential ES (models of biophysical
197 supply only; e.g. the Scholes models, WaterWorld and our new models of firewood, charcoal and
198 grazing resources) by human population (Stevens and others 2015). We also conducted like-for-like
199 comparisons of these new models for realised use of water, firewood, charcoal and grazing resources
200 with relative rural population data alone – the simplest possible model of ES use. We also assessed
201 whether these new realised ES models have higher predictive power than potential ES models when
202 compared to ES use data. We excluded urban populations for all analyses except the Poverty
203 Environment Network usage data and water use (Table 1).

204 **Validation datasets.** As we considered the performance of ES models separately for each ES, we did
205 not require locations that provided primary data on all ES together. This enabled us to access 1,675
206 data points from 16 separate validation datasets – the maximum number available to us that were
207 suitable for the purposes of this study (i.e. independent of the model calibration data; Figure 2, Table
208 1, SI-2). These data are diverse, being collected using a range of methods of varying reliability,
209 including: expert opinion (e.g. country-level statistics from the FAO), census data (e.g. district level for
210 Kenya and Ethiopia, household level for South Africa), and biophysical measurement (e.g. tree
211 inventory plots, and weir data on water flow [both from across sub-Saharan Africa]) (Table 1). As such,
212 each dataset has associated uncertainties (Grainger 2008) but, since the ‘true value’ can never be
213 absolutely determined, provides acceptable reference values for validation. Given that the datasets
214 cover a wide range of independent methods and our focus is on ranked correlative relationships
215 between models and data, there is unlikely to be systematic bias and so data quality issues should
216 impact our results minimally. In our analyses, some of the validation data required processing to
217 ensure like-for-like comparison with modelled outputs. All ES models were either run at 1 x 1 km or
218 resampled from their minimum native resolution to an exact 1 x 1 km resolution (i.e. for the Scholes
219 Firewood model [native resolution: 5 x 5km] and for LPJ-Guess [native resolution: 55.6 x 55.6 km]).
220 We then extracted a single summary value per polygon to align model outputs with polygon validation

221 data (e.g. each catchment for the Global Runoff Data Centre [GRDC]; each district for Kenya; each
222 country for FAO data; see SI-2). For forest plot point validation data (the ForestPlots.net data) we
223 compared the point data to the 1 x 1 km grid cell it was in. For the PEN data (fodder, charcoal and
224 firewood use), we buffered the point estimate of each village location by 10 km (to align with walking
225 distances for firewood or water (Agarwal 1983; Sewell and others 2016)) and calculated the summary
226 value for each model for each polygon. Hence we extracted model data to be as comparable as
227 possible to the validation data-points. This means that single values as similar in area and units as
228 possible were extracted from each model to be compared to the single validation values as provided
229 by the datasets listed in Table 1 (see SI-2 for full details of these methods). All data were normalised
230 following Verhagen and others (2017) to equalise any unit differences (SI-3-3).

231 **Statistical analyses**

232 *Calculation of model performance*

233 There is no single comprehensive measure of model performance (Bennett and others 2013). Criteria
234 commonly considered are: 1) trueness – the closeness of the agreement between the reference value
235 and the average model value, largely affected by systematic error or bias within the model); 2)
236 precision – the closeness of agreement between repeated model runs, largely affected by random
237 variables or distributions that feature within the model code; and 3) accuracy – an overall summary
238 of precision and trueness that describes the closeness of the agreement between the reference value
239 and the values obtained from the model run(s) (IOS 1994). We focussed on accuracy and trueness
240 here, as we only considered a single output dataset from each model (derived from a single set of
241 parameters) and assessed these using two metrics. The first metric was the rank correlation between
242 modelled and validation values (Spearman's ρ) – a measure of accuracy ranging from -1 to 1, with 1
243 indicating a perfect positive correlation, 0 no correlation, and -1 a perfect negative correlation. Thus,
244 ρ is a useful metric as in many cases policy makers want to rank locations by their relative ES values
245 (Willcock and others 2016). The second metric was the average absolute deviance of modelled values

246 from the 1-to-1 line representing a perfect fit of normalised model values to the normalised validation
247 values – a measure of accuracy and trueness, as it reflects the degree to which models consistently
248 reflect validation values (SI-2). In our normalised setting (with values inverted for consistency with ρ),
249 deviance ranged from 0 (poor fit) to 1 (perfect fit). For interpretation we follow the generally used
250 criteria employed in AUC (area under the curve) in which a result below 0.7 should be considered as
251 likely random (Swets and others 1979; Marmion and others 2009; Hooftman and others 2016) and a
252 value ≥ 0.7 shows a close fit between the modelled value and the validation data. It is entirely possible
253 for a model to have a high rank correlation value, but also have high deviance from the 1-1 line and
254 vice versa, so the two metrics are complementary (Table 2, Figure 3). We calculated both metrics
255 separately for every ES model for each relevant validation dataset, with the ES models run at 1 x 1 km
256 spatial resolution in most instances, giving 100 comparisons (Carbon: 12; Water supply: 21; Water use:
257 18; Charcoal use: 8; Firewood use: 15; Grazing use: 26 Figure 1).

258 *The effect of model complexity, spatial extent and adding beneficiaries on ES model performance*

259 For each ES, we assessed model complexity in terms of the number of input variables (input complexity
260 [IC]). We considered an input to include a coherent set of values covering a geographic region for a
261 single feature (e.g. land use or elevation). GIS processing without changing the feature parameter was
262 not considered an additional input, and neither were layers created by combining inputs, although the
263 parameter values of an equation could be independent, single-value datasets (see SI-4 for full details).
264 Thus, our new models (developed via GIS processing of the outputs from existing models of ES
265 potential) retained the complexity score of the associated existing model (Figure 1; SI-4). As such, our
266 complexity metric captures the generalisation that models with large numbers of equations generally
267 require more inputs (Sun and others 2016). From a user experience perspective, this complexity often
268 relates to the sourcing and processing of these required input datasets (Willcock and others 2016).
269 Our continuous complexity metric is more subtle and precise than simple categorisation of models,
270 e.g. process-based vs production function. Therefore, it allowed us to advance from previous model
271 comparisons, which often focus on identification of the best model specific to a location, by identifying

272 generalizable conclusions relating model complexity to model accuracy. We log-transformed the IC
273 value (LIC) in all instances as the data were skewed by extreme values.

274 Importantly, we considered each separate model vs validation dataset comparison a *single*
275 independent data point (e.g. InVEST stored carbon validated against carbon storage per unit area
276 derived from tree inventory plots was a single data point) (Figure 1). This is because we were
277 interested in assessing how well models performed in general in sub-Saharan Africa for different types
278 of validation data collected in different locations. This approach also enabled us to use very different
279 types of validation datasets, thereby overcoming the issue of there not being consistent validation
280 data for all ES across most of sub-Saharan Africa.

281 By considering each model vs validation dataset comparison (in terms of rank correlation or deviance)
282 a single data point, we were able to build separate generalised linear models (GLMs) for each of the
283 two model performance measures (the response variable y ; rank correlation or deviance), and for
284 each ES. In each case, the GLM was: $y \sim \text{Complexity Measure} + \text{Spatial Extent}$. Thus, LIC was chosen as
285 the complexity metric, with spatial extent (local, regional, country) modelled as a fixed factor. This
286 allowed us to test if more complex models better predict the biophysical supply or realised use of each
287 of ES, while controlling for any effects of spatial extent. Where the validation data were of realised ES,
288 we compared models of potential ES, ES demand, and realised ES.

289 **Results**

290 **Model Validation.** In general, at least one model for each ES produced outputs that represented the
291 validation data well, calculated in terms of their deviance measure and Spearman's ρ , with deviance
292 showing better fits (mean of the least squares mean [LSM] values for best fit model: $\rho = 0.43$, Deviance
293 = 0.76; Table 2; Figure 3; SI-2). *Potential ES:* The LSM value of the response variable for the best-fit
294 model showed that the best of the existing models of potential ES (carbon and water supply) matched
295 the validation data well (mean LSM value for best fit potential ES models: $\rho = 0.69$, Deviance = 0.82;

296 Table 2; Figure 3A). *Realised ES*: Whilst still producing reasonable fit to validation data, the new models
297 of realised ES did not show as good a fit as the models of potential ES to their respective validation
298 data (mean LSM value for best-fit realised ES models: $\rho = 0.30$, Deviance = 0.73; Table 2; Figure 3B).
299 When compared to realised ES data (Table 3), some (3 of 8 [38%]) of the simple models of realised ES
300 performed better than models of ES potential, and none performed worse. However, for our models
301 of realised charcoal, firewood and grazing services, a majority (45 of 47 [96%]) were predicted as well
302 by human population density alone as by our models, and in two cases (4%) population density was a
303 better predictor than our models (p -values < 0.05; Table 2). By contrast, the comparison of realised
304 water with the water use data showed population density to be a worse predictor than our new
305 realised ES models (6 of 6 [100%]; Table 2).

306 **Model Complexity.** Our comparisons showed either no (1 of 4 [25%] potential ES; 6 of 8 [75%] realised
307 ES) or a positive (3 of 4 [75%] potential ES; 2 of 8 [25%] realised ES) effect of complexity on model fit,
308 with no cases of a negative effect (Table 3). Responses to model complexity were not consistent
309 among the two model accuracy metrics (ρ and Deviance), reflecting their different properties. Notably,
310 we found positive effects across both metrics for stored carbon, but complexity was more rarely a
311 significant predictor of model fit for firewood use, charcoal use and water availability, and in these
312 cases was only detected for one of the two accuracy metrics. Grazing use and water use showed no
313 effect of complexity for either metric (Table 3).

314 **Discussion**

315 This study – the first multi-country validation of multiple ES models (to the best of our knowledge) –
316 suggests that existing ES models provide good predictions across two potential ES of high policy
317 relevance (Willcock and others 2016). But, for the ES models we investigated, models of potential ES
318 (biophysical supply) were more accurate than our new models of realised ES (use; Table 2). This
319 difference can be explained by the facts that: a) building models for realised ES is more challenging;
320 and b) there is a research bias towards the biophysical supply of a few provisioning and regulating

321 services – food supply, water availability, and stored carbon (Egoh and others 2012; Martínez-Harms
322 and Balvanera 2012; Wong and others 2014).

323 *The importance of social systems*

324 Decision-makers require information on a wide range of ES and across a variety of temporal and spatial
325 scales (Scholes and others 2013; McKenzie and others 2014; Willcock and others 2016). Meeting these
326 needs will require a shift in the focus of most models towards understanding the beneficiaries of ES
327 and quantifying their demand, access to and utilisation of services, as well as the consequences for
328 well-being (Bagstad and others 2014; Poppy and others 2014). Whilst some studies (Hamann and
329 others 2016) and models (Mulligan 2015; Suwarno and others 2018; Martínez-López and others 2019)
330 do include the demand and use of ES, our new models of realised ES (created by weighting outputs of
331 models of ES potential by human population) generally showed lower predictive power when
332 compared with the ability of existing models of ES potential to predict biophysical supply. Indeed,
333 many of our new models were unable to predict ES more accurately than human population density
334 alone (Table 2, Table 3). This suggests that rural human population density is a good proxy for ES
335 demand, and realised use of ES is closely aligned with demand in sub-Saharan Africa. The only
336 exception is water use, where our new models were better predictors of realised water use than
337 human population density (Table 2). Further combining social science theory and data to explain the
338 social-ecological processes of ES co-production, use and well-being consequences will likely result in
339 substantial improvements in our understanding and estimates of ES use (Bagstad and others 2014;
340 Díaz and others 2015; Suich and others 2015; Pascual and others 2017). This is an area of active
341 research and some modelling frameworks are beginning to address this deficiency. Socio-economic
342 data on ES use, perceptions, and well-being contributions collected over large regions can and has
343 been incorporated into models to address questions about the impacts of ecosystem change on the
344 well-being of regional and socio-economic groups (Díaz and others 2015; Hamann and others 2016;
345 Egarter Vigl and others 2017). Spatial multi-criteria analyses can be used to model how consistent

346 available potential ES are with local demand, highlighting trade-offs between beneficiary groups
347 where demand varies (Martínez-López and others 2019). Furthermore, the impact of individual
348 decision-making on ES use can be captured through the integration of agent-based models depicting
349 human behaviour with biophysical models (Villa and others 2017; Suwarno and others 2018)).
350 Coupling models of potential ES with models of demand to estimate realised ES will likely result in
351 models that are more complex than existing models (Zhang and others 2017), which our findings
352 suggest could improve accuracy, and new modelling techniques (e.g. machine learning (Willcock and
353 others 2018)) may be needed to enable this (Bryant and others 2018).

354 *The impact of model complexity*

355 The effect of model complexity on the accuracy of ES results has not been investigated in detail
356 previously. For example, a Web of Science search (20 June 2018) for ‘model’ and ‘complexity’ and
357 ‘accuracy’ and ‘ecosystem service’ resulted in only 19 studies, few of which actually assess how ES
358 model complexity affects accuracy. Our results suggested a tendency for ES model complexity to be
359 correlated with increased performance (particularly for potential ES), and strong evidence that
360 increased model complexity does not lead to worse predictions (for both potential and realised ES).
361 However, each successive increase in complexity brings diminishing returns. For example, for each
362 unit increase in LIC for models of stored carbon, ρ increased by 0.25 and deviance by 0.10 (Table 3).
363 Since LIC is log-transformed, each unit increase is achieved by a ten-fold increase in inputs.
364 Furthermore, a trade-off with benefits of additional complexity may be the feasibility of running and
365 interpreting such models (Willcock and others 2016). Results from elsewhere in the literature are
366 mixed, often dependent on the specific context of the comparison. For example, Villarino and others
367 (2014) compare simple “Tier 1” carbon accounting methods with more complex “Tier 2” methods,
368 reporting increased accuracy with model complexity. However, studies that extend this analysis to the
369 most complex “Tier 3” methods report limits in the gains in accuracy, with intermediate “Tier 2” and
370 complex “Tier 3” models producing similar predictions (Hill and others 2013; Willcock and others

2014). Furthermore, increasing model complexity does not necessarily lead to better model performance when predicting ground water recharge rates (von Freyberg and others 2015) nor agricultural yield (Quiroz and others 2017). Nevertheless, model performance should not be the only variable considered when selecting between models of differing complexity. From an ecological perspective, simple functional forms (e.g. linear or nonlinear regression equations having a sufficiently high explanatory power) can be easier to interpret and translate into applications (i.e. from science to policy). However, they may lack predictive power in novel locations and/or future time points if they insufficiently represent spatial heterogeneity in form and process (Syfert and others 2013). A certain level of complexity may be required before sufficiently reliable results can be obtained (Merow and others 2014; Salmina and others 2016), such as our observation that human population is poor predictor of water use, likely as it completely fails to capture the behaviour of the presence and flow of water. Simpler models may accurately represent more basic aspects of a system (e.g. estimating natural capital) but incorporation of additional complexity may be necessary to describe the underlying processes accurately (e.g. the interactions and feedbacks between people and ecosystems) (Merow and others 2014; Willcock and others 2014; Dunham and Grand 2016), and how different trade-offs and benefit flows can be understood and managed. Thus, model complexity should be considered in terms of how complex the ES being modelled are, what objectives need to be met, and to what end.

389 *Limitations*

390 Our analysis comes with several important caveats with respect to validation. Here, we highlight these, in part to act as a 'call to arms' for ES scientists concerning areas demanding further development.

392 Primary data collection, particularly at large scales, should be a priority for ES scientists. As validation of modelled outputs must involve like-for-like comparisons (i.e. comparing potential ES outputs to biophysical supply and realised ES outputs to observed ES use), we were unable to validate all models as well as we would have liked. For example, we were unable to include the models of realised ES

396 produced by Co\$ting Nature (Mulligan and others 2010; Mulligan 2015) due to the lack of
397 corresponding validation data.

398 Another priority for future work is to link better different types of ES models to bespoke validation
399 data to understand their performance fully. For instance, the different carbon models we used to
400 some extent model different constructs. Co\$ting Nature's stored carbon model includes both below-
401 and above-ground carbon while other models predict only above-ground carbon (see SI-1), and we
402 validated these with above-ground carbon data. As below-ground carbon can exceed that of above-
403 ground, these data are not ideal to validate the Co\$ting Nature model. As below-ground carbon in
404 forests is often consistent across forests or proportional to above-ground carbon storage (Lewis and
405 others 2013), it is unlikely that this particular issue affected our findings. However, similar issues arise
406 when linking the Costanza and others (2014) benefit transfer models with validation data, as the
407 former estimate value but are validated against either biophysical or use data (SI-1). Since benefit
408 transfer models are derived by combining global values with land cover data, one might expect the
409 values to be more indicative of the biophysical supply of services, but be poorly matched to actual ES
410 use. To reduce these issues and enable like-for-like comparisons, we generated new models (e.g. for
411 firewood, charcoal, and grazing resources) from stored carbon outputs of existing models (SI-1) in our
412 analyses. These new models used spatial masks to estimate the biomass available on relevant land
413 uses (SI-1). The outputs from these new models are likely to be overestimates as, for example, not all
414 grassland vegetation will be grazed and not all grazed land will stock at maximum capacity (Fetzel and
415 others 2017). However, since our statistical analyses focused on relative ranking (see methods), it is
416 unlikely that these uncertainties impacted our findings greatly (i.e. sites with the highest maximum
417 capacity are likely to be the sites with highest potential and/or realised grazing).

418 More generally, it may be good practice to validate models against more than one dataset, as
419 validation data have their own intrinsic inaccuracies. For example, in this study we used more than
420 one validation dataset for each ES (Table 1). More work is required to understand how best to validate

421 ES models, allowing model validation to become standard practice within the ES community,
422 increasing confidence and helping to reduce the implementation gap between ES models and policy-
423 and decision-making (Cash and others 2003; Voinov and others 2014; Wong and others 2014; Clark
424 and others 2016). However, there will always be financial and practical limits to model validation,
425 especially at large scales. Collection of high quality data is challenging and expensive, and as such
426 would require further investments; indeed, the reason ES models are used is often because of the lack
427 of primary data.

428 Finally, more work is required to develop and test more complex use models. Whilst we highlight that
429 ES models need to move beyond biophysical production to realised use by beneficiaries, our very
430 simple ES use models require substantial improvement, e.g. by incorporating flows of ES (Bagstad and
431 others 2014; Villa and others 2014). Similarly, none of the models we consider here adequately
432 represent temporal dynamics (i.e. when are ES being used?) (Scholes and others 2013; Willcock and
433 others 2016), nor can they disaggregate between beneficiary groups (i.e. who is using which services?)
434 (García-Nieto and others 2013; Bagstad and others 2014), nor estimate if such use is sustainable. All
435 three points are highly relevant to understanding if the Sustainable Development Goals
436 (<https://sustainabledevelopment.un.org/>) are being achieved, and hence represent a critical and
437 hugely challenging frontier in both ES modelling and validation. This is further complicated by the fact
438 that model reliability may differ across spatial scales (Scholes and others 2013). For example, since the
439 focus of decision-makers across sub-Saharan Africa predominantly ranges from local to national
440 scales, they require ES information at different gridcell sizes (Willcock and others 2016), and so it is
441 necessary to understand better how the accuracy of ES models varies with spatial resolution.

442 **Conclusions**

443 Our study demonstrates the feasibility of ES model validation, even in data-deficient locations such as
444 sub-Saharan Africa (Suich and others 2015; Willcock and others 2016). While this demonstration has
445 been long overdue, the lack of such large-scale, multi-model validations is perhaps reflective of the

446 complications involved. In partnership with decision-makers, the advances suggested here could help
447 to ensure ES research continues to inform ongoing policy processes (Voinov and others 2014) (such as
448 the IPBES, the Sustainable Development Goals and CBD Aichi targets). Our findings are of particular
449 relevance to sub-Saharan Africa. Whilst the continent is perceived as relatively data-deficient (Suich
450 and others 2015; Willcock and others 2016), we have shown that adequate data exist to run and
451 validate multiple models for ES of high policy relevance (Willcock and others 2016), particularly
452 relating to supplies of ES. Thus, ES models could help to meet the information demand from policy-
453 makers in sub-Saharan Africa (Willcock and others 2016).

454 **Acknowledgements**

455 This work took place under the ‘WISER: Which Ecosystem Service Models Best Capture the Needs of
456 the Rural Poor?’ project (NE/L001322/1), funded by the UK Ecosystem Services for Poverty Alleviation
457 program (ESPA; www.espa.ac.uk). ESPA receives its funding from the UK Department for International
458 Development, the Economic and Social Research Council and the Natural Environment Research
459 Council. We thank Kate Schreckenber and the anonymous reviewers, whose input substantially
460 improved the manuscript.

461 **Compliance with Ethical Standards**

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

463 **References**

- 464 Agarwal B. 1983. Diffusion of rural innovations: Some analytical issues and the case of wood-burning
465 stoves. *World Dev* 11:359–76.
- 466 Avitabile V, Herold M, Heuvelink GBM, Lewis SL, Phillips OL, Asner GP, Armston J, Ashton PS, Banin L,
467 Bayol N, Berry NJ, Boeckx P, de Jong BHJ, Devries B, Girardin CAJ, Kearsley E, Lindsell JA, Lopez-
468 Gonzalez G, Lucas R, Malhi Y, Morel A, Mitchard ETA, Nagy L, Qie L, Quinones MJ, Ryan CM,
469 Ferry SJW, Sunderland T, Laurin GV, Gatti RC, Valentini R, Verbeeck H, Wijaya A, Willcock S.
470 2016. An integrated pan-tropical biomass map using multiple reference datasets. *Glob Chang*
471 *Biol* 22:1406–20.
- 472 Bagstad KJ, Semmens DJ, Waage S, Winthrop R. 2013. A comparative assessment of decision-support
473 tools for ecosystem services quantification and valuation. *Ecosyst Serv* 5:27–39.
- 474 Bagstad KJ, Villa F, Batker D, Harrison-Cox J, Voigt B, Johnson GW. 2014. From theoretical to actual
475 ecosystem services: mapping beneficiaries and spatial flows in ecosystem service assessments.

476 Ecol Soc 19:art64.

477 Bennett ND, Croke BFW, Guariso G, Guillaume JHA, Hamilton SH, Jakeman AJ, Marsili-Libelli S,
478 Newham LTH, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD,
479 Andreassian V. 2013. Characterising performance of environmental models. *Environ Model*
480 *Softw* 40:1–20.

481 Bruijnzeel LA, Mulligan M, Scatena FN. 2011. Hydrometeorology of tropical montane cloud forests:
482 emerging patterns. *Hydrol Process* 25:465–98.

483 Bryant BP, Borsuk M, Hamela P, Olesond KLL, Schulpe CJE, Willcock S. 2018. Transparent and feasible
484 uncertainty assessment can add value to applied ecosystem services modeling. *Ecosyst Serv*.

485 Cash DW, Clark WC, Alcock F, Dickson NM, Eckley N, Guston DH, Jäger J, Mitchell RB. 2003.
486 Knowledge systems for sustainable development. *Proc Natl Acad Sci U S A* 100:8086–91.

487 Chave J, Andalo C, Brown S, Cairns MA, Chambers JQ, Eamus D, Folster H, Fromard F, Higuchi N, Kira
488 T, Lescure J-PP, Nelson BW, Ogawa H, Puig H, Riera B, Yamakura T, Fölster H, Riéra B. 2005.
489 Tree allometry and improved estimation of carbon stocks and balance in tropical forests.
490 *Oecologia* 145:87–99.

491 Clark WC, Tomich TP, van Noordwijk M, Guston D, Catacutan D, Dickson NM, McNie E. 2016.
492 Boundary work for sustainable development: Natural resource management at the
493 Consultative Group on International Agricultural Research (CGIAR). *Proc Natl Acad Sci U S A*
494 113:4615–22.

495 Costanza R, de Groot R, Sutton P, van der Ploeg S, Anderson SJ, Kubiszewski I, Farber S, Turner RK.
496 2014. Changes in the global value of ecosystem services. *Glob Environ Chang* 26:152–8.

497 Cruz-Garcia GS, Sachet E, Blundo-Canto G, Vanegas M, Quintero M. 2017. To what extent have the
498 links between ecosystem services and human well-being been researched in Africa, Asia, and
499 Latin America? *Ecosyst Serv* 25:201–12.

500 Cumming GS, Buerkert A, Hoffmann EM, Schlecht E, von Cramon-Taubadel S, Tschardt T. 2014.
501 Implications of agricultural transitions and urbanization for ecosystem services. *Nature* 515:50–
502 7.

503 Daw T, Brown K, Rosendo S, Pomeroy R. 2011. Applying the ecosystem services concept to poverty
504 alleviation: the need to disaggregate human well-being. *Environ Conserv* 38:370–9.

505 Díaz S, Demissew S, Carabias J, Joly C, Lonsdale M, Ash N, Larigauderie A, Adhikari JR, Arico S, Baldi A,
506 Bartuska A, Baste IA, Bilgin A, Brondizio E, Chan KM, Figueroa VE, Duraiappah A, Fischer M, Hill
507 R, Koetz T, Leadley P, Lyver P, Mace GM, Martin-Lopez B, Okumura M, Pacheco D, Pascual U,
508 Pérez ES, Reyers B, Roth E, Saito O, Scholes RJ, Sharma N, Tallis H, Thaman R, Watson R, Yahara
509 T, Hamid ZA, Akosim C, Al-Hafedh Y, Allahverdiyev R, Amankwah E, Asah ST, Asfaw Z, Bartus G,
510 Brooks LA, Caillaux J, Dalle G, Darnaedi D, Driver A, Erpul G, Escobar-Eyzaguirre P, Failler P,
511 Fouda AMM, Fu B, Gundimeda H, Hashimoto S, Homer F, Lavorel S, Lichtenstein G, Mala WA,
512 Mandivenyi W, Matczak P, Mbizvo C, Mehrdadi M, Metzger JP, Mikissa JB, Moller H, Mooney
513 HA, Mumby P, Nagendra H, Nesshover C, Oteng-Yeboah AA, Pataki G, Roué M, Rubis J, Schultz
514 M, Smith P, Sumaila R, Takeuchi K, Thomas S, Verma M, Yeo-Chang Y, Zlatanova D. 2015. The
515 IPBES Conceptual Framework — connecting nature and people. *Curr Opin Environ Sustain*
516 14:1–16.

517 Dunham K, Grand JB. 2016. Effects of model complexity and priors on estimation using sequential
518 importance sampling/resampling for species conservation. *Ecol Modell* 340:28–36.

519 Egarter Vigl L, Depellegrin D, Pereira P, de Groot R, Tappeiner U. 2017. Mapping the ecosystem
520 service delivery chain: Capacity, flow, and demand pertaining to aesthetic experiences in
521 mountain landscapes. *Sci Total Environ* 574:422–36.

522 Egoh B, Drakou EG, Dunbar MB, Maes J, Willemsen L. 2012. Indicators for mapping ecosystem
523 services: a review. Report EUR 25456 EN. Luxembourg, Luxembourg

524 Fetzl T, Havlik P, Herrero M, Kaplan JO, Kastner T, Kroisleitner C, Rolinski S, Searchinger T, Van
525 Bodegom PM, Wirsén S, Erb K-H. 2017. Quantification of uncertainties in global grazing
526 systems assessment. *Global Biogeochem Cycles* 31:1089–102.

527 von Freyberg J, Moeck C, Schirmer M. 2015. Estimation of groundwater recharge and drought
528 severity with varying model complexity. *J Hydrol* 527:844–57.

529 García-Nieto AP, García-Llorente M, Iniesta-Arandia I, Martín-López B. 2013. Mapping forest
530 ecosystem services: From providing units to beneficiaries. *Ecosyst Serv* 4:126–38.

531 Grainger A. 2008. Difficulties in tracking the long-term global trend in tropical forest area. *Proc Natl*
532 *Acad Sci* 105:818–23. <http://www.pnas.org/content/105/2/818.abstract>

533 Grêt-Regamey A, Weibel B, Bagstad KJ, Ferrari M, Geneletti D, Klug H, Schirpke U, Tappeiner U. 2014.
534 On the effects of scale for ecosystem services mapping. *PLoS One* 9:e112601.

535 Hamann M, Biggs R, Reyers B. 2015. Mapping social–ecological systems: Identifying ‘green-loop’ and
536 ‘red-loop’ dynamics based on characteristic bundles of ecosystem service use. *Glob Environ*
537 *Chang* 34:218–26.

538 Hamann M, Biggs R, Reyers B, Pomeroy R, Abunge C, Galafassi D. 2016. An Exploration of Human
539 Well-Being Bundles as Identifiers of Ecosystem Service Use Patterns. Cebrian J, editor. *PLoS*
540 *One* 11:e0163476.

541 Handley G, Higgins K, Sharma B, Bird K, Cammack D. 2009. Poverty and poverty reduction in sub-
542 Saharan Africa: An overview of the issues. London, UK

543 Hill TC, Williams M, Bloom AA, Mitchard ETA, Ryan CM. 2013. Are Inventory Based and Remotely
544 Sensed Above-Ground Biomass Estimates Consistent? Bond-Lamberty B, editor. *PLoS One*
545 8:e74170.

546 Hooftman DAP, Edwards B, Bullock JM. 2016. Reductions in connectivity and habitat quality drive
547 local extinctions in a plant diversity hotspot. *Ecography (Cop)* 39:583–92.

548 Houerou HN Le, Hoste CH. 1977. Rangeland Production and Annual Rainfall Relations in the
549 Mediterranean Basin and in the African Sahelo-Sudanian Zone. *J Range Manag* 30:181.

550 IOS. 1994. ISO 5725-1:1994 Accuracy (trueness and precision) of measurement methods and results
551 - Part 1: General principles and definitions.

552 IPBES. 2016. Methodological assessment of scenarios and models of biodiversity and ecosystem
553 services. (Ferrier S, Ninan KN, Leadley P, Alkemade R, Acosta LA, Akçakaya HR, Brotons L,
554 Cheung WWL, V.Christensen, Harhash KA, Kabubo-Mariara J, Lundquist C, Obersteiner M,
555 Pereira H, Peterson G, Pichs-Madruga R, Ravindranath N, Rondinini C, Wintle BA, editors.).
556 Bonn, Germany: Secretariat of the Intergovernmental Platform for Biodiversity and Ecosystem
557 Services

558 Kareiva PM. 2011. Natural capital : theory & practice of mapping ecosystem services. Oxford
559 University Press

560 Kolmogorov AN. 1998. On tables of random numbers. *Theor Comput Sci* 207:387–95.

561 Laporte N, Merry F, Baccini A, Goetz S, Stabach J, Bowman M. 2008. Réduire les émissions de CO2 du
562 déboisement et de la dégradation dans la République Démocratique du Congo: Un premier
563 aperçu. Falmouth, MA, USA.

564 Lee H, Lautenbach S. 2016. A quantitative review of relationships between ecosystem services. *Ecol*
565 *Indic* 66:340–51.

566 Levin S, Xepapadeas T, Crépin A-S, Norberg J, de Zeeuw A, Folke C, Hughes T, Arrow K, Barrett S,
567 Daily G, Ehrlich P, Kautsky N, Mäler K-G, Polasky S, Troell M, Vincent J, Walker B. 2013. Social-
568 ecological systems as complex adaptive systems: modeling and policy implications. *Environ Dev*
569 *Econ* 18:111–32.

570 Lewis SL, Sonké B, Sunderland T, Begne SK, Lopez-Gonzalez G, van der Heijden GMF, Phillips OL,
571 Affum-Baffoe K, Baker TR, Banin L, Bastin J-F, Beeckman H, Boeckx P, Bogaert J, De Cannière C,
572 Chezeaux E, Clark CJ, Collins M, Djangbletey G, Djuikouo MNK, Droissart V, Doucet J-L, Ewango
573 CEN, Fauset S, Feldpausch TR, Foli EG, Gillet J-F, Hamilton AC, Harris DJ, Hart TB, de Haulleville
574 T, Hladik A, Hufkens K, Huygens D, Jeanmart P, Jeffery KJ, Kearsley E, Leal ME, Lloyd J, Lovett JC,
575 Makana J-R, Malhi Y, Marshall AR, Ojo L, Peh KS-H, Pickavance G, Poulsen JR, Reitsma JM, Sheil
576 D, Simo M, Steppe K, Taedoumg HE, Talbot J, Taplin JRD, Taylor D, Thomas SC, Toirambe B,
577 Verbeeck H, Vleminckx J, White LJ, Willcock S, Woell H, Zeng L. 2013. Above-ground

578 biomass and structure of 260 African tropical forests. *Philos Trans R Soc B Biol Sci* 368.

579 Malinga R, Gordon LJ, Jewitt G, Lindborg R. 2015. Mapping ecosystem services across scales and
580 continents – A review. *Ecosyst Serv* 13:57–63.

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

583 Martínez-Harms MJ, Balvanera P. 2012. Methods for mapping ecosystem service supply: a review.
584 *Int J Biodivers Sci Ecosyst Serv Manag* 8:17–25.

585 Martínez-López J, Bagstad KJ, Balbi S, Magrath A, Voigt B, Athanasiadis I, Pascual M, Willcock S, Villa
586 F. 2019. Towards globally customizable ecosystem service models. *Sci Total Environ* 650:2325–
587 36.

588 McKenzie E, Posner S, Tillmann P, Bernhardt JR, Howard K, Rosenthal A. 2014. Understanding the
589 use of ecosystem service knowledge in decision making: lessons from international experiences
590 of spatial planning. *Environ Plan C Gov Policy* 32:320–40.

591 McKenzie E, Rosenthal A, Bernhardt J, Girvetz E, Kovacs K, Olwero N, Tof J. 2012. *Guidance and Case
592 Studies for InVEST Users*. (WWF, editor.). Washington, USA: World Wildlife Fund

593 Merow C, Smith MJ, Edwards TC, Guisan A, McMahon SM, Normand S, Thuiller W, Wüest RO,
594 Zimmermann NE, Elith J. 2014. What do we gain from simplicity versus complexity in species
595 distribution models? *Ecography (Cop)* 37:1267–81.

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

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

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

603 Mulligan M, Guerry A, Arkema K, Bagstad K, Villa F. 2010. Capturing and quantifying the flow of
604 ecosystem services. In: Silvestri S, Kershaw F, editors. *Framing the flow: Innovative Approaches
605 to Understand, Protect and Value Ecosystem Services Across Linked Habitats*. Cambridge, UK:
606 UNEP World Conservation Monitoring Centre. pp 26–33.

607 Ochoa V, Urbina-Cardona N. 2017. Tools for spatially modeling ecosystem services: Publication
608 trends, conceptual reflections and future challenges. *Ecosyst Serv* 26:155–69.

609 Pachzelt A, Forrest M, Rammig A, Higgins SI, Hickler T. 2015. Potential impact of large ungulate
610 grazers on African vegetation, carbon storage and fire regimes. *Glob Ecol Biogeogr* 24:991–
611 1002.

612 Pascual U, Balvanera P, Díaz S, Pataki G, Roth E, Stenseke M, Watson RT, Başak Dessane E, Islar M,
613 Kelemen E, Maris V, Quaas M, Subramanian SM, Wittmer H, Adlan A, Ahn S, Al-Hafedh YS,
614 Amankwah E, Asah ST, Berry P, Bilgin A, Breslow SJ, Bullock C, Cáceres D, Daly-Hassen H,
615 Figueroa E, Golden CD, Gómez-Baggethun E, González-Jiménez D, Houdet J, Keune H, Kumar R,
616 Ma K, May PH, Mead A, O'Farrell P, Pandit R, Pengue W, Pichis-Madruga R, Popa F, Preston S,
617 Pacheco-Balanza D, Saarikoski H, Strassburg BB, van den Belt M, Verma M, Wickson F, Yagi N.
618 2017. Valuing nature's contributions to people: the IPBES approach. *Curr Opin Environ Sustain*
619 26–27:7–16.

620 Poppy GM, Chiotha S, Eigenbrod F, Harvey CA, Honzák M, Hudson MD, Jarvis A, Madise NJ,
621 Schreckenberk K, Shackleton CM, Villa F, Dawson TP. 2014. Food security in a perfect storm:
622 using the ecosystem services framework to increase understanding. *Philos Trans R Soc London
623 B Biol Sci* 369.

624 Quiroz R, Loayza H, Barreda C, Gavilán C, Posadas A, Ramírez DA. 2017. Linking process-based potato
625 models with light reflectance data: Does model complexity enhance yield prediction accuracy?
626 *Eur J Agron* 82:104–12.

627 Redhead JW, May L, Oliver TH, Hamel P, Sharp R, Bullock JM. 2018. National scale evaluation of the
628 InVEST nutrient retention model in the United Kingdom. *Sci Total Environ* 610–611:666–77.

629 Redhead JW, Stratford C, Sharps K, Jones L, Ziv G, Clarke D, Oliver TH, Bullock JM. 2016. Empirical
630 validation of the InVEST water yield ecosystem service model at a national scale. *Sci Total*
631 *Environ*:1–9.

632 Salmina ES, Wondrousch D, Kühne R, Potemkin VA, Schüürmann G. 2016. Variation in predicted
633 internal concentrations in relation to PBPK model complexity for rainbow trout. *Sci Total*
634 *Environ* 550:586–97.

635 Scholes R, Reyers B, Biggs R, Spierenburg M, Duriappah A. 2013. Multi-scale and cross-scale
636 assessments of social–ecological systems and their ecosystem services. *Curr Opin Environ*
637 *Sustain* 5:16–25.

638 Scholes RJ. 1998. The South African 1: 250 000 maps of areas of homogeneous grazing potential.

639 Sewell SJ, Sewell SJ, Desai S. 2016. The impacts of undeveloped roads on the livelihoods of rural
640 women. *Rev Soc Sci* 1:15–29.

641 Shackleton SE, Shackleton CM. 2012. Linking poverty, HIV/AIDS and climate change to human and
642 ecosystem vulnerability in southern Africa: consequences for livelihoods and sustainable
643 ecosystem management. *Int J Sustain Dev World Ecol* 19:275–86.

644 Sharps K, Masante D, Thomas A, Jackson B, Redhead J, May L, Prosser H, Cosby B, Emmett B, Jones L.
645 2017. Comparing strengths and weaknesses of three ecosystem services modelling tools in a
646 diverse UK river catchment. *Sci Total Environ* 584:118–30.

647 Smith B, Prentice IC, Sykes MT. 2001. Representation of vegetation dynamics in the modelling of
648 terrestrial ecosystems: comparing two contrasting approaches within European climate space.
649 *Glob Ecol Biogeogr* 10:621–37.

650 Smith B, Wårlind D, Arneeth A, Hickler T, Leadley P, Siltberg J, Zaehle S. 2014. Implications of
651 incorporating N cycling and N limitations on primary production in an individual-based dynamic
652 vegetation model. *Biogeosciences* 11:2027–54.

653 van Soesbergen A, Mulligan M. 2018. Uncertainty in data for hydrological ecosystem services
654 modelling: Potential implications for estimating services and beneficiaries for the CAZ
655 Madagascar. *Ecosyst Serv* 33:175–86.

656 Spake R, Lasseur R, Crouzat E, Bullock JM, Lavorel S, Parks KE, Schaafsma M, Bennett EM, Maes J,
657 Mulligan M, Mouchet M, Peterson GD, Schulp CJE, Thuiller W, Turner MG, Verburg PH,
658 Eigenbrod F. 2017. Unpacking ecosystem service bundles: Towards predictive mapping of
659 synergies and trade-offs between ecosystem services. *Glob Environ Chang* 47:37–50.

660 Steffen W, Richardson K, Rockström J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de
661 Vries W, de Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers
662 B, Sörlin S. 2015. Planetary boundaries: Guiding human development on a changing planet.
663 *Science* (80-) 347.

664 Stevens FR, Gaughan AE, Linard C, Tatem AJ, Jarvis A, Hashimoto H. 2015. Disaggregating Census
665 Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data.
666 Amaral LAN, editor. *PLoS One* 10:e0107042.

667 Suich H, Howe C, Mace G. 2015. Ecosystem services and poverty alleviation: A review of the
668 empirical links. *Ecosyst Serv* 12:137–47.

669 Sun Z, Lorscheid I, Millington JD, Lauf S, Magliocca NR, Groeneveld J, Balbi S, Nolzen H, Müller B,
670 Schulze J, Buchmann CM. 2016. Simple or complicated agent-based models? A complicated
671 issue. *Environ Model Softw* 86:56–67.

672 Suwarno A, van Noordwijk M, Weikard H-P, Suyamto D. 2018. Indonesia’s forest conversion
673 moratorium assessed with an agent-based model of Land-Use Change and Ecosystem Services
674 (LUCES). *Mitig Adapt Strateg Glob Chang* 23:211–29.

675 Swetnam RD, Fisher B, Mbilinyi BP, Munishi PKT, Willcock S, Ricketts T, Mwakalila S, Balmford A,
676 Burgess ND, Marshall AR, Lewis SL. 2011. Mapping socio-economic scenarios of land cover
677 change: A GIS method to enable ecosystem service modelling. *J Environ Manage* 92:563–74.

678 Swets J, Pickett R, Whitehead S, Getty D, Schnur J, Swets J, Freeman B. 1979. Assessment of
679 diagnostic technologies. *Science* (80-) 205:753–9.

680 Syfert MM, Smith MJ, Coomes DA, Meagher T, Roberts D. 2013. The Effects of Sampling Bias and
681 Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models.
682 Roberts DL, editor. *PLoS One* 8:e55158.

683 Turner KG, Anderson S, Gonzales-Chang M, Costanza R, Courville S, Dalgaard T, Dominati E,
684 Kubiszewski I, Ogilvy S, Porfirio L, Ratna N, Sandhu H, Sutton PC, Svenning J-C, Turner GM,
685 Varennes Y-D, Voinov A, Wratten S. 2016. A review of methods, data, and models to assess
686 changes in the value of ecosystem services from land degradation and restoration. *Ecol Modell*
687 319:190–207.

688 Verhagen W, Kukkala AS, Moilanen A, van Teeffelen AJA, Verburg PH. 2017. Use of demand for and
689 spatial flow of ecosystem services to identify priority areas. *Conserv Biol* 31:860–71.

690 Villa F, Bagstad KJ, Voigt B, Johnson GW, Portela R, Honzák M, Batker D. 2014. A methodology for
691 adaptable and robust ecosystem services assessment. *PLoS One* 9:e91001.

692 Villa F, Balbi S, Athanasiadis IN, Caracciolo C. 2017. Semantics for interoperability of distributed data
693 and models: Foundations for better-connected information. *F1000Research* 6:686.

694 Villarino SH, Studdert GA, Lattera P, Cendoya MG. 2014. Agricultural impact on soil organic carbon
695 content: Testing the IPCC carbon accounting method for evaluations at county scale. *Agric*
696 *Ecosyst Environ* 185:118–32.

697 Voinov A, Seppelt R, Reis S, Nabel JEMS, Shokravi S. 2014. Values in socio-environmental modelling:
698 Persuasion for action or excuse for inaction. *Environ Model Softw* 53:207–12.

699 Willcock S, Hooftman D, Sitas N, O'Farrell P, Hudson MD, Reyers B, Eigenbrod F, Bullock JM. 2016. Do
700 ecosystem service maps and models meet stakeholders' needs? A preliminary survey across
701 sub-Saharan Africa. *Ecosyst Serv* 18:110–7.

702 Willcock S, Martínez-López J, Hooftman DAP, Bagstad KJ, Balbi S, Marzo A, Prato C, Sciandrello S,
703 Signorello G, Voigt B, Villa F, Bullock JM, Athanasiadis IN. 2018. Machine learning for ecosystem
704 services. *Ecosyst Serv*.

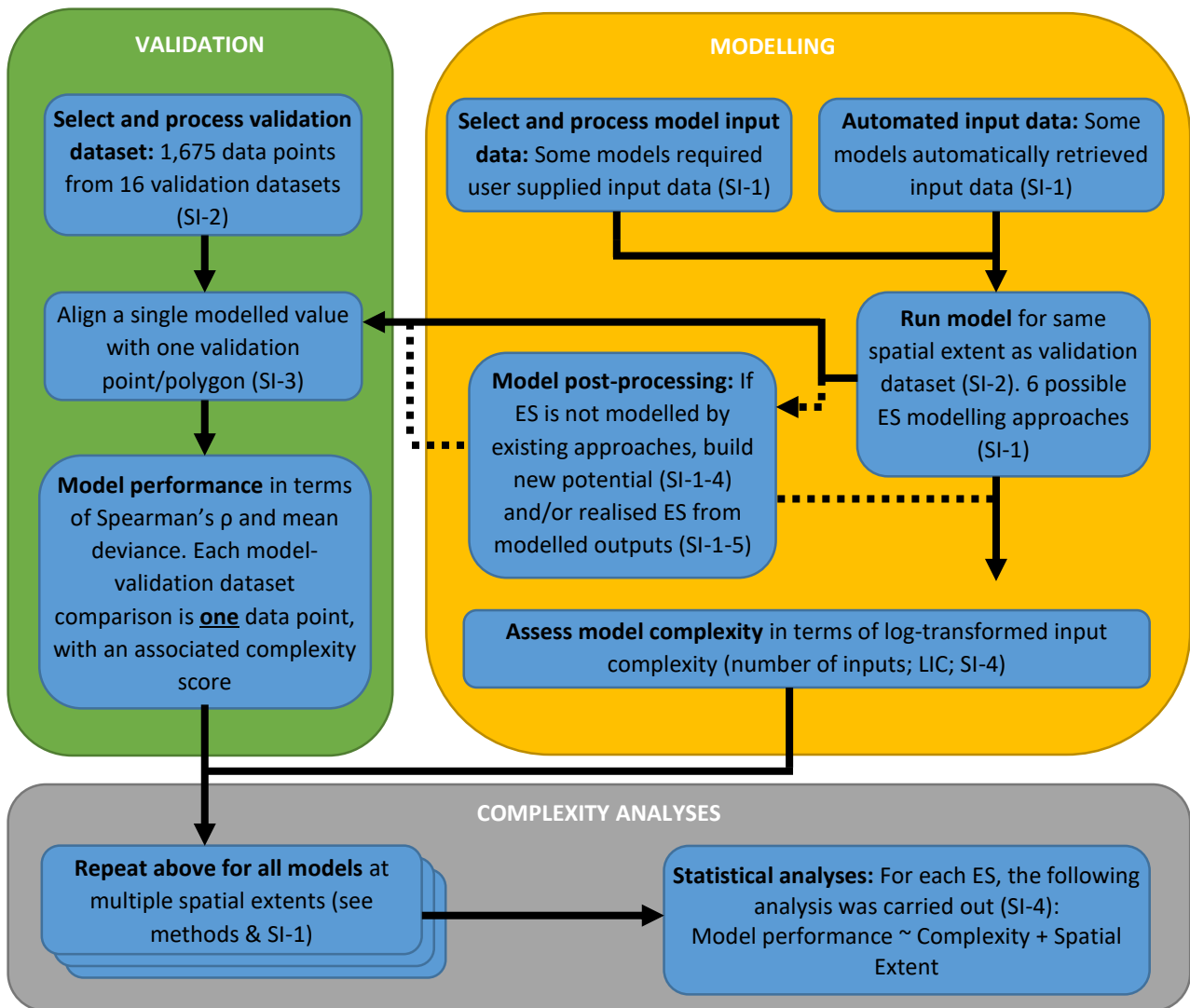
705 Willcock S, Phillips OL, Platts PJ, Balmford A, Burgess ND, Lovett JC, Ahrends A, Bayliss J, Doggart N,
706 Doody K, Fanning E, Green JM, Hall J, Howell KL, Marchant R, Marshall AR, Mbilinyi B, Munishi
707 PK, Owen N, Swetnam RD, Topp-Jorgensen EJ, Lewis SL. 2014. Quantifying and understanding
708 carbon storage and sequestration within the Eastern Arc Mountains of Tanzania, a tropical
709 biodiversity hotspot. *Carbon Balance Manag* 9:2.

710 Wong CP, Jiang B, Kinzig AP, Lee KN, Ouyang Z. 2014. Linking ecosystem characteristics to final
711 ecosystem services for public policy. Knops J, editor. *Ecol Lett* 18:108–18.

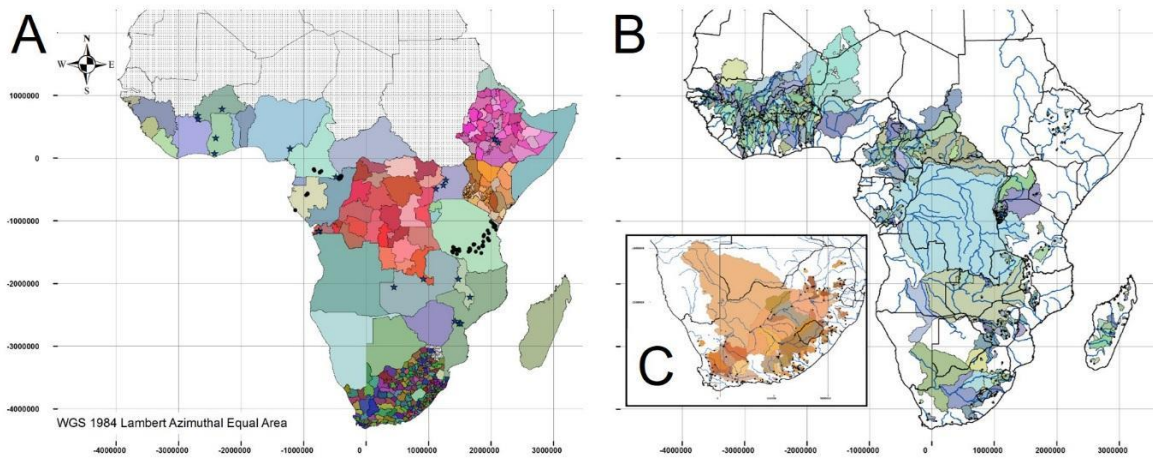
712 Zhang L, Peng J, Liu Y, Wu J. 2017. Coupling ecosystem services supply and human ecological
713 demand to identify landscape ecological security pattern: A case study in Beijing–Tianjin–Hebei
714 region, China. *Urban Ecosyst* 20:701–14. <http://link.springer.com/10.1007/s11252-016-0629-y>.
715 Last accessed 29/01/2019

716

717

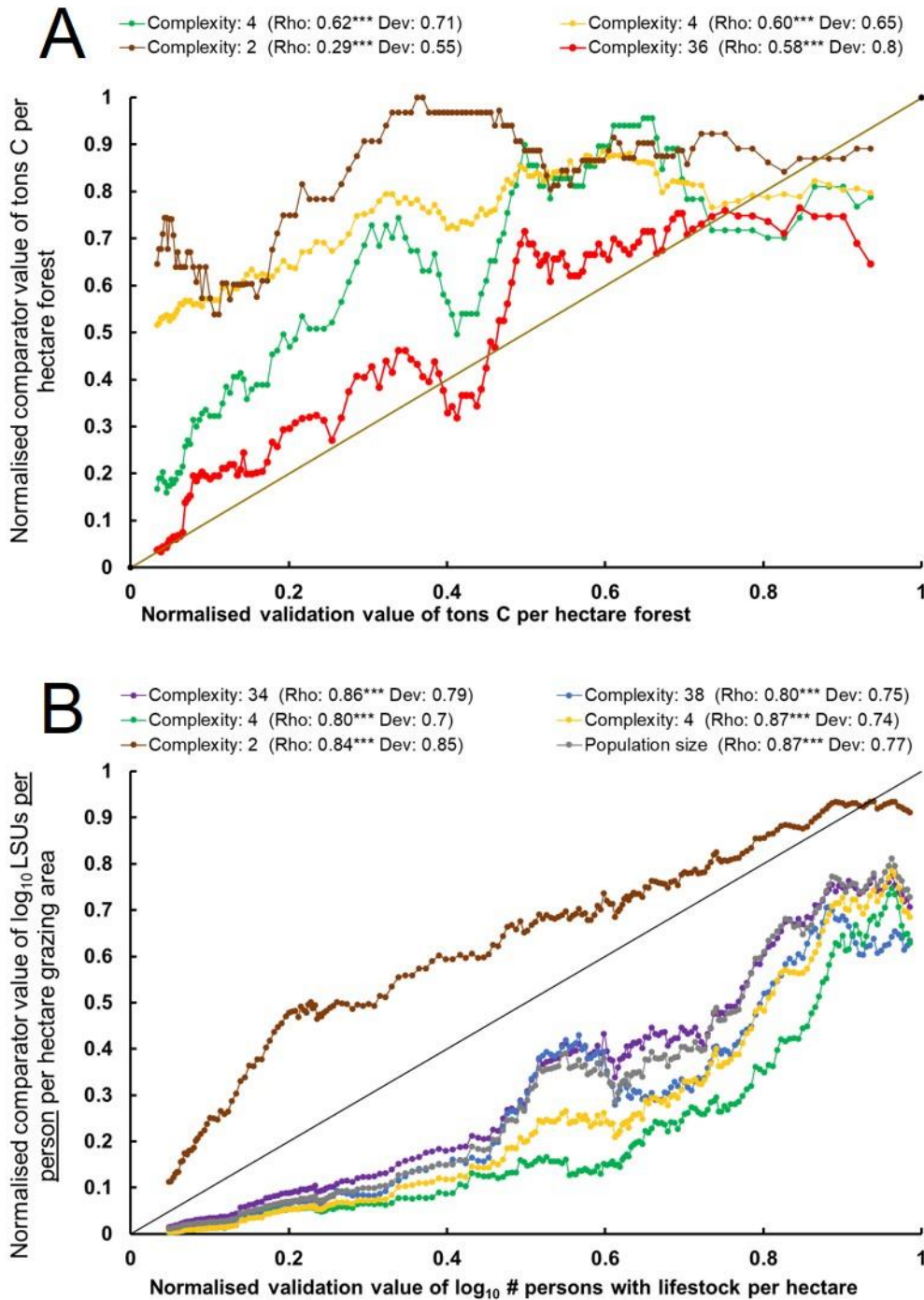


719 **Figure 1** - A summary of the analytical framework, divided into validation, modelling and analysis
 720 subsets.



722
 723 **Figure 2** – Locations at which validation datasets were gathered (SI-2). A) Coloured countries show
 724 our study area and our validation data at the country scale; dots represent standing carbon plots;
 725 stars represent PEN sites used for charcoal, firewood and grazing; districts in the Democratic
 726 Republic of the Congo are used for standing carbon; counties in Ethiopia and Kenya for grazing; and
 727 municipalities in South Africa for charcoal, firewood and grazing. B) Catchments used through the
 728 Global Runoff Data Centre managed weir dataset. C) Catchments through the South African weir
 729 data managed by the Department for Water and Sanitation. Colours in all figures are present only to
 730 allow distinction among different units within data-sets.

731



732

733 **Figure 3** – Examples of ecosystem service model validation for (A) potential biophysical carbon
 734 supply and (B) realised grazing use. X-axis is (A) tons carbon per hectare forest in ForestPlot.net
 735 (Willcock and others 2014; Avitabile and others 2016) and (B) the validation set of South African data
 736 (Hamann and others 2015), being the normalised \log_{10} number of people with livestock per hectare.
 737 Y-axis is the normalised modelled value. Different lines are different models, characterised by their
 738 complexity score. The lines are added to the graphs for visual clarity only, to allow the reader to see
 739 trends; we smoothed the lines with a 10% running average.

740 **Table 1** - Datasets used to validate the ecosystem service models included in this study. Each separate model-validation dataset comparison is a single
 741 independent data point within our analysis.

	Source	Type	Period	Selected # data-points	Unit of validation	Type of area	Service	Link to data source
1	Global Runoff Data Centre (GRDC)	Flow through Weirs	Running period of Weir	512	m ³ year ⁻¹ *	Catchments¶	Water Supply	www.bafg.de/GRDC ¶¶
2	South Africa, Department of Water and Sanitation (DWS)			188	m ³ year ⁻¹ *	Catchments¶	Water Supply	www.dwa.gov.za/Hydrology
3	Aquastat	Water use per capita	1999-2011†	36	m ³ year ⁻¹ *	Countries	Water use	www.fao.org
4	Stanford Library	Ethiopia livestock Census	1999	70	LSU's*	Zones	Grazing	library.stanford.edu
5	Kenyan Government database	Kenya livestock Census	2009	46	LSU's*	Counties	Grazing	www.opendata.go.ke/
6	South-African Statistics department (Hamann and others 2015, 2016)	South-Africa per household census data	2011	234	Usage*	Municipalities	Grazing	Hamann and others 2015
7				234	Usage*	Municipalities	Firewood use	
8	FAO databases	Animal stocks	2012	36	LSU's*	Countries	Grazing	www.fao.org/faostat
9		Carbon stock in living biomass in forest areas	2010	36	Tons above ground Carbon	Countries	Stored Woody Carbon	www.fao.org/faostat
10		Total usage per year	2013	36	Tons biomass	Countries	Charcoal use	www.fao.org/faostat
11		Total usage per year	2013	36	m ³ wood fuel	Countries	Firewood use	www.fao.org/faostat
12	Carbon stocks via Laporte and others (2008)	Carbon stock in DRC in forest areas	2007	31	Tons above ground Carbon	Districts	Stored Woody Carbon	whrc.org
13	ForestPlots.Net (Willcock and others 2014; Avitabile and others 2016)	Above ground biomass using Chave and others (2005) approximation	Last available	147	Tons above ground Carbon	Plots	Stored woody Carbon	www.forestplots.net ¶¶
14	CIFOR, Poverty	Fodder usage	2004-2008	23	Kg biomass year ⁻¹ ‡	Villages‡	Grazing	www1.cifor.org/pen ¶¶
15	Environment Network (PEN)	Charcoal production	2004-2008	23	Kg biomass year ⁻¹ ‡	Villages‡	Charcoal use	www1.cifor.org/pen ¶¶
16		Firewood production	2004-2008	23	m ³ year ⁻¹ ‡	Villages‡	Firewood use	www1.cifor.org/pen ¶¶

742 *Unit converted from original source, see SI-2; ‡ quarterly data summed into years and all families summed per village; ¶ Catchments bespoke DEM-
 743 delineated to weir points. †Latest values as was available per country in the data-base in July 2015. ¶¶ Under permission of use for this project only

745 **Table 2** - Comparison of individual ecosystem service model performance (rank correlation [ρ] or
 746 mean deviance) with that of the model ensemble and human population density. For both measures
 747 a value closer to 1 indicates better fit. n/a not applicable. p is given in superscript for significant
 748 effects (otherwise $p > 0.05$).

Ecosystem service	Response variable	Least squares mean value for the best fitting individual model	Is human population density a significantly better predictor than all individual models?
Stored carbon (biophysical supply)*	ρ	0.677	n/a
	Deviance	0.790	n/a
Water (biophysical supply)*	ρ	0.695	n/a
	Deviance	0.855	n/a
Water (use)†	ρ	0.137	Worse ^{0.0341}
	Deviance	0.800	Worse ^{<0.0014}
Charcoal (use)†	ρ	0.185	Equal
	Deviance	0.660	Equal
Firewood (use)†	ρ	0.407	Equal
	Deviance	0.731	Equal
Grazing (use)†	ρ	0.463	Equal
	Deviance	0.728	Equal

749 *Modelled using existing approaches (SI-1-1); † Modelled using our new approaches (SI-1-4, SI-1-5)

750

751 **Table 3** – The effects of variables on ecosystem service model performance, derived from generalised
752 linear models as follows: Model performance (rank correlation [ρ] or mean deviance) ~ Complexity
753 Measure + Spatial Extent. Log-transformed input complexity is the measure of complexity, and spatial
754 extent (local [L], regional [R], country [C]) are fixed effects. Additionally, whether potential ES (V1) or
755 realised ES (V2) models types performed better is indicated for ES use. 0 indicates no significant effect
756 ($p>0.05$) and n/a not applicable. Significant effects are shown with p in superscript and the direction
757 of effect as +/- for continuous variables (i.e. complexity) or in terms of differences among factor levels
758 (determined using the *lsmeans* statement in SAS 9.3).

Ecosystem Service	Response variable	Complexity (LIC)		Spatial Extent	Realised vs potential ES model performance
		Direction of effect	Coefficient		
Stored carbon (biophysical supply)*	ρ	+ ^{0.0055}	0.2089	0	n/a
	Deviance	+ ^{0.0030}	0.0979	0	n/a
Water (biophysical supply)*	ρ	+ ^{0.0332}	0.2010	n/a	n/a
	Deviance	0	n/a	n/a	n/a
Water (use)†	ρ	0	n/a	0	0
	Deviance	0	n/a	0	0
Charcoal (use)†	ρ	0	n/a	0	0
	Deviance	+ ^{0.0139}	0.0732	L>C ^{<0.0001}	V2>V1 ^{<0.0001}
Firewood (use)†	ρ	0	n/a	R>L,C ^{0.0389}	0
	Deviance	+ ^{0.0508}	0.0457	R,L>C ^{0.00331}	V2>V1 ^{<0.0001}
Grazing (use)†	ρ	0	n/a	C,R>L ^{<0.0001}	0
	Deviance	0	n/a	C,R>L ^{<0.0001}	V2>V1 ^{0.0243}

759 *Modelled using existing approaches (SI-1-1); † Modelled using our new approaches (SI-1-4, SI-1-5)

760