Hooftman, Danny A.P.; Bullock, James M.; Jones, Laurence; Eigenbrod, Felix; Barredo, José I.; Forrest, Matthew; Kindermann, Georg; Thomas, Amy; Willcock, Simon. 2022. Reducing uncertainty in ecosystem service modelling through weighted ensembles.

© 2020 Elsevier B.V.
This manuscript version is made available under the CC BY-NC-ND 4.0 license
https://creativecommons.org/licenses/by-nc-nd/4.0/

This version is available at http://nora.nerc.ac.uk/id/eprint/532479

Copyright and other rights for material on this site are retained by the rights owners. Users should read the terms and conditions of use of this material at https://nora.nerc.ac.uk/policies.html#access.

This is an unedited manuscript accepted for publication, incorporating any revisions agreed during the peer review process. There may be differences between this and the publisher’s version. You are advised to consult the publisher’s version if you wish to cite from this article.

The definitive version was published in Ecosystem Services, 53, 101398. https://doi.org/10.1016/j.ecoser.2021.101398

The definitive version is available at https://www.elsevier.com/

Contact UKCEH NORA team at noraceh@ceh.ac.uk

The NERC and UKCEH trademarks and logos (‘the Trademarks’) are registered trademarks of NERC and UKCEH in the UK and other countries, and may not be used without the prior written consent of the Trademark owner.
Reducing Uncertainty in Ecosystem Service Modelling through Weighted Ensembles

Hooftman, Danny A.P.\textsuperscript{1,2}, James M. Bullock\textsuperscript{2}, Laurence Jones\textsuperscript{3}, Felix Eigenbrod\textsuperscript{4}, José I. Barredo\textsuperscript{5}, Matthew Forrest\textsuperscript{6}, Georg Kindermann\textsuperscript{7}, Amy Thomas\textsuperscript{3} & Simon Willcock\textsuperscript{8,9*}

\* Corresponding author

Affiliations:

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

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

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

Highlights:

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

Abstract: (150 words)
Over the last decade many ecosystem service (ES) models have been developed to inform sustainable land and water use planning. However, uncertainty in the predictions of any single model in any specific situation can undermine their utility for decision-making. One solution is creating ensemble predictions, which potentially increase accuracy, but how best to create ES ensembles to reduce uncertainty is unknown and untested. Using ten models for carbon storage and nine for water supply, we tested a series of ensemble approaches against measured validation data in the UK. Ensembles had at minimum a 5-17% higher accuracy than a randomly selected individual model and, in general, ensembles weighted for among model consensus provided better predictions than unweighted ensembles. To support robust decision-making for sustainable development and reducing uncertainty around these decisions, our analysis suggests various ensemble methods should be applied depending on data quality, for example if validation data are available.

Graphical Abstract:

<table>
<thead>
<tr>
<th>Accuracy compared to mean ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Models</td>
</tr>
<tr>
<td>Unweighted Ensembles</td>
</tr>
<tr>
<td>Mean Ensemble</td>
</tr>
<tr>
<td>Median Ensemble</td>
</tr>
<tr>
<td>Untrained Weighted Ensembles</td>
</tr>
<tr>
<td>Deterministic Consensus</td>
</tr>
<tr>
<td>Iterated Consensus</td>
</tr>
<tr>
<td>Attribute based</td>
</tr>
<tr>
<td>Trained Weighted Ensembles</td>
</tr>
<tr>
<td>Accuracy weighted</td>
</tr>
<tr>
<td>Regressed consensus</td>
</tr>
</tbody>
</table>

Keywords: Carbon; Committee averaging; Prediction Error; Accuracy; United Kingdom; Validation; Water supply; Weighted averaging

Video Summary: (see attached file)
1. Introduction

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

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

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

Across the sciences, one solution to address uncertainty surrounding the accuracy of any single model is to use an ensemble of models (Araújo & New 2007; Willcock et al. 2020) – using individual models as replicates with different input parameters and boundary conditions (Araújo & New 2007; Dormann et al. 2018). Variation among models in the assumptions and formats can result in large differences in predictions, in terms of predicted values and how they vary over space, especially when there is uncertainty as to the state and processes of the system being modelled (van Soesbergen & Mulligan 2018; Willcock et al. 2019). Ensembles of models are hypothesised to have enhanced accuracy over individual models due to fewer overall errors in prediction by reducing the influence of idiosyncratic outcomes from single models (Araújo & New 2007; Dormann et al. 2018). Individual models rarely capture all potentially relevant processes or are often tuned to particular ecosystem characteristics. A combination of models might provide a more comprehensive coverage of processes and their forms, and avoids the chance of (unknowingly) selecting a model with a high prediction error at the location and scale of interest for a particular study (Willcock et al. 2020).
Model ensembles are common in other disciplines – e.g. in niche modelling (Araújo & New 2007, Grenouillet et al. 2011), agroecology (Refsgaard et al. 2014), hydrology and water resources management (Wang et al. 2019; He et al. 2021), and climate and weather modelling (Knutti et al. 2013), as well as market forecasting (He et al. 2012). However, ensembles have been largely neglected in ES studies (Bryant et al. 2018). The only current exception is the simplest ensemble approach (i.e. ‘committee averaging’ – taking the unweighted mean of a group of individual models per location –) which was applied to ES models in Sub-Saharan Africa, and gave higher accuracy in terms of fit to validation data (Willcock et al. 2020).

Approaches that use more information might yield even more accurate estimates. Thus, here we explore the outstanding question of “what are the best ways to build ES model ensembles to realise the benefits such ensembles can bring to sustainability science?”

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

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

However for ES, the lack of a priori validation data in many cases means that the distributions of accuracy among ES models are unknown. Furthermore, given that inferences about model accuracy at one location may not be transferable to others (Willcock et al. 2019), weighting using validation results from a separate study may not improve outcomes. Therefore where validation data are not available, the consensus among models could be used to weight their individual contribution to the ensemble value (Marmion et al. 2009; Grenouillet et al. 2011). This approach follows the logic that models whose output values are more different to those of the other models (i.e. are more distinct) are more likely to be incorrect. Therefore, weighting by consensus reduces the impact of outputs from more idiosyncratic models (i.e. those with extreme values, outliers or badly comparable processes) by comparison with the other models (Araújo & New 2007; Dormann et al. 2018), but does not exclude their information fully. The opposite may also be true – i.e. more distinct models are more accurate – for example in cases where more similar models have common inaccuracies.
Here, we implement 10 alternative ensemble methods, restricting ourselves to methods feasible for a wide range of users, to evaluate whether weighting provides higher accuracy and if so which type of method produces the most accurate predictions against validation data. We focus on two services, water supply and carbon storage, in the United Kingdom. To support decision-making, we map the results for potential further use, which are available via https://doi.org/10.5285/a9ae773d-b742-4d42-ae42-2b594bae5d38. We use post-processing – specifically normalisation and per area correction – developed in earlier work (Willcock et al. 2019; Willcock et al. 2020) to make outputs among models comparable.

2. Methods

We developed and validated unweighted average and weighted average ensembles of models for a provisioning service (water supply; subsequently referred to as ‘water’) and a regulating service (aboveground carbon storage; subsequently referred to as ‘carbon’), for which there is a large quantity of reliable validation data; allowing us to assess ensemble accuracies. We compared accuracy (i.e. fit to validation data) of these individual models with those of the ensembles generated from them via multiple approaches, assessed if weighted ensembles were an improvement on the unweighted mean-averaged ensemble, and identified the methods of weighting ensembles that gave the highest accuracy.

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

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>Description</th>
<th>Details &amp; Matlab Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Accuracy approaches</td>
<td>Correlation coefficient between ranked variables $V$ and $T$.</td>
<td>$T$ is either $Y$ or $E$, depending on ensemble method</td>
</tr>
<tr>
<td>• Spearman $\rho$</td>
<td>$D^4 = 1 - \left( \frac{1}{m} \sum_{x=1}^{m}</td>
<td>y(x) - T(x)</td>
</tr>
<tr>
<td>• Inverse Deviance ($D^4$)</td>
<td>$T(x)$ is either $Y_{i(x)}$ or $E_{i(x)}$</td>
<td></td>
</tr>
<tr>
<td>B. Ensemble approaches</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unweighted Averaging:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>En-1. Mean</td>
<td>$E_{(x)} = \langle Y_{i(x)} \rangle_{(x)}$</td>
<td></td>
</tr>
</tbody>
</table>
### Deterministic consensus

**En-2. Median**

\[ E(x) = \left( \bar{Y}_i \right)_{(x)} \]

Hypothesised to perform better than mean for skewed distributions.

**Untrained Weighted Ensembles:**

\[ E(x) = \sum_i^n \left( \frac{\omega_i}{\sum_j \omega_j} \times Y_j \right)_{(x)} \]

with \( \omega_i \) following:

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Formula</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-3. PCA</td>
<td>( \omega_i = \text{loadings of first Principal Component axis} )</td>
<td></td>
<td>Princomp-tool</td>
</tr>
<tr>
<td>En-4. Correlation coefficients</td>
<td>( \omega_i = \frac{1}{n} \times \sum_j c_{(i,j)} ) for all ( j \in i ) with ( c_{(i,j)} = \frac{1}{m-1} \times \sum^n_x \left( (Y_{(i,x)} - \bar{Y}<em>i) \times (Y</em>{(j,x)} - \bar{Y}_j) \right) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Iterated consensus**

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Formula</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-5. Regression to the median</td>
<td>( Y_{(x)} \sim \left( \frac{\sum_i^n \omega_i Y_i}{(x)} \right) )</td>
<td>nlmefit-tool, maximising Log Likelihood</td>
<td></td>
</tr>
<tr>
<td>En-6. Exhaustive leave-one-out cross-validation</td>
<td>( Y_{(x)} \sim \sum_i^n \omega_i Y_i ) for all ( j \in i ) subsequently: ( \omega_i = \frac{1}{n} \times \sum_i^n \left( \frac{1}{m-1} \times \sum_i \omega_{ij} \right) )</td>
<td>nlmefit-tool, maximising Log Likelihood</td>
<td></td>
</tr>
</tbody>
</table>

**Attribute-based**

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Formula</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-7. Upweighted finer spatial resolution</td>
<td>( \omega_i = \frac{1}{\log_{10}(\text{spatial resolution})} )</td>
<td>Finer spatial resolution: smaller grid size in 1-dimensional meters (e.g. 25 m)</td>
<td></td>
</tr>
<tr>
<td>En-8. Attribute weighting: distinctiveness</td>
<td>( \omega_i = \left( \frac{m^g}{n} \right) ) when upweighted with ( n^g = i \in g )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \omega_i = \left( \frac{n}{m^g} \right) ) when downweighted with ( n^g = i \in g )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Trained Weighted Ensembles: o-transfer via jack-knife training**

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Formula</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-9. Accuracy-weighted</td>
<td>( \omega_i = A_i ) with ( A_i = \frac{V(Y_{(x)},Y_{(y=x)})}{V(Y_{(y=x)},Y_{(y=x)})} )</td>
<td>With ( A_i ), either Spearman ( \rho ) or ( D ) accuracy</td>
<td></td>
</tr>
<tr>
<td>En-10. Log-likelihood regressions</td>
<td>( V(Y_{(x,y)}) \sim \left( \sum_i^n \omega_i Y_i \right)_{(y=x)} )</td>
<td>Using nlmefit-tool, maximising Log Likelihood</td>
<td></td>
</tr>
</tbody>
</table>

---

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Grid size (spatial resolution)</th>
<th>Model Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>InvEst v3.7.0&lt;sup&gt;11&lt;/sup&gt; (Integrated Valuation of Ecosystem Services and Trade-offs)</td>
<td>Carbon module: above ground stocks</td>
<td>25 × 25 meters</td>
<td>Look-up table</td>
</tr>
<tr>
<td></td>
<td>Water yield module: run-off per cell</td>
<td></td>
<td>Process</td>
</tr>
<tr>
<td>LPJ-GUESS&lt;sup&gt;2,3†&lt;/sup&gt; (Lund-Potsdam-Jena General Ecosystem Simulator)</td>
<td>Vegetation biomass stocks per cell, mean for years 2009-2018</td>
<td>0.5° (∼ 46 × 46 km)</td>
<td>Process</td>
</tr>
<tr>
<td></td>
<td>Water run-off per cell, mean for years 2009-2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LUCI&lt;sup&gt;4†&lt;/sup&gt; (Land Utilisation Capability Indicator)</td>
<td>Above ground carbon stocks</td>
<td>10 × 10 meters</td>
<td>Look-up table</td>
</tr>
<tr>
<td></td>
<td>Accumulated water run-off</td>
<td>5 × 5 meters</td>
<td>Process</td>
</tr>
<tr>
<td>$$-benefit transfer using The Economics of Ecosystems and Biodiversity database&lt;sup&gt;5,6†&lt;/sup&gt;</td>
<td>Above ground carbon stock as monetary value</td>
<td>25 × 25 meters</td>
<td>Look-up table</td>
</tr>
<tr>
<td></td>
<td>Water run-off as monetary value per cell</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aqueduct v2.1 Total Blue Water&lt;sup&gt;7§&lt;/sup&gt;</td>
<td>Accumulated water run-off</td>
<td>138 flow areas</td>
<td>Deterministic</td>
</tr>
<tr>
<td>ARIES k-Explorer&lt;sup&gt;8‡&lt;/sup&gt; (Artificial Intelligence for Environment &amp; Sustainability)</td>
<td>Joined above and below ground carbon stocks</td>
<td>1-hectare</td>
<td>Look-up table</td>
</tr>
<tr>
<td>Barredo et al. (2012)&lt;sup&gt;9&lt;/sup&gt;</td>
<td>A European map of above ground biomass stocks</td>
<td>1 km&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Look-up table</td>
</tr>
<tr>
<td>Copernicus, Tree Cover Density&lt;sup&gt;9§&lt;/sup&gt;</td>
<td>Proxy for carbon: tree Cover Density 2015 from MODIS satellite imagery.</td>
<td>20 × 20 meters</td>
<td>Deterministic</td>
</tr>
<tr>
<td>DECIPHeR&lt;sup&gt;10§&lt;/sup&gt; (Dynamic fluxEs and ConnectIVITY for Predictions of HydRology)</td>
<td>Accumulated water run-off through NRFA delineated catchment outlets, mean for years 1995-2015</td>
<td>387 catchments in common with validation</td>
<td>Process</td>
</tr>
<tr>
<td>Grid-to-Grid&lt;sup&gt;1118&lt;/sup&gt;</td>
<td>Accumulated water run-off, mean for years 1995-2015</td>
<td>1 km&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Process</td>
</tr>
<tr>
<td>Henrys et al. (2016)&lt;sup&gt;8&lt;/sup&gt;</td>
<td>Above ground carbon stocks</td>
<td>1 km&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Look-up table</td>
</tr>
<tr>
<td>Kindermann et al. (2008)&lt;sup&gt;9&lt;/sup&gt;</td>
<td>A global map of above ground forest biomass stocks</td>
<td>1 hectare</td>
<td>Deterministic</td>
</tr>
<tr>
<td>National Forest Inventory (2018)&lt;sup&gt;112‡&lt;/sup&gt;</td>
<td>Woodland Land Cover Map&lt;sup&gt;15&lt;/sup&gt; with above ground carbon stocks based on added Look-up table (Table. SI-1-4)</td>
<td>20 × 20 meters</td>
<td>Look-up table</td>
</tr>
<tr>
<td>Scholes Growth Days&lt;sup&gt;13,14†&lt;/sup&gt;</td>
<td>Proxy for water run off per cell: # Days precipitation exceeds evapotranspiration</td>
<td>1 km&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Deterministic</td>
</tr>
<tr>
<td>WaterWorld v2&lt;sup&gt;15,51&lt;/sup&gt;</td>
<td>Accumulated water run-off</td>
<td>0.0083° (∼ 1 km&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>Process</td>
</tr>
</tbody>
</table>

<sup>1</sup>Output generated for this work; <sup>2</sup>online tool; <sup>3</sup>existing dataset; <sup>4</sup>Kareiva et al. (2011); <sup>5</sup>Smith et al. (2014); <sup>6</sup>Ahlström et al. (2015); <sup>7</sup>Thomas et al. (2020); <sup>8</sup>de Groot et al. (2012); <sup>9</sup>Costanza et al. (2014); <sup>10</sup>Gassert et al. (2015) <sup>9</sup>Martínez-López et al. (2019); <sup>10</sup>land.copernicus.eu/tree-cover-density/status-maps/2015; <sup>10</sup>Coxon et al. (2019a; 2019b);
Bell et al. (2018a; 2018b); Forestry Commission (2018); Scholes (1998); Willcock et al. (2019); Mulligan (2013); following Ding & Bullock (2018), Willcock et al. (2019).
2.2. Validation datasets (step 2)

Our carbon stock validation dataset was provided by Forest Research and comprises species inventories in all forest estates in England and Scotland in 2019 (data-forestry.opendata.arcgis.com; density shown in Figure 3; locations in Figure SI-1-2). In 201,143 forest compartments of varying size (mean: 4.4 hectares. median 1.6 hectares, ± 22.1), tree species, stand age and thinning regime were recorded for three vegetation layers. For each compartment and layer therein, the unique combination of stand age, thinning regime and tree species of the inventory data was searched in the UK Carbon Code tables (woodlandcarboncod.org.uk) and life-time accumulated biomass was converted to total standing carbon per hectare estimates per compartment, with the layers summed per compartment (SI-1-2). Subsequently, compartments were spatially joined into 2078 polygons of ‘forest’ that were separated if more than 25 meters distance from each other.

Our water supply validation dataset comprised 519 hydrometric gauging stations from the National River Flow Archive of the UK (NRFA; nrfa.ceh.ac.uk), with associated catchments representing a variety of sizes distributed across the whole of the UK (Figure 3). From the 1598 potential catchments in NRFA, we selected those that were >100 km² to get a robust mean run-off from the catchments. In cases where multiple gauging stations were found along the same river, based on name, only the largest was chosen to avoid pseudoreplication. An additional set of 41 Welsh catchments was included which did not meet this size criterion. Wales contains mainly small catchments due its geography – mountain ranges close to the sea – and so we selected catchments >25 km² to avoid this part of the UK being underrepresented. The data were polygons encompassing these catchments. Details are provided in SI-1-2.

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

For each individual model, predictions were obtained for each polygon in the validation dataset using the ArcGIS spatial analyst Zonal tool with a forced 2.5 m grid size environmental setting to minimise edge effects; i.e. all predicted values were obtained by resampling into 2.5 × 2.5 m grid cells. In most cases the modelled value per polygon was obtained by taking the sum of all constituent grid cell values, corrected for both actual grid size and the resampling to 2.5 m. In the case of accumulated flow models, we corrected for potential small scale differences in flow routing among these models by taking the maximum flow value within both a 2 km range of the NRFA reported location of the gauging station and the polygon associated with that gauging station.

To ensure comparability among model outputs, we standardised by normalising among the outputs for each individual model and for the validation data-sets. Prior to this step all outputs were area corrected as either mean carbon stock – or proxy thereof – per hectare or water supply per hectare of catchment (with accumulated run-off estimates post-processed to give net run-off per cell; SI-1-1). This normalisation followed Willcock et al. (2019), and allowed us to address differences in units among models (such as monetary benefit transfer vs. satellite-based tree cover densities or run-off, and equalised carbon and biomass). To avoid impacts of extreme values without eliminating such data-points, we employed a double-sided Winsorising protocol for normalisation (Willcock et al. 2019; Verhagen et al. 2017), using the values associated to the 2.5% and 97.5% percentiles of number of datapoints to define the 0 and 1 values (values below or above these percentiles became 0 or 1 respectively). This winsorising normalisation protocol assumes outlier data are valid, but skewed values, in our case mainly by per area averaging, and corrects for this by compressing the variance tails rather than trimming them (Keselman et al. 2008; Erceg & Mirosevich 2008). Hence, we trade-off an even data distribution over the full 0-1 normalised range against the chance of having a true far outlier maximum (see SI-5 for a full investigation into the impact of the Winsorising protocol over standard normalisation for the validation data distribution). For each model, normalisation was done prior to creating ensembles.

For validation, we employed two accuracy measures (Willcock et al. 2019; Willock et al. 2020), which are related to different aims in modelling ES (Table 1):
1) Comparing the rank order of predicted and validation data using Spearman \( \rho \). This is relevant where modelling is used to discover, for example, the most important locations for delivering an ES, or conversely, those areas whose development may have least impact on ES delivery.

2) Ascertaining the absolute difference of each modelled value from its validation value using the inverse of the deviance \( (D^1)\). This is relevant where modelled values are important, e.g. when testing where ES levels exceed a minimum threshold. We used the inverse of the deviance so that, like \( \rho \), a higher value indicated greater accuracy.

### 2.4. Generate ensembles (step 5) and compare accuracy among ensemble types (step 6)

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

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

For weighted ensembles we calculated:

\[
E_{(\chi)} = \sum_{i}^{n} \left( \frac{w_i}{\sum_{i}^{n} w_i} \times Y_i \right)_{(\chi)} \]  

with positive weights \( w_i \) for model \( i \) of validation polygon \( \chi \), weights \( w_i \) are normalised to sum to 1, \( Y_i \) the modelled values for \( i \) per polygon (step 3), and \( n \) the total number of models per service.

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

1) Untrained ensembles (En-3 to En-8) represent a situation in which there is no validation data. To generate uncertainty estimates allowing statistical comparison with the models and among ensembles we jack-knifed (Araújo & New 2007; Refsgaard et al. 2014) with 50% of the spatial data polygons for 250 runs, i.e. every run contained a new selection of half the dataset. We tested three approaches to produce the ensembles:

- **Deterministic consensus** among models can be calculated using several approaches, including the fit to a common consensus axis such as from a Principal Components Analysis (Marmion et al. 2009; Grenouillet et al. 2011) or weighting by correlation coefficients (En-3 & En-4; ensemble numbering follows Table 1).

- **Iterative approaches** might more accurately quantify consensus among models through using structured trial-and-error (Dormann et al. 2018; Tebaldi & Knutti 2007). We use two regression techniques: between the individual models and the median (En-5) and leave-one-out cross-validation (En-6) following the suggestion in Dormann et al. (2018).

- One might a priori place value on a particular model attribute and use this to create weights (Englund et al. 2017; Willcock et al. 2019; Brun et al. 2020; En-7, En-8 & En-9). For example, one could up- or down-weight more distinct model types through a binary matrix of differences (En-8 & En-9; S1-
1-4) in land cover map used, grid-size, measured or modelled climate, model extent, presence of 
time-series, time step-size and model type (i.e. look-up table, deterministic or process based).
Alternatively models that run at coarser spatial resolutions are penalised (En-7): smaller grid sizes 
are deemed more useful for decision-making (Willcock et al. 2016).

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

After creating the ensembles, their accuracy was assessed following step 4 using the two measures (see 2.3):
Spearman $\rho$ and the inverse of the deviance ($D^i$). We assessed any improvement over the unweighted mean-
averaged ensemble as the reference with pairwise t-tests against the null hypothesis of equal accuracy 
(Matlab ttest-tool). A similar analysis against the median-averaged ensemble as reference can be found in 
SI-2. To avoid spurious findings of significance through having a large number of replicates, we assessed 
improvement using bootstrapped tranches of 50 runs each with 250 replicates, and averaging the P-values.
Since we used the same statistical test 12-times per service per accuracy estimate, we employed a full 
conservative Bonferroni correction; ($\alpha = 0.05/12$) on the resulting average P-values. To compare the 
ensembles with the individual models we calculated per replicate the mean difference in accuracy among 
all models ($A_i$) against accuracy of an ensemble ($A_E$) following: 
$$\left(\sum_i^n \frac{A_i - A_E}{A_i - 1} \right) \times \frac{1}{n},$$ 
with $n$ the number 
models and $i$ an individual model.

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

2.5. Spatial representation of ensembles and uncertainty (step 7)
To better support decision-making, we mapped our ES ensembles for the UK. For all the water ensembles, 
the mean normalised value across jack-knifed ensemble predictions per ensemble method were mapped as 
catchment polygons (step 5, $N = 519$). For all carbon ensembles we mapped as 1 km$^2$ grid cells. Here, for 
each ensemble approach, the estimated weights as calculated for the validation polygons – mean averaged 
among jack-knife runs– were transferred to the full area, with the result aggregated to a 1 km$^2$ resolution 
based on the mean value among 1 hectare grid cells. In total, this carbon dataset has 253,802 cells that 
(partially) contain non-sea land cover. We transferred the weights calculated for the forests since running 
cross-validation approaches on over 250K data points would extremely time consuming to compute.
However, since our validation data are only from forests/woodlands, we are aware of introducing a potential 
bias that could skew non-forested areas to lower values. Furthermore, we generated UK-scale maps of 
spatial variation in the differences among the untrained ensemble approaches, by calculating the standard 
error of the mean (SEM) among these spatial outputs. These maps are freely available online 
(https://doi.org/10.5285/a9ae773d-b742-4d42-ae42-2b594bae5d38), and spatial patterns of uncertainty are 
discussed in SI-4.

3. Results

3.1. Ensembles are more accurate than individual models
The average accuracy of individual models, represented by the mean of accuracy values taken across all 
models, was lower than that for any of the ensembles we created. The accuracy of the unweighted averaged 
ensembles (of modelled values at each location, e.g. ‘mean ensemble’) was appreciably higher than the
mean value for accuracy of the individual models for both carbon and water: \(19\% \pm 1.1\% \) [sd] for \(\rho\) and

\(12.1\% \pm 0.5\%\) for \(D^i\) improvement in fit to the validation data for carbon and \(5.7\% \pm 0.4\%\) for \(\rho\) and \(9.5\%\)

\(\pm 1.7\%\) for \(D^i\) for water (Figure 2). Untrained weighted ensembles showed large improvements — for most, larger than the unweighted ensembles — over the mean accuracy of the individual models of \(17\%\) to \(27\%\)

\((\rho)\) and \(7.6\%\) to \(15\%\) \((D^i)\) for carbon (Figure 2A and B), and \(5.3\%\) to \(6.5\%\) \((\rho)\) and \(7.7\%\) to \(18\%\) \((D^i)\) for

water (Figure 2C and D). In all cases, pairwise t-tests indicated highly significant differences between each ensemble and the mean value of accuracy of individual models (all \(P<1E^{-10}\)). Thus, creating an ensemble improves prediction accuracy against a randomly chosen individual model irrespective of the ensemble approach chosen.
Figure 2. Accuracy of above ground carbon stock ensembles (10 models; A and B), and of water supply ensembles (9 models; C and D) against validation data. The mean of accuracy values across the containing models – i.e. a randomly chosen model – is provided for comparison. For detail on the different ensemble types see Table 1 and SI-1-3. We show the average accuracy of 250 bootstrap runs with 50% of the dataset. The vertical dashed line indicates the reference unweighted mean-averaged ensemble (black dot, ‘mean ensemble’). Error bars indicate the standard deviation among runs in terms of proportional difference to the mean ensemble, calculated per bootstrap run as the difference in accuracy to the mean ensemble divided by the accuracy of the mean ensemble. The coefficient of variation among bootstraps for the mean carbon ensemble was 4% and 1%, for ρ and D↑ respectively, and 1 % and 2% for water (not shown). Blue coloured ensemble accuracies are significantly higher than the unweighted mean ensemble (Bonferroni corrected α = (0.05/12)); Red coloured bars are significantly lower; Black dashed bars are not significantly different to the mean ensemble.
3.2. Weighted ensembles are more accurate than unweighted ensembles

All weighted ensembles, whether trained or untrained, significantly outperformed the reference unweighted mean ensemble (Figure 2), with the exception of $D^3$ for carbon. In all cases, pairwise t-tests indicated these differences were highly significant ($P<1E^{-10}$; see Figure SI-2-1 for similar analyses against the median-averaged ensemble).

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

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

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

We validated the robustness of our results using independent data and models from a different area (Sub-Saharan Africa; Willcock et al. 2019), which gave similar results of weighted ensembles outperforming the reference mean ensemble (Figure SI-2-2).
Figure 3. Spatial distribution of validation points and the reference mean ecosystem service value. A the Distribution of 2078 carbon validation forests as coverage of 10 × 10 km cells – many individual forest fragments would be too small to be clear at this scale, see SI SI-1-2 –, white cells are empty. B the reference unweighted mean ensemble of carbon across 10 models, normalised on scale 0-1. C the 519 catchments used for water validation and ensemble calculations coloured by their size – smaller watersheds that overlap larger ones are displayed on top; lines show underlying largest catchment level. D the reference unweighted mean ensemble of water supply across 9 models, normalised on scale 0-1. All maps here, in SI-3 (all ensembles) and SI-4 (uncertainty) could support landscape decisions in the UK and are available via https://doi.org/10.5285/a9ae773d-b742-4d42-ae42-2b594bae5d38.

4. Discussion

We have shown that predictions from ensembles of models have substantially higher accuracy than a randomly selected single ES model, and especially that weighting approaches increase ensemble accuracy. Finding increased performance through use of ensemble approaches is common in other fields. For example, the increased accuracy of ensemble species distribution models ranges from 1-2% (Crossman et al. 2012; Abrahms et al. 2019) to 12% (Grenouillet et al. 2011), although an increase is not universal (Hao et al. 2020). Similarly, 2% accuracy increases were found for market forecasting ensembles (He et al. 2012), and neural network ensemble averaging resulted in up to 7% improvements in accuracy (Inoue & Narisha 2000).
Specific to ES, unweighted averaged ensembles have been shown to be 5.0–6.1% more accurate than individual models (Willcock et al. 2020). Our improvements with ES ensembles are at minimum 5%-17%, suggesting substantial differences among models in their adequacy (Dormann et al. 2018), but also that ensemble approaches that use more information offer greater increases in accuracy. We found that taking the median generally outperforms a mean ensemble, probably because the latter is more influenced by outliers. Our results provide evidence that weighted ES ensembles created using consensus techniques produce more accurate outputs than unweighted ensembles. This finding is supported by our additional analysis using independent models and data from Sub-Saharan Africa (in a biome with very different climatic and soil characteristics; SI-2), suggesting our findings may be generalisable, although investigating this specifically (e.g., for different ES, regions and validation datasets) is an important avenue for future research.

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

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

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

5. Conclusion

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

References

https://doi.org/10.1111/ddi.12940


Bell, V.A. et al. (2018b). Grid-to-Grid model estimates of monthly mean flow and soil moisture for Great Britain (1891 to 2015): observed driving data [MaRIUS-G2G-Oudin-monthly]. [Data Set] (NERC Environmental Information Data Centre). [https://doi.org/10.5285/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0](https://doi.org/10.5285/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0)


Coxon, G. et al. (2019b). DECIPHER model estimates of daily flow for 1366 gauged catchments in Great Britain (1962-2015) using observed driving data. [Data Set] (NERC Environmental Information Data Centre). [https://doi.org/10.5285/d770b12a-3824-4e40-8da1-930cf9470858](https://doi.org/10.5285/d770b12a-3824-4e40-8da1-930cf9470858)


Forestry Commission, United Kingdom. (2018). *National Forest Inventory Woodland GB 2018*. [Data Set] (Forestry Commission Open Data). [http://data-forestry.opendata.arcgis.com/datasets/d3d7bfba1cda4a3b83a948f33c5777c0_0](http://data-forestry.opendata.arcgis.com/datasets/d3d7bfba1cda4a3b83a948f33c5777c0_0)

https://doi.org/10.1016/j.ecolmodel.2018.11.013


https://doi.org/10.1038/495305a


https://doi.org/10.1111/ecog.04890


https://doi.org/10.5285/9be652e7-d5ce-44c1-a5fc-8349f76f5f5c


https://www.oxforduniversitypressscholarship.com/view/10.1093/acprof:oso/9780199588992.001.0001/a cprof-9780199588992

https://doi.org/10.1037/1082-989X.13.2.110


https://doi.org/10.1029/2011GL046864


