Ensembles of ecosystem service models can improve accuracy and indicate uncertainty

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

Science of the Total Environment

DOI:
10.1016/j.scitotenv.2020.141006

Published: 10/12/2020

Peer reviewed version

Supplementary Information for Ensembles of ecosystem service models can improve accuracy and indicate uncertainty

Contents

SI1. Supporting tables and Figures

Table SI-1-1. Preliminary Table of Best model per combination per metric. For validation data sets see Table SI-1-2

ES	Validation data set	Best performing model		Ensemble accuracy (mean)		Ensemble accuracy (median)	
		Spearman's ρ	***Inverted Deviance ($D^{\downarrow}$)***	***ρ***	***$D^{\downarrow}$***	***ρ***	***$D^{\downarrow}$***
Water supply	Global Runoff Data Centre	Co$ting Nature	WaterWorld	0.73	0.89	0.72	0.88
Water supply	South Africa DWS	Scholes water Surplus	Scholes water Surplus	0.63	0.90	0.65	0.88
Water use	Aquastat	Benefit transfer	Benefit transfer	-0.03	0.79	0.07	0.80
Grazing use	Ethiopia livestock Census	Benefit transfer	LPJ-GUESS	0.46	0.72	0.43	0.72
Grazing use	Kenya livestock Census	New model from Co$ting Nature	New model from Co$ting Nature	0.51	0.73	0.48	0.74
Grazing use	South-Africa per household census data	New model from Co$ting Nature	Benefit transfer	0.88	0.88	0.87	0.8
Firewood use	South-Africa per household census data	New model from Co$ting Nature	Benefit transfer	0.67	0.79	0.62	0.76
Grazing use	FAO databases: Animal stocks	Scholes model	New model from Co$ting Nature	0.51	0.66	0.58	0.69
Stored Woody Carbon	FAO databases: Carbon stock in living biomass in forest areas	LPJ-GUESS	InVEST	0.60	0.83	0.61	0.81
Charcoal use	FAO databases: Total Usage	Benefit transfer	LPJ-GUESS	0.39	0.64	0.39	0.62

	per year						
Firewood use	FAO databases: Total Usage per year	Benefit transfer	New model from Co$ting Nature	0.58	0.76	0.60	0.76
Stored carbon	Carbon stock in DRC in forest areas	Co$ting Nature	LPJ-GUESS	0.74	0.74	0.74	0.74
Stored carbon	ForestPlots.Net	InVEST	LPJ-GUESS	0.67	0.69	0.68	0.72
Grazing use	Poverty Environment Network: Fodder usage	Scholes model & Benefit transfer	Scholes model	-0.28	0.51	-0.31	0.47
Charcoal use	Poverty Environment Network: Charcoal production	InVEST	New model from Co$ting Nature	0.01	0.72	0.06	0.74
Firewood use	Poverty Environment Network: Firewood production	Scholes Firewood model	LPJ-GUESS	-0.03	0.75	-0.04	0.76

Table SI1-2. Datasets used to validate the ecosystem service models included in this study, taken from Willcock et al. (2019). Each separate model-validation dataset comparison is a single independent data point within our analysis. The location and geographical spread of these data are shown in

*Unit converted from original source, see SI-3; ‡ quarterly data summed into years and all families summed per village; ¶ Catchments delineated to weirs locations. †Latest values that were available per country in the database in July 2015. ¶¶ Under permission of use for this project only

	Source	Type	Period	Selected # data-points	Unit of validation	Type of area	Service	Link to data source
1	Global Runoff Data Centre	Flow through Weirs	Running period of Weir	512	m^3 $year^{-1}$*	Catchments¶	Water Supply	www.bafg.de/GRDC¶¶
2	South Africa 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 Stats department (Hamann et al., 2016, 2015)	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 (Avitabile et al., 2016; Willcock et al., 2014)	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 Environment Network	Fodder usage	2004-2008	23	Kg biomass $year^{-1}$‡	Villages‡	Grazing	www1.cifor.org/pen¶¶
15		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¶¶

Table SI1-3. Mean proportional increases (standard error of the mean) for each one-by-one comparison among ensemble types for 16 pairwise validation sets, including the individual models as a group. The first column corresponds to Fig 2. Student's T-test statistics with df = 15: *: P< 0.05; ** P<0.01; *** P<0.001; †reported in SI3, not in main text.

Ensemble method	Models	Mean	Median	Deviance weighted†	Rho weighted†
Spearman's *ρ*					
Models ρ = 0.39 (0.04)	-				
Mean	5.03% (2.2)*	-			
Median	5.75% (2.4)*	0.64% (0.9)	-		
Deviance weighted†	7.02% (1.9)**	2.04% (0.8)*	1.49% (1.2)	-	
Best model	14.3% (3.3)***	12.6% (5.5)*	11.9% (5.6)*	10.2% (5.0)$^{P=0.06}$	-
Inverse Deviance ($D^{\downarrow}$)					
Models $D^{\downarrow}$ = 0.72 (0.01)	-				
Mean	6.09% (2.0)**	-			
Median	4.99% (1.6)**	-0.87% (1.0)	-		
Rho weighted†	6.97% (2.0)**	0.87% (0.6)	1.86% (0.8)*	-1.17% (0.6)$^{P=0.07}$	-
Best model	11.3% (1.5)***	6.29% (1.7)**	7.25% (1.5)***	-	5.34% (1.4)**

Table SI1-4. An alternative analysis to Table SI-1-3 for mean proportional increases (standard error of the mean) in which individual model improvements per validation dataset are not subject averaging. As a consequence, compared to Table-SI-1-3 the degrees of freedom for the Student T-test are expanded to N = 79 -1 = 78. The results are nearly identical in improvement ratios, SEM-variance and t-test statistics as those reported in the main text and Table-SI-1-3. *: P< 0.05; ** P<0.01; *** P<0.001; †reported in SI3, not in main text.

Ensemble method	Models
Spearman's *ρ*	
Models ρ = 0.39 (0.04)	-
Mean	5.37% (2.2)*
Median	6.16% (2.2)**
Complexity weighted†	5.93% (2.2)**
Deviance weighted†	7.49% (2.1)***
Best model	17.3% (2.8)***
Inverse Deviance ($D^{\downarrow}$)	
Models $D^{\downarrow}$ = 0.72 (0.01)	-
Mean	6.79% (1.6)***
Median	5.49% (1.5)***
Complexity weighted†	6.87% (1.7)***
Rho weighted†	7.60% (1.6)***
Best model	12.7% (1.6)***

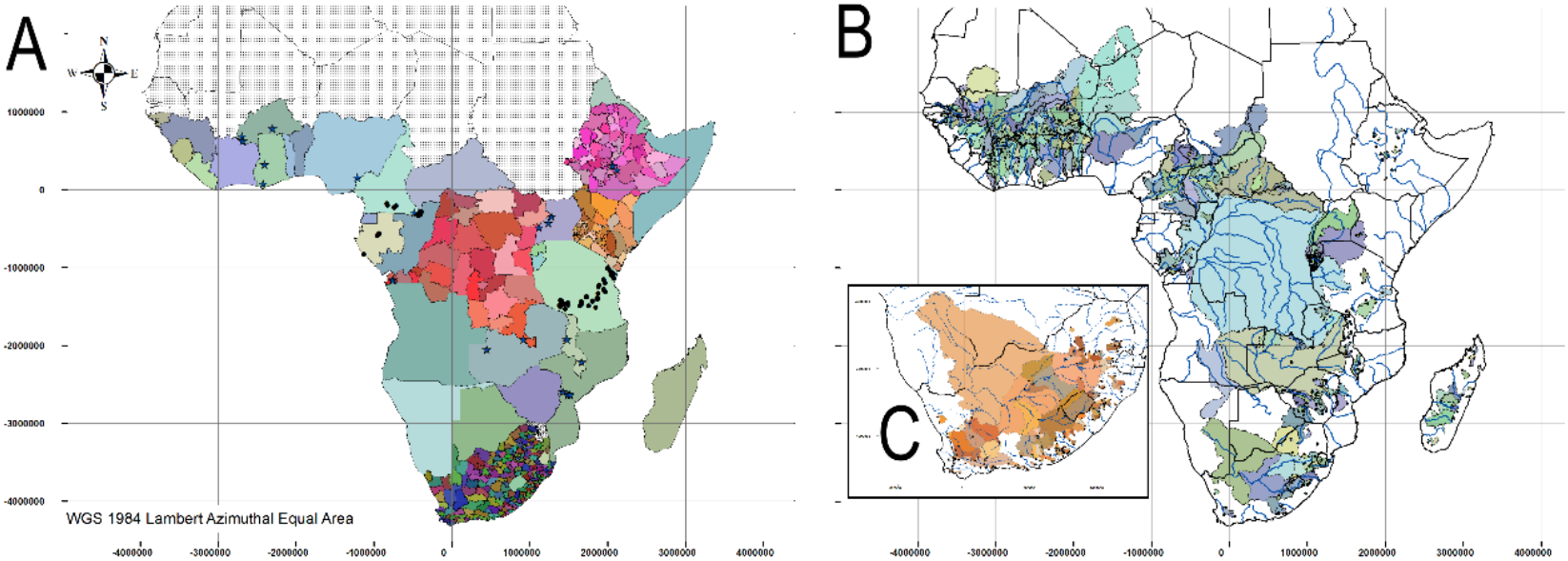

Figure SI1-1. Location of validation sets, taken from Willcock *et al.* (2019). A) Coloured countries show our study area and our validation data at the country scale; dots represent standing carbon plots; stars represent PEN sites used for charcoal, firewood and grazing; districts in DRC are used for standing carbon; counties in Ethiopia and Kenya for grazing; and municipalities in South-Africa for charcoal, firewood and grazing. B) Catchments used through the GRDC managed weir dataset. C) Catchments through the South-African weir data managed by the Department for Water and Sanitation.

SI2 – Hotspots

Introduction

Hotspot analyses can be used to support more robust decisions (Egoh et al., 2009). A 'hotspot' defines an area where high values of the spatial estimates of a number of variables (e.g. species occurrence, biodiversity, water availability) overlap and so synergies covering multiple benefits are suggested. For example, a single modelling framework (e.g. InVEST) could be used to identify areas important for multiple ES to support sustainable management decisions (Díaz et al., 2018). Using multiple models to identify hotspots increases robustness by reducing the likelihood of additional evidence altering the positions of hotspots and so the decisions that might be made. For example, Schulp et al (2014) used four models to estimate the provision of five ES across Europe, identifying hotspots and coldspots across each ES.

Methods

We graphically depicted relative hotspots and coldspots, where coldspots are co-occurring low values of multiple services. We divided all modelled areas into km^2 gridcells – except water, which is in m^3 ha^{-1} per polygon – counting the number of models that correspond to quartiles of their normalised range of ES delivery. We defined the upper quartile of each individual model as 'hot' and the lower quartile as 'cold', with the middle quartiles being 'in between' – simplifying individual model outputs into three categories. Depending on the ES, 4-6 individual models were available, with some ES models being spatially restricted (masked) into by forest, grazing or woody habitats as performed in Willcock *et al.* (2019). The few areas with missing values (i.e. where <3 model estimates were available) were marked as missing. For each ES, the individual model quartiles were combined into a graphical ensemble for every km^2-grid cell. Each grid cell was assigned as 'hot spots', 'cold spots' or 'in-between' when all individual models agreed in quartile assignment or all-individual-models-minus-one as one single model disagreed in quartile assignment. If not a gridcell was labelled as 'disagreement'.

Results

For sub-Saharan Africa, we found large areas for which most individual models agree, having the same quartile categories. Figure SI2-1 shows ensemble-wide areas important for ES delivery (hotspots - red), as well as less important areas (coldspots – blue). Hotspots indicate areas which should be high-priority for decision makers (i.e. to ensure sustainable management).

Areas of disagreement (black) within ensembles of models might be of interest for researchers in order to improve the next generation of ES models. Disagreement is foremost among water models and in the non-dense forest areas for above ground carbon storage. Firewood and Charcoal have four models in common that are equal at normalised scales. However, Firewood contains an additional bespoke Firewood model (Willcock et al. (2019)) that generates more disagreement areas.

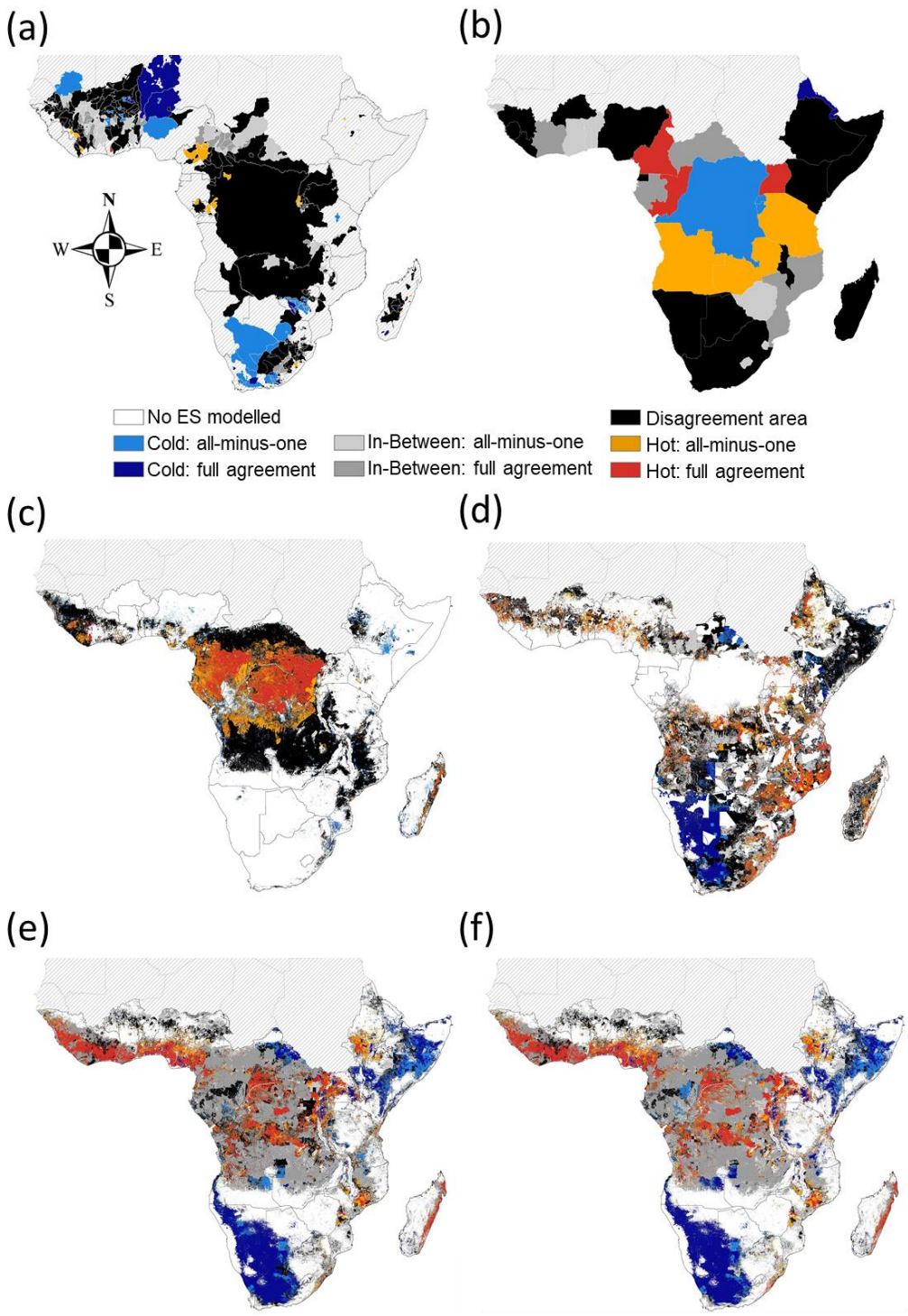

Figure SI-2-1. Among-model agreement levels for hot (upper quartile) and cold (lower quartile) areas of ES delivery, based on normalised model data. Red, grey and blue colours represent agreement among all models and all-models-minus-one (lighter tone), including agreement for mid-levels (middle quartiles; i.e. 'in-between'; grey). Black areas show disagreement among models. White areas were not modelled. A) Water supply per hectare catchment (6 models); B) water usage (6 models) per hectare per country; C) Carbon storage in forest vegetation (4 models); D) Grazing (6 models); E) Firewood usage (5 models); F) Charcoal (4 models).

SI3 – Alternative ensemble techniques

Introduction

In some situations, certain models or types of models may be perceived as more reliable, and ensemble averages can then be weighted by some measure of reliability. This may seem counter-intuitive – if the most accurate model is known, shouldn't that model be used in isolation? The answer to this question depends on whether validation data exists to support the application in question. For example, one might wish to model ES in the present day, using validation data to identify the most accurate model (e.g. Willcock et al (2019)) and using that model to support near real-time decision-making. However, scenarios are often used to support decisions (Willcock et al., 2016). Scenarios, by their very definition, are used to represent possible futures and, as such, cannot be validated in the present day. Thus, when running scenarios, it may be preferable to validate the models against the present day, before extrapolating the results into the future. But there is no guarantee that the most accurate model for present day estimates will continue to be the most accurate model in the future, so ensemble modelling is sensible. The scenarios produced could be calculated via ensemble averages weighted by model accuracy (in the present day). Furthermore, if validation has taken place in one geographic region, it may be risky to assume the most accurate model in that region is so in other parts of the World. This protocol has the advantage of giving most weight to the model currently thought to be the most accurate, but guards against any changes to this relationship over time or space.

Methods

We complemented our committee averaging (mean [Eqn. SI1] and median [Eqn. SI2]) by creating ensembles using methods requiring validation data. Specifically, we incorporated accuracy information through weighting by ρ (Eqn. SI3) and $D^{\downarrow}$ (Eqn. SI4), using values from Willcock et al (2019) (Figure 1).

Equation SI1: $E_{mn(x)} = (\overline{X_i})_{(x)}$ Mean of all models (i) for 1 km^2 grid-cell *x*

Equation SI2: $E_{md(x)} = (\widetilde{X_i})_{(x)}$ Median of models (i) for 1 km^2 grid-cell *x*

Equation SI3: $E_y = \left(\frac{\sum_i (Y_i \times \rho_i)}{\sum_i \rho_i}\right)_y$ Weighting by accuracy, ρ, where y represents the geographic locations where validation data were

Equation SI4: $E_y = \left(\frac{\sum_i (Y_i \times D\downarrow_i)}{\sum_i D\downarrow_i}\right)_y$ Weighting by accuracy, $D^{\downarrow}$

To evaluate ensemble accuracy, we compared the ensemble estimate (E) to the validation data from Willcock et al (2019) as described in the main text.

Results

We hypothesised that weighting ensembles by model accuracy would improve the accuracy of the ensemble further. However, weighting by ρ could improve predictive quality as measured by $D^{\downarrow}$ (to avoid circularity) and *vice* versa from the median (Figure SI3-1; Table SI-New2). Thus, including validation evidence when creating the ensemble improves the likelihood of a higher accuracy but not with a large improvement.

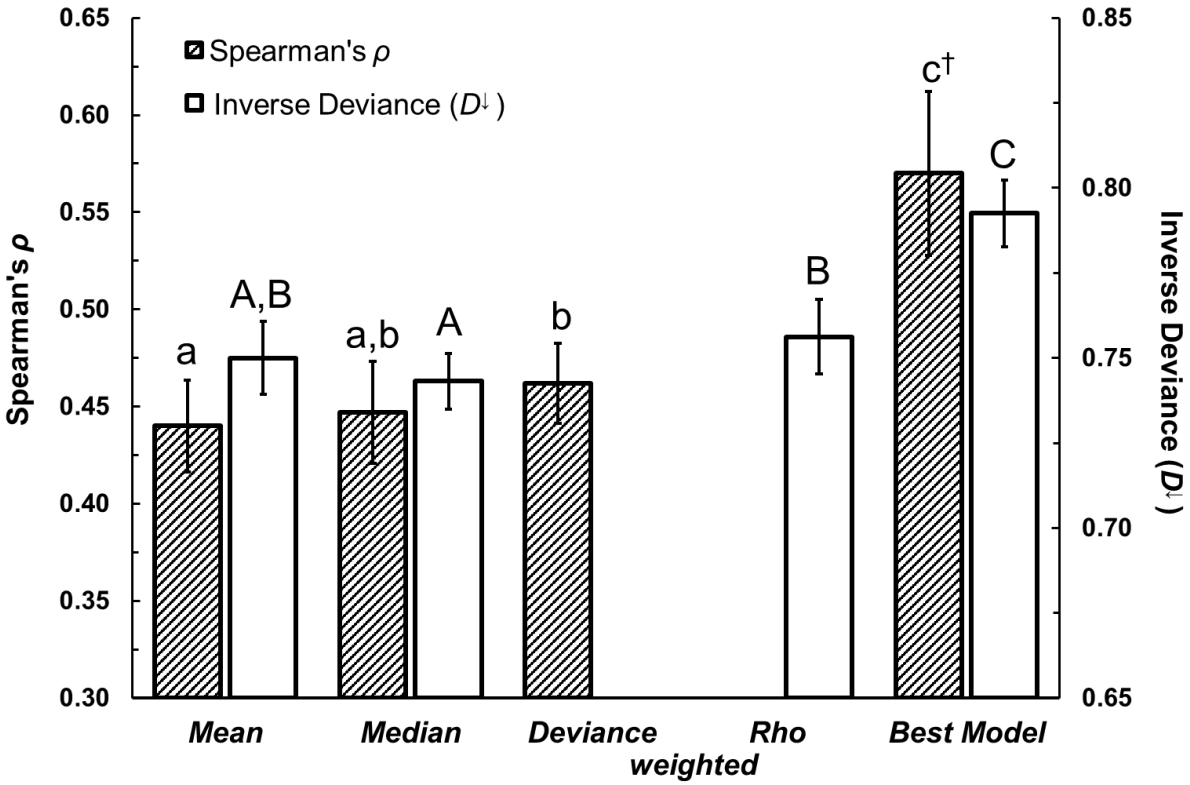

Figure SI3-1. Mean ρ and $D^{\downarrow}$ of the mean and median ensembles, accuracy weighted ensembles and best-fit individual model. Dark bars = Spearman's ρ; Light bars = Inverse Deviance $D^{\downarrow}$. Circular weighting is excluded. Error bars indicate variation in proportional improvement against the individual models, calculated as $SEM_{imp} = CV_{imp}$ x absolute difference, with CV the coefficient of variation of proportional improvement based on standard error of the mean (SEM). Thus, error bars indicate the variation in improvement against individual models as a group to highlight the range of improvement of ensemble techniques. N = 16 per bar. Significance indicating lettering (α <0.05) are based on proportional differences among single categories comparisons (Table SI-New2; capitals are for $D^{\downarrow}$). ‡ marginally significant different from deviance weighted ensemble (P=0.06).

Discussion

We show that ensembles weighted using accuracy data are more accurate than those with no *a priori* information (e.g. committee averages), with weighted ensembles being 7.0% more accurate than individual models for both output statistics. Thus, weighting by goodness of fit might be preferred to committee averaging when similar data are available (Refsgaard et al., 2014), but validation of the exact desired model output is not possible. This might be the case when estimating ES in a data poor area immediate adjacent to a data rich one, or when running future scenarios in data rich locations as performance in one geographic space/time may not be indicative of performance elsewhere.

SI References

Avitabile, V., Herold, M., Heuvelink, G.B.M., Lewis, S.L., Phillips, O.L., Asner, G.P., Armston, J., Ashton, P.S., Banin, L., Bayol, N., Berry, N.J., Boeckx, P., de Jong, B.H.J., Devries, B., Girardin, C.A.J., Kearsley, E., Lindsell, J.A., Lopez-Gonzalez, G., Lucas, R., Malhi, Y., Morel, A., Mitchard, E.T.A., Nagy, L., Qie, L., Quinones, M.J., Ryan, C.M., Ferry, S.J.W., Sunderland, T., Laurin, G.V., Gatti, R.C., Valentini, R., Verbeeck, H., Wijaya, A., Willcock, S., 2016. An integrated pan-tropical biomass map using multiple reference datasets. Glob. Chang. Biol. 22, 1406–1420. https://doi.org/10.1111/gcb.13139

Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chambers, J.Q., Eamus, D., Folster, H., Fromard, F., Higuchi, N., Kira, T., Lescure, J.-P.P., Nelson, B.W., Ogawa, H., Puig, H., Riera, B., Yamakura, T., Fölster, H., Riéra, B., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145, 87–99. https://doi.org/DOI 10.1007/s00442-005-0100-x

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

Egoh, B., Reyers, B., Rouget, M., Bode, M., Richardson, D., 2009. Spatial congruence between biodiversity and ecosystem services in South Africa. Biol. Conserv. 142, 553–562. https://doi.org/10.1016/j.biocon.2008.11.009

Hamann, M., Biggs, R., Reyers, B., 2015. Mapping social–ecological systems: Identifying 'green-loop' and 'red-loop' dynamics based on characteristic bundles of ecosystem service use. Glob. Environ. Chang. 34, 218–226. https://doi.org/10.1016/j.gloenvcha.2015.07.008

Hamann, M., Biggs, R., Reyers, B., Pomeroy, R., Abunge, C., Galafassi, D., 2016. An Exploration of Human Well-Being Bundles as Identifiers of Ecosystem Service Use Patterns. PLoS One 11, e0163476. https://doi.org/10.1371/journal.pone.0163476

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

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

Schulp, C.J.E., Burkhard, B., Maes, J., Van Vliet, J., Verburg, P.H., 2014. Uncertainties in Ecosystem Service Maps: A Comparison on the European Scale. PLoS One 9, e109643. https://doi.org/10.1371/journal.pone.0109643

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

Willcock, S., Hooftman, D.A.P., Balbi, S., Blanchard, R., Dawson, T.P., O'Farrell, P.J., Hickler, T., Hudson, M.D., Lindeskog, M., Martinez-Lopez, J., Mulligan, M., Reyers, B., Shackleton, C., Sitas, N., Villa, F., Watts, S.M., Eigenbrod, F., Bullock, J.M., 2019. A Continental-Scale Validation of Ecosystem Service Models. Ecosystems 22, 1902–1917. https://doi.org/10.1007/s10021-019-

00380-y

Willcock, S., Phillips, O.L., Platts, P.J., Balmford, A., Burgess, N.D., Lovett, J.C., Ahrends, A., Bayliss, J., Doggart, N., Doody, K., Fanning, E., Green, J.M., Hall, J., Howell, K.L., Marchant, R., Marshall, A.R., Mbilinyi, B., Munishi, P.K., Owen, N., Swetnam, R.D., Topp-Jorgensen, E.J., Lewis, S.L., 2014. Quantifying and understanding carbon storage and sequestration within the Eastern Arc Mountains of Tanzania, a tropical biodiversity hotspot. Carbon Balance Manag. 9, 2. https://doi.org/10.1186/1750-0680-9-2