Title: A continental scale validation of ecosystem service models

Shortened Title: Validation of ecosystem service models

Authors: Simon Willcock*,1,2, Danny A.P. Hooftman*,3,4, Stefano Balbi5, Ryan Blanchard6, Terence P. Dawson7, Patrick J. O’Farrell6,8, Thomas Hickler9,10, Malcolm D. Hudson11, Mats Lindeskog12, Javier Martinez-Lopez5, Mark Mulligan7, Belinda Reyers13, Charlie Shackleton14, Nadia Sitas6, Ferdinando Villa5,15, Sophie M. Watts2, Felix Eigenbrod2,16 & James M. Bullock4.

* Joint first authors (contributed equally)
† Corresponding author

Contributions

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

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

Keywords: Africa; beneficiary; carbon; charcoal; complexity; firewood; grazing; natural capital; water.

Highlights:

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

Introduction

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

To meet demand for an enhanced understanding of ES flows, many spatial modelling methods and tools for mapping ES have been developed, ranging from very simple land-cover based proxies to sophisticated process-based models (IPBES 2016). Whilst a growing literature is comparing the outputs and features of the different tools (Bagstad and others 2013; Turner and others 2016), validation of these models is challenging and thus rare in the literature (Bennett and others 2013). Few studies have validated single ES models against independent datasets, and then only rarely at a larger, country scales (Mulligan and Burke 2005; Bruijnzeel and others 2011; Redhead and others 2016, 2018). Even more rare are studies that explicitly validate multiple ES models simultaneously, and these generally involve small areas at catchment scale (Sharps and others 2017). As a consequence, the uncertainties associated with most ES models and the datasets that underpin them remain largely unknown (Bryant and others 2018; van Soesbergen and Mulligan 2018). This is a particular issue as the results of local-scale validation are likely not to be transferable to new locations (Redhead and others 2016) or to the regional and national scales at which ES model outputs are most widely used (Willcock and others 2016). As a result, attempts at validation by those applying models in new settings are all the more important (Bryant and others 2018). Indeed, rescaling social-ecological
patterns and processes to different spatial resolutions and extents can induce substantial systematic bias (Grêt-Regamey and others 2014), providing challenges to decision making in situations where model results are the only source of information. Lack of proven credibility, salience and legitimacy are the major reasons for the ‘implementation gap’ between all ES research (not just ES models) and its incorporation into policy- and decision-making (Cash and others 2003; Voinov and others 2014; Wong and others 2014; Clark and others 2016).

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

The unknown credibility of ES models (Voinov and others 2014) is most pronounced where they are arguably most needed – in many developing countries, where data collection and model development efforts are least advanced (Suich and others 2015). Such ES information is important because the rural and urban poor are often the most dependent on ES (either directly or indirectly (Cumming and others 2014)), both for their livelihoods (Daw and others 2011; Suich and others 2015) and as a coping strategy for buffering shocks (Shackleton and Shackleton 2012). A major barrier to the understanding and management of these benefit flows to the poor is a lack of information on the potential supply and realised use of ES, particularly in the developing world (Wong and others 2014; Willcock and others 2016; Cruz-Garcia and others 2017). Indeed, the comparisons of ES models to primary data that do exist are all focused on potential and not realised ES (i.e. biophysical supplies only and not actual
Analyses need to be disaggregated to focus on how people use ES, from which ecosystems, and how such benefits contribute to the people’s well-being (Daw and others 2011; Bagstad and others 2014; Cruz-Garcia and others 2017).

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

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

Description of ecosystem service models. We selected ES models to test, focussing on: 1) ES models capable of estimating some of our selected potential ES (stored carbon, available water) and providing inputs to our new models of firewood, charcoal, and grazing resources within our study area; 2) the subset of these models for which adequate validation data could be identified, allowing like-for-like comparisons between modelled outputs and validation data; and 3) models representing a range of complexities from simple production functions to process-based models. As such, we used six existing ES modelling frameworks that contain one or more models meeting these criteria (Table SI-1-1, SI-1-1). We selected InVEST (Kareiva 2011; McKenzie and others 2012), CoSting Nature (Mulligan and others 2010; Mulligan 2015), WaterWorld (Mulligan 2013) and benefits transfer (based on coupling the Costanza and others (2014) values with GlobCover 2009 landcover categories; SI-1-2) due to their widespread use and global applicability (Bagstad and others 2013). We also included the well-known and partially validated (Pachzelt and others 2015) dynamic vegetation model LPJ-GUESS (Smith and
Although LPJ-GUESS is not traditionally considered an ES model and has a relatively coarse native resolution (0.5 x 0.5 degrees, but constrained mainly by the resolution of environmental input variables), it is increasingly used for ES modelling applications (including implementation within the ARIES technology (Villa and others 2014)) and it is a process-based model that gives outputs that effectively track the biophysical supply of many potential ES (Bagstad and others 2014; Lee and Lautenbach 2016). Furthermore, we included the Scholes models (comprising two grazing models and a rainfall surplus model) as it is the only large-scale ES models designed specifically for use in sub-Saharan Africa (Scholes 1998) (SI-1-3). Ideally, we would also have compared bespoke local models with local data. However, such models simply do not exist in sub-Saharan Africa in most places. Moreover, as the global models we compare run at fine spatial resolutions (except LPJ-GUESS), it is reasonable to investigate how well they perform in terms of accuracy against local data collected in many locations in many different ways (as is the case here).

At time of analysis (March 2017), InVEST, Co$ting Nature, and LPJ-Guess did not have models that focus on firewood, charcoal or grazing resources, but they did explicitly output stored vegetation carbon. As the supply of these three ES is directly dependent on the amount of biomass present, which is what underpins estimates of stored vegetation carbon in all three models, we built eight new predictors using the outputs from these three existing carbon modules (to estimate the potential supply of these three additional ES (SI-1-4). These new models used spatial masks to estimate the biomass available on relevant land uses (SI-1). For example, we applied a “grazing” spatial mask to derive grassland carbon from InVEST and Co$ting Nature standing carbon outputs. We excluded areas in which little to no grazing activity was expected (e.g. protected areas) and included areas in which most of the above-ground carbon is assumed to be available for livestock grazing (Figure SI-1-1A; Table SI-1-5). For LPJ-Guess we used C3/C4 carbon outputs as estimate for grazing resources. Thereafter, we converted grazing biomass to FAO livestock units for sub-Saharan Africa using the conversion factors from Houerou & Hoste (1977). Henceforth, we refer to these carbon-based predictors as ES models.
Finally, we created models of realised ES by weighting models of potential ES (models of biophysical supply only; e.g. the Scholes models, WaterWorld and our new models of firewood, charcoal and grazing resources) by human population (Stevens and others 2015). We also conducted like-for-like comparisons of these new models for realised use of water, firewood, charcoal and grazing resources with relative rural population data alone – the simplest possible model of ES use. We also assessed whether these new realised ES models have higher predictive power than potential ES models when compared to ES use data. We excluded urban populations for all analyses except the Poverty Environment Network usage data and water use (Table 1).

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

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

**Statistical analyses**

**Calculation of model performance**

There is no single comprehensive measure of model performance (Bennett and others 2013). Criteria commonly considered are: 1) trueness – the closeness of the agreement between the reference value and the average model value, largely affected by systematic error or bias within the model); 2) precision – the closeness of agreement between repeated model runs, largely affected by random variables or distributions that feature within the model code; and 3) accuracy – an overall summary of precision and trueness that describes the closeness of the agreement between the reference value and the values obtained from the model run(s) (IOS 1994). We focussed on accuracy and trueness here, as we only considered a single output dataset from each model (derived from a single set of parameters) and assessed these using two metrics. The first metric was the rank correlation between modelled and validation values (Spearman’s $\rho$) – a measure of accuracy ranging from -1 to 1, with 1 indicating a perfect positive correlation, 0 no correlation, and -1 a perfect negative correlation. Thus, $\rho$ is a useful metric as in many cases policy makers want to rank locations by their relative ES values (Willcock and others 2016). The second metric was the average absolute deviance of modelled values
from the 1-to-1 line representing a perfect fit of normalised model values to the normalised validation values – a measure of accuracy and trueness, as it reflects the degree to which models consistently reflect validation values (SI-2). In our normalised setting (with values inverted for consistency with $\rho$), deviance ranged from 0 (poor fit) to 1 (perfect fit). For interpretation we follow the generally used criteria employed in AUC (area under the curve) in which a result below 0.7 should be considered as likely random (Swets and others 1979; Marmion and others 2009; Hooftman and others 2016) and a value $\geq 0.7$ shows a close fit between the modelled value and the validation data. It is entirely possible for a model to have a high rank correlation value, but also have high deviance from the 1-1 line and vice versa, so the two metrics are complementary (Table 2, Figure 3). We calculated both metrics separately for every ES model for each relevant validation dataset, with the ES models run at 1 x 1 km spatial resolution in most instances, giving 100 comparisons (Carbon: 12; Water supply: 21; Water use: 18; Charcoal use: 8; Firewood use: 15; Grazing use: 26 Figure 1).

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

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

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

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

**Results**

**Model Validation.** In general, at least one model for each ES produced outputs that represented the validation data well, calculated in terms of their deviance measure and Spearman’s $\rho$, with deviance showing better fits (mean of the least squares mean [LSM] values for best fit model: $\rho = 0.43$, Deviance = 0.76; Table 2; Figure 3; SI-2). **Potential ES:** The LSM value of the response variable for the best-fit model showed that the best of the existing models of potential ES (carbon and water supply) matched the validation data well (mean LSM value for best fit potential ES models: $\rho = 0.69$, Deviance = 0.82;
Whilst still producing reasonable fit to validation data, the new models of realised ES did not show as good a fit as the models of potential ES to their respective validation data (mean LSM value for best-fit realised ES models: $\rho = 0.30$, Deviance = 0.73; Table 2; Figure 3B). When compared to realised ES data (Table 3), some (3 of 8 [38%]) of the simple models of realised ES performed better than models of ES potential, and none performed worse. However, for our models of realised charcoal, firewood and grazing services, a majority (45 of 47 [96%]) were predicted as well by human population density alone as by our models, and in two cases (4%) population density was a better predictor than our models (p-values < 0.05; Table 2). By contrast, the comparison of realised water with the water use data showed population density to be a worse predictor than our new realised ES models (6 of 6 [100%]; Table 2).

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

**Discussion**

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

Decision-makers require information on a wide range of ES and across a variety of temporal and spatial scales (Scholes and others 2013; McKenzie and others 2014; Willcock and others 2016). Meeting these needs will require a shift in the focus of most models towards understanding the beneficiaries of ES and quantifying their demand, access to and utilisation of services, as well as the consequences for well-being (Bagstad and others 2014; Poppy and others 2014). Whilst some studies (Hamann and others 2016) and models (Mulligan 2015; Suwarno and others 2018; Martínez-López and others 2019) do include the demand and use of ES, our new models of realised ES (created by weighting outputs of models of ES potential by human population) generally showed lower predictive power when compared with the ability of existing models of ES potential to predict biophysical supply. Indeed, many of our new models were unable to predict ES more accurately than human population density alone (Table 2, Table 3). This suggests that rural human population density is a good proxy for ES demand, and realised use of ES is closely aligned with demand in sub-Saharan Africa. The only exception is water use, where our new models were better predictors of realised water use than human population density (Table 2). Further combining social science theory and data to explain the social-ecological processes of ES co-production, use and well-being consequences will likely result in substantial improvements in our understanding and estimates of ES use (Bagstad and others 2014; Díaz and others 2015; Suich and others 2015; Pascual and others 2017). This is an area of active research and some modelling frameworks are beginning to address this deficiency. Socio-economic data on ES use, perceptions, and well-being contributions collected over large regions can and has been incorporated into models to address questions about the impacts of ecosystem change on the well-being of regional and socio-economic groups (Díaz and others 2015; Hamann and others 2016; Egarter Vigl and others 2017). Spatial multi-criteria analyses can be used to model how consistent
available potential ES are with local demand, highlighting trade-offs between beneficiary groups where demand varies (Martínez-López and others 2019). Furthermore, the impact of individual decision-making on ES use can be captured through the integration of agent-based models depicting human behaviour with biophysical models (Villa and others 2017; Suwarno and others 2018)). Coupling models of potential ES with models of demand to estimate realised ES will likely result in models that are more complex than existing models (Zhang and others 2017), which our findings suggest could improve accuracy, and new modelling techniques (e.g. machine learning (Willcock and others 2018)) may be needed to enable this (Bryant and others 2018).

The impact of model complexity

The effect of model complexity on the accuracy of ES results has not been investigated in detail previously. For example, a Web of Science search (20 June 2018) for ‘model’ and ‘complexity’ and ‘accuracy’ and ‘ecosystem service’ resulted in only 19 studies, few of which actually assess how ES model complexity affects accuracy. Our results suggested a tendency for ES model complexity to be correlated with increased performance (particularly for potential ES), and strong evidence that increased model complexity does not lead to worse predictions (for both potential and realised ES). However, each successive increase in complexity brings diminishing returns. For example, for each unit increase in LIC for models of stored carbon, $\rho$ increased by 0.25 and deviance by 0.10 (Table 3).

Since LIC is log-transformed, each unit increase is achieved by a ten-fold increase in inputs. Furthermore, a trade-off with benefits of additional complexity may be the feasibility of running and interpreting such models (Willcock and others 2016). Results from elsewhere in the literature are mixed, often dependent on the specific context of the comparison. For example, Villarino and others (2014) compare simple “Tier 1” carbon accounting methods with more complex “Tier 2” methods, reporting increased accuracy with model complexity. However, studies that extend this analysis to the most complex “Tier 3” methods report limits in the gains in accuracy, with intermediate “Tier 2” and complex “Tier 3” models producing similar predictions (Hill and others 2013; Willcock and others
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.

Limitations

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

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

More generally, it may be good practice to validate models against more than one dataset, as validation data have their own intrinsic inaccuracies. For example, in this study we used more than one validation dataset for each ES (Table 1). More work is required to understand how best to validate
ES models, allowing model validation to become standard practice within the ES community, increasing confidence and helping to reduce the implementation gap between ES models and policy- and decision-making (Cash and others 2003; Voinov and others 2014; Wong and others 2014; Clark and others 2016). However, there will always be financial and practical limits to model validation, especially at large scales. Collection of high quality data is challenging and expensive, and as such would require further investments; indeed, the reason ES models are used is often because of the lack of primary data.

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

**Conclusions**

Our study demonstrates the feasibility of ES model validation, even in data-deficient locations such as sub-Saharan Africa (Suich and others 2015; Willcock and others 2016). While this demonstration has been long overdue, the lack of such large-scale, multi-model validations is perhaps reflective of the
In partnership with decision-makers, the advances suggested here could help to ensure ES research continues to inform ongoing policy processes (Voinov and others 2014) (such as the IPBES, the Sustainable Development Goals and CBD Aichi targets). Our findings are of particular relevance to sub-Saharan Africa. Whilst the continent is perceived as relatively data-deficient (Suich and others 2015; Willcock and others 2016), we have shown that adequate data exist to run and validate multiple models for ES of high policy relevance (Willcock and others 2016), particularly relating to supplies of ES. Thus, ES models could help to meet the information demand from policy-makers in sub-Saharan Africa (Willcock and others 2016).

Acknowledgements

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

Compliance with Ethical Standards

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

References


Ecol Soc 19:art64.


Last accessed 29/01/2019
Figures & Tables

Figure 1 - A summary of the analytical framework, divided into validation, modelling and analysis subsets.
Figure 2 – Locations at which validation datasets were gathered (SI-2). A) Coloured countries show our study area and our validation data at the country scale; dots represent standing carbon plots; stars represent PEN sites used for charcoal, firewood and grazing; districts in the Democratic Republic of the Congo are used for standing carbon; counties in Ethiopia and Kenya for grazing; and municipalities in South Africa for charcoal, firewood and grazing. B) Catchments used through the Global Runoff Data Centre managed weir dataset. C) Catchments through the South African weir data managed by the Department for Water and Sanitation. Colours in all figures are present only to allow distinction among different units within data-sets.
Figure 3 – Examples of ecosystem service model validation for (A) potential biophysical carbon supply and (B) realised grazing use. X-axis is (A) tons carbon per hectare forest in ForestPlot.net (Willcock and others 2014; Avitabile and others 2016) and (B) the validation set of South African data (Hamann and others 2015), being the normalised log₁₀ number of people with livestock per hectare. Y-axis is the normalised modelled value. Different lines are different models, characterised by their complexity score. The lines are added to the graphs for visual clarity only, to allow the reader to see trends; we smoothed the lines with a 10% running average.
Table 1 - Datasets used to validate the ecosystem service models included in this study. Each separate model-validation dataset comparison is a single independent data point within our analysis.

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th>Period</th>
<th>Selected # data-points</th>
<th>Unit of validation</th>
<th>Type of area</th>
<th>Service</th>
<th>Link to data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Runoff Data Centre (GRDC)</td>
<td>Flow through Weirs</td>
<td>Running period of Weir</td>
<td>512</td>
<td>m³ year⁻¹*</td>
<td>Catchments*</td>
<td>Water Supply</td>
<td><a href="http://www.bafg.de/GRDC">www.bafg.de/GRDC</a>**</td>
</tr>
<tr>
<td>South Africa, Department of Water and Sanitation (DWS)</td>
<td>Water use per capita</td>
<td>1999-2011†</td>
<td>36</td>
<td>m³ year⁻¹*</td>
<td>Countries</td>
<td>Water use</td>
<td><a href="http://www.dwa.gov.za/Hydrology">www.dwa.gov.za/Hydrology</a></td>
</tr>
<tr>
<td>Aquastat</td>
<td>Ethiopia livestock Census</td>
<td></td>
<td>70</td>
<td>LSU’s*</td>
<td>Zones</td>
<td>Grazing</td>
<td>library.stanford.edu</td>
</tr>
<tr>
<td>Stanford Library</td>
<td>South-Africa per household census data</td>
<td></td>
<td>234</td>
<td>Usage*</td>
<td>Municipalities</td>
<td>Grazing</td>
<td></td>
</tr>
<tr>
<td>Kenyan Government database</td>
<td>Kenya livestock Census</td>
<td>2009</td>
<td>46</td>
<td>LSU’s*</td>
<td>Counties</td>
<td>Grazing</td>
<td><a href="http://www.opendata.go.ke/">www.opendata.go.ke/</a></td>
</tr>
<tr>
<td>South-African Statistics department (Hamann and others 2015, 2016)</td>
<td>South-Africa per household census data</td>
<td></td>
<td>234</td>
<td>Usage*</td>
<td>Municipalities</td>
<td>Firewood use</td>
<td></td>
</tr>
<tr>
<td>FAO databases</td>
<td>Animal stocks</td>
<td>2012</td>
<td>36</td>
<td>LSU’s*</td>
<td>Countries</td>
<td>Grazing</td>
<td><a href="http://www.fao.org/faostat">www.fao.org/faostat</a></td>
</tr>
<tr>
<td>Carbon stock in living biomass in forest areas</td>
<td>Carbon stock</td>
<td>2010</td>
<td>36</td>
<td>Tons above ground Carbon</td>
<td>Countries</td>
<td>Stored Woody Carbon</td>
<td><a href="http://www.fao.org/faostat">www.fao.org/faostat</a></td>
</tr>
<tr>
<td>Total usage per year</td>
<td></td>
<td>2013</td>
<td>36</td>
<td>Tons biomass</td>
<td>Countries</td>
<td>Charcoal use</td>
<td><a href="http://www.fao.org/faostat">www.fao.org/faostat</a></td>
</tr>
<tr>
<td>Total usage per year</td>
<td></td>
<td>2013</td>
<td>36</td>
<td>m³ wood fuel</td>
<td>Countries</td>
<td>Firewood use</td>
<td><a href="http://www.fao.org/faostat">www.fao.org/faostat</a></td>
</tr>
<tr>
<td>Carbon stocks via Laporte and others (2008)</td>
<td>Carbon stock in DRC in forest areas</td>
<td>2007</td>
<td>31</td>
<td>Tons above ground Carbon</td>
<td>Districts</td>
<td>Stored Woody Carbon</td>
<td>whrc.org</td>
</tr>
<tr>
<td>ForestPlots.Net (Willcock and others 2014; Avitabile and others 2016)</td>
<td>Above ground biomass using Chave and others (2005) approximation</td>
<td>Last available</td>
<td>147</td>
<td>Tons above ground Carbon</td>
<td>Plots</td>
<td>Stored woody Carbon</td>
<td><a href="http://www.forestplots.net">www.forestplots.net</a>**</td>
</tr>
<tr>
<td>CIFOR, Poverty Environment Network (PEN)</td>
<td>Fodder usage</td>
<td>2004-2008</td>
<td>23</td>
<td>Kg biomass year⁻¹‡</td>
<td>Villages‡</td>
<td>Grazing</td>
<td>www1.cifor.org/pen**</td>
</tr>
<tr>
<td>Charcoal production</td>
<td></td>
<td>2004-2008</td>
<td>23</td>
<td>Kg biomass year⁻¹‡</td>
<td>Villages‡</td>
<td>Charcoal use</td>
<td>www1.cifor.org/pen**</td>
</tr>
<tr>
<td>Firewood production</td>
<td></td>
<td>2004-2008</td>
<td>23</td>
<td>m³ year⁻¹‡</td>
<td>Villages‡</td>
<td>Firewood use</td>
<td>www1.cifor.org/pen**</td>
</tr>
</tbody>
</table>

*Unit converted from original source, see SI-2; † quarterly data summed into years and all families summed per village; ¶ Catchments bespoke DEM-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.
Table 2 - Comparison of individual ecosystem service model performance (rank correlation [$\rho$] or mean deviance) with that of the model ensemble and human population density. For both measures a value closer to 1 indicates better fit. n/a not applicable. $p$ is given in superscript for significant effects (otherwise $p>0.05$).

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

*Modelled using existing approaches (SI-1-1); † Modelled using our new approaches (SI-1-4, SI-1-5)
Table 3 – The effects of variables on ecosystem service model performance, derived from generalised linear models as follows: Model performance (rank correlation [\(\rho\]) or mean deviance) \(\sim\) Complexity Measure + Spatial Extent. Log-transformed input complexity is the measure of complexity, and spatial extent (local [L], regional [R], country [C]) are fixed effects. Additionally, whether potential ES (V1) or realised ES (V2) models types performed better is indicated for ES use. 0 indicates no significant effect \((p>0.05)\) and n/a not applicable. Significant effects are shown with \(p\) in superscript and the direction of effect as +/- for continuous variables (i.e. complexity) or in terms of differences among factor levels (determined using the \textit{lsmeans} statement in SAS 9.3).

<table>
<thead>
<tr>
<th>Ecosystem Service Variable</th>
<th>Response Variable</th>
<th>Complexity (LIC)</th>
<th>Spatial Extent</th>
<th>Realised vs potential ES model performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stored carbon (biophysical supply)*</td>
<td>(\rho) +0.0055</td>
<td>0.2089</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Water (biophysical supply)*</td>
<td>Deviance +0.0030</td>
<td>0.0979</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Water (use)†</td>
<td>(\rho) 0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Charcoal (use)†</td>
<td>(\rho) 0</td>
<td>n/a</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Firewood (use)†</td>
<td>Deviance +0.0139</td>
<td>0.0732</td>
<td>L&gt;C&lt;0.0001</td>
<td>V2&gt;V1&lt;0.0001</td>
</tr>
<tr>
<td>Grazing (use)†</td>
<td>(\rho) 0</td>
<td>n/a</td>
<td>R&gt;L,C&lt;0.0389</td>
<td>0</td>
</tr>
<tr>
<td>Deviance +0.0508</td>
<td>0.0457</td>
<td>R,L&gt;C&lt;0.0031</td>
<td>V2&gt;V1&lt;0.0001</td>
<td></td>
</tr>
</tbody>
</table>

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