Reducing Uncertainty in Ecosystem Service Modelling through Weighted Ensembles

Hooftman, Danny; Bullock, James; Jones, Laurence; Eigenbrod, Felix; Barredo, Jose; Forrest, Matthew; Kinderman, George; Thomas, Amy; Willcock, Simon

Ecosystem Services

DOI:
https://doi.org/10.1016/j.ecoser.2021.101398

Published: 01/02/2022

Peer reviewed version

Supplementary Information
to

Reducing Uncertainty in Ecosystem Service Modelling through Weighted Ensembles

All Matlab codes written for this manuscript can be found in the GitHub account: https://github.com/EnsemblesTypes/

SI-1 Validation sets, models and ensembles

Table SI-1-1. Validation sets and Models (across 2 pages), including their coverage within the United Kingdom. The UK consists of 4 countries: England, Scotland, Wales –together referred to as Great Britain (GB) – and Northern Ireland (NI; Figure SI-2-1); Self-governing dependencies of the UK are not included (such as the Isle of Man and Channel Islands); *i.e.* <GB + NI> indicates the full scale studied (242,495 km^2). For model descriptions see below. Model type terminology follows Mulligan & Wainwright (2013), Ding & Bullock (2018) and Willcock *et al.* (2019). References to these models and examples of usage can be found at the descriptions below.

Validation sets	Service/Proxy	Coverage	Grid Size & timeframe if applicable	Type
National River Flow Archive	Flow volume through gauging stations	GB + NI	519 selected Watersheds, annual average 1995-2015	Measurements
Forest Research, inventories	Forest estates carbon stocks	England & Scotland	2078 continuous forest fragments in 2019	Inventories
Models run especially for this work (SI-1-1a)				
InVest	Above ground carbon stocks	GB + NI	25 × 25 meter	Look-up table
	Water run-off per cell	GB + NI	Idem	Process model
$-benefit transfer	Above ground carbon stock monetary value	GB + NI	25 × 25 meter	Look-up table
	Water run-of monetary value per cell	GB + NI	Idem	Look-up table
Scholes Growth Days	# Days precipitation exceeds evapo-transpiration; proxy Water for run-off	GB + NI	1-km^2	Deterministic model
LPJ-GUESS	Vegetation biomass stocks	GB + NI	0.5 Degrees (≈ 45.6 × 45.6 km), average for years 2009-2018	Process model
	Water run-off per cell	GB + NI	Idem	Process model
LUCI	Above ground carbon stocks	GB	10 × 10 meter	Look-up table
	Flow volume in rivers: accumulated run-off	England & Wales	5 × 5 meter	Process model
National Forest Inventory Woodland GB 2018 with added Look-up table	Above ground carbon stocks with partial input selection	GB	20 × 20 meter	Look-up table
Models run through online tools without input selection options (SI-1-1b)				
WaterWorld	Flow volume in rivers: accumulated run-off	GB + NI	0.0083 degrees (≈ 1-km^2)	Process model

ARIES k-Explorer[S14]	Joined above and below ground carbon stocks	GB + NI	1-hectare	Look-up table
Existing modelled outputs from online repositories or existing with co-authors (SI-1-1c)				
Grid-to-Grid	Flow volume in rivers: accumulated run-off	GB	1-km^2, annual average 1995-2015	Process model
DECIPHeR	Flow volume through watershed outlet: accumulated run-off	GB	487 catchments overlapping with validation set, annual average 1995-2015	Process model
Aqueduct v2.1 Total Blue Water	Water run-off	GB + NI	138 flow areas	Deterministic model
Henrys *et al.* (2016)	Above ground carbon stocks	England	1-km^2	Look-up table
Barredo *et al.* (2012)	Above ground biomass stocks	GB + NI	1-km^2	Look-up table
Kindermann *et al.* (2008)	Above ground biomass stocks	GB + NI	1-hectare	Deterministic model
Copernicus, Tree Cover Density	Proxy: tree cover density per cell	GB + NI	20 × 20 meter	Deterministic model

SI-1-1 Model inputs and outputs used

SI-1-1a Models run especially for this work

InVest

InVest is a suite of stand-alone, free and open-source models from the Natural Capital Project (Kareiva *et al.* 2011; McKenzie *et al.* 2014) and are downloaded as one package from the website (naturalcapitalproject.org/invest/). Extended descriptions of each model are provided in the online user guide (data.naturalcapitalproject.org/ nightly-build/invest-users-guide/html/).

InVest comprises independent modules, each module covering one ecosystem service. In this study we used two of the more widely used modules; the water yield module (*e.g.* Redhead *et al.* 2016) and the carbon module (*e.g.* Goldstein *et al.* 2012) of release v3.7.0, which was the current version at the time of conducting this part of our research in 2020. Although software-based, the two InVest modules we used do not contain autonomously drawn-in data sources. Instead, all datasets need to be provided manually. The output generated is at the spatial resolution equal of the provided land cover map, which was the LCM2015 (Rowland *et al.* 2017) in our case.

InVest water yield module.

The InVest water yield model is a process model, built as a hydropower module, identifying quantitatively how much water or economic value each part of the landscape contributes to hydropower production. This is done by estimating water run-off through a single point. The model has three components: water yield, water consumption, and hydropower valuation. We employed the first component here, using the gridded outputs of water run-off per grid cell, allowing standardising extraction per validation polygon among model data sets. Parametrisation followed Redhead *et al.* (2016) as far as feasible.

As input data we used:

- Annual total precipitation at 1-km gridded estimates of monthly rainfall for Great-Britain and Northern Ireland from 1890 to 2017 (CEH-GEAR: Tanguy er al. 2019). The rainfall estimates are derived from the Met Office national database of observed precipitation and are downloaded from CEH- EIDC (catalogue.ceh.ac.uk/documents/ee9ab43d-a4fe-4e73-afd5-cd4fc4c82556).

The data were extracted per month, summed per year and averaged for the period 1996-2015. The employed end date of 2015 matches the used Grid-to-Grid (Bell *et al.* 2018b and DECIPHeR (Coxon *et al.* 2019b) existing model outputs.

- Global Potential Evapotranspiration from CGIAR-CSI on a 0.009 degree raster: (csi.cgiar.org/Aridity/), clipped to the UK and resampled to an exact 1-km raster.
- Root restricting raster was obtained from the European Soil Database (ESDB) version 2: esdac.jrc.ec.europa.eu/content/european-soil-database-v20-vector-and-attribute-data. Soils for the UK were extracted. The soil depths was calculated as to the minimum of the rock depth (DR) or the depth to a gleyed horizon (DGH) with a maximum of 1500 cm. Subsequently, this polygon layer was converted to an exact 1-km grid.
- Land use was following the LCM2015 (Rowland *et al.* 2017), the leading UK land cover map when conducting this research, with 21 classes at a 25 × 25 meter resolution. catalogue.ceh.ac.uk/documents/bb15e200-9349-403c-bda9-b430093807c7
- Plant Available Water Content (PAWC) raster was obtained from the European Soil Database (ESDB) version 2 (link see above). Soils for the UK were extracted. PAWC was calculated based on Easily Available Water Capacity (EAWC) as weighted average of two layers, the topsoil of 20 cm depth and subsoil up to the root restriction depth calculated above. Subsequently, this polygon layer was converted to an exact 1-km grid.
- The seasonality factor (Z) was set at 30, following (Redhead *et al.* 2016)
- The maximum rooting depth and evapotranspiration coefficient (Kc), as look-up table per Land Cover class, were provided by John Redhead as used in Redhead *et al.* (2016).

InVest carbon module

This InVest module is a look-up table based model, which uses maps of land use and land cover types and data on wood harvest rates, harvested product degradation rates, and stocks in four carbon pools (aboveground biomass, belowground biomass, soil, dead organic matter) to estimate the amount of carbon currently stored in a landscape or the amount of carbon sequestered over time. We did not employ the sequestration functions and restricted ourselves to above ground, standing, carbon pool only to match our Forest Research (FR; SI-1-2) validation set. The model generates gridded maps of standing carbon per land use based on the carbon pools at the spatial resolution equal of the provided land cover map, which was the LCM2015 (Rowland *et al.* 2017) in our case.

As input data we used:

- Land use was following the LCM2015 (Rowland *et al.* 2017), the leading UK land cover map when conducting this research, with 21 classes at a 25 × 25 meter resolution. catalogue.ceh.ac.uk/documents/bb15e200-9349-403c-bda9-b430093807c7
- Ecofloristic zones via CDIAC: (cdiac.ess-dive.lbl.gov/ftp/global_carbon/ecofloristic_zones.zip)
- Carbon stocks per land use class per ecofloristic zone via CDIAC (Ruesch & Gibbs 2008; Table SI-1-2): (cdiac.ornl.gov/epubs/ndp/global_carbon/carbon_tables.pdf). We only used above ground stored carbon values. In the model carbon stocks for the other layers, below ground, soil and dead material were set to 0.

Table. SI-1-2. Look-up table above ground carbon estimates in tonnes per hectare per ecofloristic region (Ruesch & Gibbs 2008) per LCM2015 class (Rowland *et al.* 2017) used in the InVest Carbon module

LCM2015 class	Temperate Oceanic Forest	Temperate mountain systems	Boreal coniferous forest	Boreal mountain systems
Broadleaved woodland	69	58	33	9
Coniferous Woodland	73	61	33	9
Arable and Horticulture	5	5	5	5
Improved Grassland	5	5	5	5
Neutral Grassland	4.5	4.5	5	5
Calcareous Grassland	4.5	4.5	5	5
Acid grassland	4.5	4.5	5	5
Fen, Marsh and Swamp	7.4	7.4	3	3
Heather	7.4	7.4	3	3
Heather grassland	4.5	4.5	5	5
Bog	7.4	7.4	3	3
Inland Rock	1	1	1	1
Saltwater	0	0	0	0
Freshwater	0	0	0	0
Supra-littoral Rock	1	1	1	1
Supra-littoral Sediment	1	1	1	1
Littoral Rock	1	1	1	1
Littoral sediment	1	1	1	1
Saltmarsh	7.4	7.4	3	3
Urban	0	0	0	0
Suburban	0	0	0	0

$- Benefit transfer

$- Benefit transfer is a look-up table based model employing Costanza *et al.* (2014), who provides estimates of the monetary value of global ecosystem services based on the TEEB study (de Groot *et al.* 2012) and associated SVD-TEEB-database. Following Costanza *et al.* (2014) and Willcock *et al.* (2019), we associated LCM 2015 classes (Rowland *et al.* 2017) with benefit values for water supply and Climate regulation (carbon) at a 25 x25 meter grid size (Table SI-2-2). Here water was defined as use per person by dividing by population density at a 1-hectare resolution using WorldPop (2018).

Table. SI-1-3. Look-up Table values in US-$ per hectare per year as used for $-benefit transfer

LCM2015 class	Climate Regulation	Water Supply	Costanza *et al.* Category
Broadleaved woodland	152	191	Temperate Forest
Coniferous woodland	152	191	Temperate Forest
Arable and horticulture	0†	400	Cropland
Improved grassland	0†	400	Cropland
Neutral grassland	40	60	Grassland
Calcareous grassland	40	60	Grassland
Acid grassland	40	60	Grassland
Fen, marsh and swamp	2‡	408	Swamps/Floodplains
Heather	2‡	408	Swamps/Floodplains
Heather grassland	2‡	408	Swamps/Floodplains
Bog	2‡	408	Swamps/Floodplains
Inland rock	0	0	n/a
Saltwater	0	0	n/a
Freshwater	0	1808	Lakes/Rivers
Supra-littoral rock	0	0	n/a
Supra-littoral sediment	0	0	n/a
Littoral rock	0	0	n/a
Littoral sediment	0	0	n/a
Saltmarsh	65	1217	Tidal Marsh/Mangroves
Urban	54¶	0	Urban
Suburban	145§	0	Urban

†Assumed to be fully harvested = 0 stock; ‡ from (de Groot *et al.* 2012) for the UK; ¶ adaptation as: [904.7 × 6%] with the 6% the tree density ((land.copernicus.eu/pan-european/high-resolution-layers/forests/tree-cover-density/status-maps/2015) within Urban cell in CM2015 cells; § idem as ¶ but with 16% tree density in suburban cells.

Scholes Growth days

The annual number of days rainfall exceeds evapotranspiration was calculated in monthly bins (Scholes 1998; Willcock *et al.* 2019) as statistical deterministic model and is calculated as:

$$GD_x = \sum_{m=1}^{12} \left(\frac{d_m * P_{(m,x)}}{E_{(m,x)}} \right) \qquad \text{Eq. SI-1-1}$$

With P = precipitation in grid cell x in month m; E = potential evapotranspiration in grid cell x in month m; m = month (1 to 12), x = 1-km^2 grid cell and d = number of days per month.

We used the follow source data:

1) Monthly precipitation using WorldClim, version 2.1 (Fick *et al.* 2017), on a 0.009 degree resolution (30 degree seconds) and resampled to exactly 1000 × 1000 meter cells from (worldclim.org/).

2) Monthly global Potential Evapotranspiration from CGIAR-CSI (Zomer *et al.* 2006) on a 0.009 degree resolution (30 degree seconds) and resampled to exactly 1000 × 1000 meter cells from (csi.cgiar.org/Aridity/).

LPJ-GUESS

The Lund–Potsdam–Jena General Ecosystem Simulator (LPJ-GUESS, www.nateko.lu.se/lpj-guess; *e.g.* Smith *et al.* 2014; Ahlström *et al.* 2015) is a dynamic vegetation/ecosystem process model designed for regional to global applications. The model combines process-based representations of terrestrial vegetation dynamics and land–atmosphere carbon and water exchanges in a modular framework. LPJ-GUESS explicitly considers key vegetation and ecosystem processes such as photosynthesis, vegetation growth, resource (light, water and nitrogen) competition between plant functional types (PFTs), disturbance by fires, and carbon storage.

Compared to other dynamic global vegetation models, LPJ-GUESS represents vegetation dynamics and canopy structure at a high level of detail and simulates individual trees with a forest gap model approach (Bugmann 2001). Stochastic establishment and mortality of individual trees is simulated at the so-called patch level (1000-m^2), and the results of a number of replicate patches (here 10) are averaged to represent the overall vegetation of a model grid cell with homogenous environmental input. This detail is an advantage for regional or local applications (Smith *et al.* 2001; Hickler *et al.* 2012).

The model has been successfully evaluated against a large number of benchmarks at local to global scales (*e.g.* Smith *et al.* 2014; Ahlström *et al.* 2015; see more examples at : iis4.nateko.lu.se/lpj-guess/LPJ-GUESS_bibliography.pdf). The version applied here described fully in Smith *et al.* (2014) and the references therein, with the additional inclusion of the SIMFIRE-BLAZE fire model (Knorr *et al.* 2019). The input datasets for climate, atmospheric CO_2 concentration, human population density, nitrogen deposition and land use are identical to those used by the dynamic global vegetation models (including LPJ-GUESS) contributing to the Global Carbon Project 2019 (Friedlingstein *et al.* 2019). For the climate data input, CRU-JRA 6-hourly forcing (was aggregated to daily resolution and used (Viovy 2009; Harris *et al.* 2014; Kobayashi *et al.* 2015), rather than the alternative monthly climate input dataset.

A more general popular-science description of the model also addressing non-experts, but not including the most recent parameterization, the nitrogen cycle or the SIMFIRE-BLAZE fire model, can be found here: iis4.nateko.lu.se/lpj-guess/guess.pdf.

LPJ-GUESS output was provided by co-author Matthew Forrest. We used:

- The vegetation carbon stocks;
- The total simulated water runoff.

LPJ-GUESS was applied at 0.5° spatial resolution, which approximates to 45.6 × 45.6 km grid cells in the UK.

LUCI

The Land Utilisation Capability Indicator (www.lucitools.org/; *e.g.* Sharps *et al.* 2017; Trodahl *et al.* 2017; Thomas *et al.* 2020) is an ecosystem services modelling tool which illustrates the impacts of land use on various ecosystem services. It runs at fine spatial scales modelling a range of ecosystem services, and produces outputs as both physical units, and a status classification. LUCI combines status classification data for multiple services to identify areas where landscape usage change might be beneficial, and where maintenance of the status quo might be desirable. The model automatically applies spatial discretisation at the same resolution as the topographic data, which may be considered false precision for highly spatially aggregated climate datasets, but is more appropriate for hydrological routing functions, and allows for minimal loss of spatial information from the polygon land use dataset. The carbon stock model is look-up table based and is estimated in this case at a 10 × 10 meter scale. The water yield module provides accumulated water run-off at a 5 × 5 meter scale as process model, and also calculates in-stream annual average flow. Existing runs were collated for this study.

LUCI water yield module.
The LUCI (Land Utilisation & Capability Index) model simulates flow accumulation over the landscape, using GIS functions for calculating flow direction and accumulating water through the landscape to the watercourse, to give annual average streamflow at all points on the river network. For this model run, evapotranspiration was a user provided input which does not change with changing land cover; development is underway to modify evapotranspiration according to changing land cover and soil properties but this functionality was not parameterised for the UK at the time of writing.

As input data we used:

- Annual total precipitation at 1-km gridded estimates of monthly rainfall for Great-Britain and Northern Ireland from 1890 to 2014 (CEH-GEAR). The rainfall estimates are derived from the Met Office national database of observed precipitation and are downloaded from CEH- EIDC (https://catalogue.ceh.ac.uk/documents/f2856ee8-da6e-4b67-bedb-590520c77b3c). The data were extracted per month, summed per year and averaged for the period 2000-2010.
- Potential Evapotranspiration (PET) was calculated using the CHESS (Climate hydrology and ecology research support system) meteorological dataset (Robinson *et al.* 2017) and the Penman-Monteith equation (Monteith 1965) parameterised as a reference grass crop everywhere.
- Digital terrain model (DTM) with 5m spatial resolution from NextPerspectives. Licensed to: UK Centre for Ecology & Hydrology for PGA, through Next Perspectives™

LUCI carbon module
The LUCI carbon model is look up table based, and maps carbon in soils and vegetation using input data on soil type and land use. Soil carbon was not included here, but is generally mapped according to England and Wales average values for land use and soil combinations at two depths: a) 0-30 cm depth and b) 0-1m depth. Vegetation carbon was calculated based on Tier 2 IPCC partitions of aboveground live, aboveground dead and belowground.

As input data we used:

- Land use was assigned using LCM2007 (Morton *et al.* 2014), in polygon format (LCM2007 © and database right NERC (CEH) 2011. All rights reserved. Contains Ordnance Survey data © Crown copyright and database right 2007.)
- A lookup table provided as part of the LUCI model toolboxes

National Forest Inventory Woodland GB 2018
The NFI woodland map covers all forest and woodland area over 0.5-hectare with a minimum of 20% canopy cover (or the potential to achieve it) and a minimum width of 20 metres, including areas of new planting, clear-fell, windblow and restocked areas. This grid layer was downloaded via data-forestry.opendata.arcgis.com/datasets/d3d7bfba1cba4a3b83a948f33c5777c0_0 at a 20 × 20 meter scale, to which we added a self-collated look-up table (Table SI2-3).

Table. SI-1-4. Look-up table for carbon stock in tons per hectare for National Forest Inventory woodland

FC Inventory Class	Carbon stock	Description/Justification
Grass	4.50	From CDIAC†: Grassland
Other vegetation	3.00	From CDIAC†: Bogs etc.
Open water	0	Assumed 0
Assumed woodland	48.2	Mean broadleaved and conifer
Broadleaved woodland	64.4	Weighted CDIAC† values across all 4 UK ecozones
Conifer woodland	32.0	Weighted CDIAC† values across all 4 UK ecozones
Ground prep	0	Assumed 0
Low density woodland	12.05	Assumed 25% of full woodland
Mixed mainly broadleaved woodland	64.4	= Broadleaved woodland
Mixed mainly conifer woodland	32.0	= Conifer woodland
Young trees	2.89	[Fraction of cumulative carbon for all species as: mean <=15 years vs. mean >=50 year ^ < 100 years)]‡ × mean(Broadleaved and Conifer)
Shrub	2.89	= young trees
Agriculture land	5.00	From CDIAC†: Arable
Bare area	0	Assumed 0
River	0	Assumed 0
Urban	0	Assumed 0
Felled	0	To be removed: 0 stock
Coppice	5.78	= young trees × 2
Cloud shadow	0	Assumed 0
Quarry	0	Assumed 0
Road	0	Assumed 0
Coppice with standards	5.78	= coppice
Windfarm	0	Assumed 0
Uncertain	0	Assumed 0
Power line	0	Assumed 0

†Ruesch & Gibbs 2008; ‡From Forest Research sequestration cumulative tables (see validation set)

SI-1-1b. Models run through online tools without input selection options

WaterWorld

WaterWorld (version 2; Mulligan 2013) is an internally parameterised, process-based model of water accumulation, which we used to model water supply in our study. The model as we used it is readily available via (www.policysupport.org/waterworld). A freely available description is provided on the website. We run and downloaded accumulated water supply at a ≈ 1-km^2 scale (0.0083°), the runs were conducted web-based.

ARIES

ARIES (ARtificial Intelligence for Ecosystem Services, but recently renamed to ARtificial Intelligence for Environment and Sustainability; *e.g.* Villa *et al.* 2014; Martínez-López *et al.* 2019) is a networked collaborative software technology designed for rapid ecosystem service assessment and valuation (http://aries.integratedmodelling.org/). It gives equal emphasis to ecosystem service supply, demand and

flow in order to quantify actual service provision and use by society (as opposed to quantifying potential service benefits). It aims to provide a suite of models that support science-based decision-making. We used the web-based ARIES Explorer (k.Explorer) that aims to allow non-technical users for carbon stocks at a 1-hectare scale as look-up table based model. The water supply module was not finished at the time of conducting this research (last check 07-04-2021).

SI-1-1c. Existing modelled outputs from online repositories or existing with co-authors

Grid-to-Grid

Grid-to-Grid (Bell *et al.* 2009; 2018a) is an hydrological accumulated flow process model for water supply in Great Britain (England, Scotland &Wales). This national-scale hydrological model runs on an exact 1000 × 1000 meter grid aligned with the GB national grid, at a 15-minute time-step, and is parameterised using digital datasets (*e.g.* soil types, land-cover). The effect of urban and suburban land-cover on runoff and downstream flows is accounted for in the model. The downloaded dataset (Bell *et al.* 2018b: catalogue.ceh.ac.uk/documents/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0) was produced as part of MaRIUS (Managing the Risks, Impacts and Uncertainties of drought and water Scarcity), which was a UK NERC-funded research project (2014-2017) that developed a risk-based approach to drought and water scarcity (mariusdroughtproject.org/).

We used the 1996-2015 predictions, averaging monthly $m^3 s^{-1}$ values per grid cell, and summing those into annual flows, weighted with seconds per month. For details see our anonymous GitHub account: https://github.com/EnsemblesTypes/DataExtractionTools, module *ExtractDataG2G*.

DECIPHeR

DECIPHeR (Dynamic fluxEs and ConnectIvity for Predictions of HydRology; Coxon *et al.* 2019a) is a process model framework that simulates and predicts hydrologicflows from spatial scales of small headwater catchments to entire continent. Here we downloaded DECIPHeR model estimates of daily flow for 1366 gauged catchments in Great Britain (1962-2015). The downloaded dataset (Coxon *et al.* 2019b: catalogue.ceh.ac.uk/documents/d770b12a-3824-4e40-8da1-930cf9470858) was produced as part of MaRIUS (Managing the Risks, Impacts and Uncertainties of drought and water Scarcity) to provide national scale probabilistic flow simulations and predictions for UK drought risk analysis. MaRIUS was a UK NERC-funded research project (2014-2017) that developed a risk-based approach to drought and water scarcity (mariusdroughtproject.org/).

We selected the 487 catchments corresponding between Coxon *et al.* (2019b) and our NRFA validation dataset. We used the 1996-2015 predictions, summing daily values into years, translating from $m^3 s^{-1}$ to days per grid cell. Subsequently we averaged among years. For details see our GitHub code at: https://github.com/EnsemblesTypes/DataExtractionTools, module *ExtractDataDECIPHeR*.

Aqueduct v2.1

From the world-wide maps for Water of the World resources Institute (Gassert et al 2015: wri.org/resources/data-sets/aqueduct-global-maps-21-data) we used the total annual Blue Water (BA) per main catchment, which is statistically deterministic modelled data. The total blue water estimate approximates natural river discharge and does not account for withdrawals or consumptive use. Since the 138 main catchments do not match NRFA aligned catchments, we resampled into 1-km^2 grid cells of equal value enabling to extract based on our smaller validation catchments.

Henrys *et al.* (2016)

Model estimates of *aboveground carbon for Great Britain* (Henrys *et al.* 2016) at a 1-km^2 scale downloaded from: catalogue.ceh.ac.uk/documents/9be652e7-d5ce-44c1-a5fc-8349f76f5f5c. This look-up table based spatial dataset presents estimates of total carbon stored in vegetation across England, not Great Britain as claimed in the title, and is based upon methodology developed in Milne & Brown (1997). Presented as carbon density (tonnes per hectare) at a 1-km scale, this map was produced using estimates of the average amount of carbon stored in each land cover type and upscaling to a full England

coverage, based on the estimated spatial distribution of land cover across England using the 2007 Land Cover Map (Morton *et al.* 2014). Estimates of carbon in tonnes per hectare for each land cover category as a whole, not being woodland, was estimated using Milne & Brown (1997) as well were carbon density for woodlands based on among species differences and age specific estimates. Countryside Survey data from 2007 (Maskell *et al.* 2008) was used to derive these species and age structure.

Barredo *et al.*
From the JRC report, *A European map of living forest biomass and carbon* stock (Barredo *et al.* 2012; Avitabile & Camia 2018), using the map of above ground forest living biomass at a 1-km^2 scale, provided into the project by co-author José I. Barredo. The report can be downloaded via op.europa.eu/en/publication-detail/-/publication/b9345574-a96f-4417-87ed-1a85d2252834/language-en and is described in Avitabile & Camia (2018).

The methodology for creating the map, based on IPCC (2006) and Ruesch & Gibbs (2008), is a look-up table associated with a Corine land cover map (EEA 1993; 2000) and the FAO's map of Global Ecological Zones for the Global Forest (FAO 2001). The method spatializes biomass density values in forest, sourced from the IPCC report on Guidelines for National Greenhouse Gas Inventories (IPCC 2006: tables 4.7, 4.8).

Biomass values were allocated to each grid cell taking into consideration the forest area and the biomass density corresponding to the ecological zone of the grid cell. Then, the forest biomass map was adjusted at grid cell level by applying ratios to match the country-level biomass values reported in the FAO's Forest Resource Assessment (FAO 2010). Therefore, the adjusted map is in agreement with reported biomass at country level.

The adjusted biomass map was validated using sub-national data from National Forest Inventories of four European countries. The validation indicated that the map represents faithfully the amounts of biomass at subnational level. Additionally, Avitabile & Camia (2018) implemented an assessment of forest biomass maps in Europe using harmonized national statistics and inventory plots. Their results indicate that the forest biomass map of Barredo *et al.* (2012) ranked high in comparison to other commonly used biomass maps of Europe.

Kindermann *et al.*
From Kindermann *et al.* (2008), updated for Avitabile & Camia (2018): *A global forest growing stock, biomass and carbon map based on FAO statistics*, using the carbon forest map at a 1-hectare scale, provided into the project by co-author Georg Kindermann. This model is a deterministic model downscaling country statistics given by FAO. Therefore it estimates site productivity using temperature, precipitation, radiation, altitude and soil characteristics. With this site productivity the standing biomass of full stocked normal forest stand with an increment optimal rotation time is estimated, weighted with a forest cover map and scaled to fit the country statistics.

Tree Cover Density 2015
As a proxy of carbon stock, the density of trees deterministically modelled from Modis satellite imagery at a 20 × 20 meter scale using the pan-European Copernicus data-base (land.copernicus.eu/pan-european/high-resolution-layers/forests/tree-cover-density/status-maps/2015).

SI-1-2 Validation datasets

UK National River Flow Archive
This validation set provides annual flow in m^3 through a catchment pour point.

The National River Flow Archive (NRFA: nrfa.ceh.ac.uk/; *e.g.* Dixon *et al.* 2013; Harvey *et al.* 2012) is the UK's focal point for river flow data. The NRFA collates, quality controls, and archives hydrometric data from gauging station networks across the UK including the extensive networks operated by the Environment Agency (England), Natural Resources Wales, the Scottish Environment Protection Agency and for Northern Ireland, the Department for Infrastructure - Rivers. The NRFA data underpin much of the hydrological research and water resources development and management activity in the UK. Peak flow datasets provided by the archive form the basis of UK industry standard flood frequency estimates for planning and development purposes. One of the key features of the NRFA is providing a central database and retrieval service of hydrometric measurements for 1,597 gauging stations associated to delineated catchments in the UK of which 1,239 stations were active in 2019. Their associated catchments vary in size from <1-km^2 to 9948 km^2 (larger Thames area).

We selected 519 selected hydrometric gauging stations from the NRFA database, with associated catchments varying in size and spatially spread across the UK (Figure S1-1). Spatial representation is as polygons of these watersheds. We used the following criteria to select gauging stations:

- The gauging station opened before 1996
- The station was still in use in 2015
- A delineated catchment was present as polygon in the database
- The catchment fitted the size criteria below and was not subsequent along the same river.

For the latter, firstly we selected all gauging stations with an associated catchment > 1000 km^2 (82 stations). Thereafter, we selected stations with associated catchment size > 100 km^2, with the prerequisite of not taking multiple gauging stations along the same river, given the river's name and first 2 digits of station code, which are river specific. In those cases the station with the largest catchment (*i.e.* most downstream) was selected. We additionally included a set of 41 Welsh catchments > 25km^2 to assure spatial fit with the LUCI model.

The provided daily flows (in $m^3\ s^{-1}$) were calculated into daily totals and subsequently summed in annual values for the period 1996-2015 and averaged among years. The employed end date of 31-12-2015 matches the used existing model outputs for Grid-to-Grid (Bell *et al.* 2018b) and DECIPHeR (Coxon *et al.* 2019b).
For details see our GitHub code at: https://github.com/EnsemblesTypes/DataExtractionTools, module *ExtractDataNRFA.*

Because of partial overlaps of the associated catchments, model data extractions for these polygons were conducted per single polygon using the ArcGIS Zonal tool, afterwards the results are combined. See GitHub: https://github.com/EnsemblesTypes/DataExtractionTools, modules *Water_Extractions(arcpy)* and *WaterModelDataCombine*.

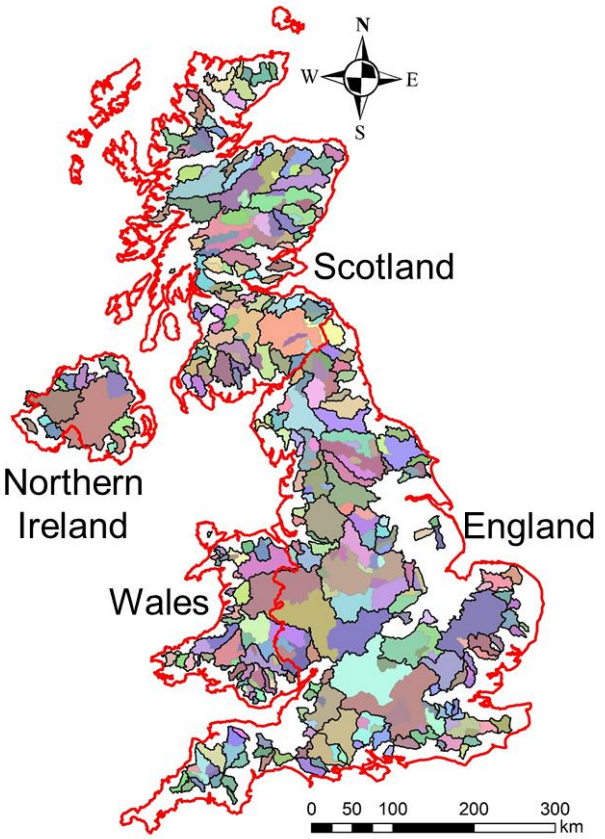

Figure SI-1-1. Catchment coverage of the UK. The 519 catchments used for water validation and ensemble calculations. Smaller catchments overlapping bigger ones are put on top; lines show underlying largest catchment level. Colours are for depiction purpose only. Included are the four countries of the United Kingdom, catchments can cross these borders.

National Forest Estate Sub-compartments for England and Scotland 2019
This validation set provides a close estimate for above ground standing carbon stocks.

The carbon stock validation dataset was retrieved from UK Forest Research (FR) open data with the guidance of Kevin Watts and consists of species inventories in all forest estates in England and Scotland in 2019. The FR open data sources, among others, holds a Sub-Compartment Database (SCDB) and forest stock maps. The goal is to provide an authoritative data source, providing information for forest stock recording, monitoring, analysis and reporting. Through this, the dataset supports decision-making on the whole of the estates. Information from the inventory is used by the Forestry Commission, wider government, industry and the public for economic, environmental and social forest-related decision-making. The total covered area by these estates is 8894 km^2 (889,410 hectares). See Figure SI-2-2. The data can be found at:
data-forestry.opendata.arcgis.com/datasets/3993555ec8124b1e91b55a4a8b84567c_0
data-forestry.opendata.arcgis.com/datasets/1a971b7b3e14439f8481d016f46d99d3_0

In 201,143 present 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 forest layers. Accumulated carbon was calculated per hectare per compartment using the UK species specific accumulated CO_2 sequestration with stand age, with a 12/44 CO_2 to carbon conversion rate woodlandcarboncode.org.uk/images/Spreadsheets/WCC_CarbonCalculationSpreadsheet_Version2.3_12May2020.xlsx (sheet: Biomass Carbon Look Up Table), linking to the full list of the 135 UK tree species (forestresearch.gov.uk/documents/2783/PF2011_Tree_Species.pdf).

For all 201,143 sub-compartments, all mentioned species could, after an initial cross-check, relate to the species in the UK tree species list. For each compartment for each of the three layers, the unique combination of stand age, thinning regime and species was searched in the Biomass Carbon Look Up Table and area converted to carbon per hectare. Subsequently the three layers were summed to get an overall carbon estimate per compartment, as carbon per hectare. For details see our GitHub code at: https://github.com/EnsemblesTypes/DataExtractionTools, module *CarbonCompartments*, with the actual search and area conversions in Lines 75-83, 109 and 116; most remaining code is to initiate the input data (up to line 34) and detect potential table errors and inconsistencies in stand ages ('Weird_list'), and correct those where possible. Finally, only 4 sub-compartments were removed because of unresolvable inconsistencies in stand ages.

Subsequently, sub-compartments were spatially joined into 2078 polygons of forest with a large size range (mean: 409-hectares. median 34.1-hectares, ± 2569), summing the carbon estimates of the different units within. Afterwards model prediction results are extracted using codes on GitHub: https://github.com/EnsemblesTypes/DataExtractionTools, module *Carbon_Extractions(arcpy)*.

Though not being measurements per se, these data are as close as feasible to such large scale estimates and similar as used for forest plots in *e.g.* Africa (Willcock *et al.* 2014; Avitabile *et al.* 2016).

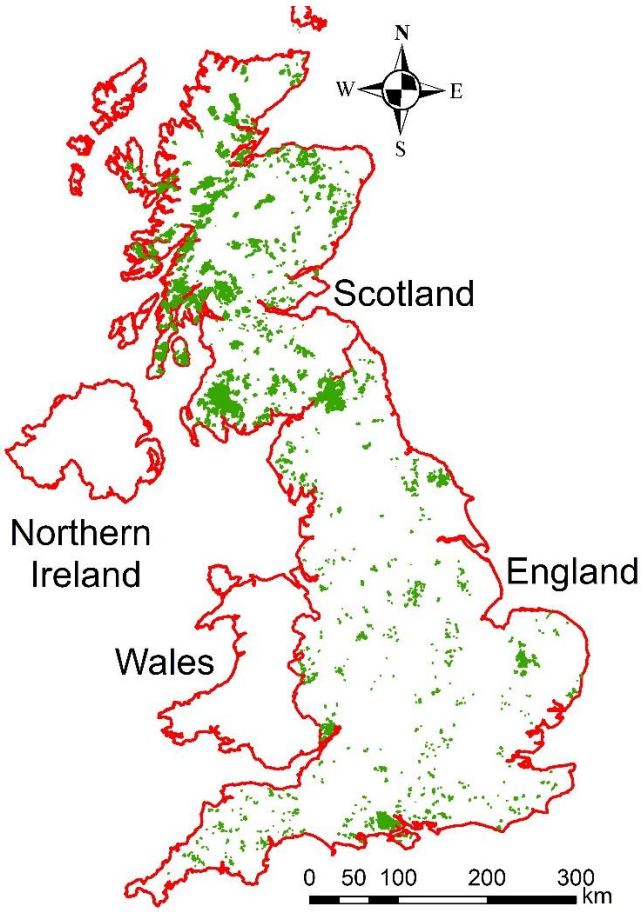

Figure SI-1-2. Forest Estate Locations in England and Scotland provided by Forest Research. (data-forestry.opendata.arcgis.com/). Included are the UK country divisions.

SI-1-3 Ensemble calculations and comparisons

There is a knowledge gap as to which ensemble approach is best for ES modelling to support decision-making: should ES model ensembles be produced by unweighted averaging, or weighting using either trained or untrained approaches? Here, we address this by implementing a range of ensemble methods to evaluate which produces the most accurate predictions against validation data.

Many approaches exist to calculate ensembles (Araújo & New 2007; Marmion *et al.* 2009; Dormann *et al.* 2018) – we compare 10 ways of doing this here. Broadly, *weighted* ensemble approaches fall into two categories: 'untrained' where ensembles are obtained without any validation data; and 'trained' where validation data are incorporated. *Unweighted* ensembles (mean and median) have been most generally used, in both species distribution (Araújo & New 2007; Marmion *et al.* 2009) - and climate modelling (Refsgaard *et al.* 2014).

However, weighting unequally among model outputs is suggested as a possible approach to acquire better accuracy (Marmion *et al.* 2009; Knutti *et al.* 2013; Dormann *et al.* 2018; Willcock *et al.* 2020). Still, as yet there are no guidelines to combine ES model outputs of various forms and scales using weighting approaches aiming to reduce uncertainty among models by weighting them unequally (Marmion *et al.* 2009; Dormann *et al.* 2018). Based on ideas developed in other fields we test several approaches of determining weights. Especially where validation data are not available, the consensus among models, can be used to weight their individual contribution to the ensemble value. This approach follows the logic that models whose output values differ more from those of the other models (*i.e.* are more distinct) are more likely to be incorrect. We restrict ourselves to weighting methods feasible for a wide group of users with standard statistics packages available – such as R, SPSS, SAS, Matlab – focussing on consensus. The selected weighting types should be feasible to replicate without much coding ability.

We use examples of three type of methods to determine weights:

1. Deterministic consensus (*i.e.* determining a consensus axis: Marmion *et al.* 2009; Grenouillet *et al.* 2011);
2. Iterated consensus (through regression and cross-validation: Araújo & New 2007; Dormann *et al.* 2018);
3. Attribute-based (*e.g.* Marmion *et al.* 2009; Masson & Knutti 2011; Englund *et al.* 2017; Willcock *et al.* 2019; Brun *et al.* 2020).

Deterministic consensus among models can be calculated using several approaches, including the fit to a common consensus axis such as through Principal Components Analysis, or weighting by correlation coefficients (En-3 & En-4; ensemble numbering follows Table 1 main text and below). However, through using structured trial-and-error, iterative approaches might more accurately quantify consensus among models by reducing uncertainty. We will test here whether that is true using examples of two regression techniques: between the individual models and the median (En-5) and leave-one-out cross-validation (En-6), as well as trained regression between validation data and modelled values, (En-10). By contrast, one might a priori place value on particular model attribute and use these to create weights (En-7, En-8 & En-9). For example, by up- or down-weighting more distinct model types emphasising models that may contain processes not captured in others, or by penalising those models that go against the convention (En-8), or penalising models that run at coarser spatial resolutions (since smaller grid sizes are deemed more useful for decision-making: En-7; Willcock *et al.* 2016). The attribute matrix we developed for the latter is explained in SI-1-4. As a last category, when validation data are available, it is possible to weight individual models by model accuracy (En-9).

Most employed weighting techniques are from species distribution modelling ensembles (*e.g.* from Araújo & New 2007, Marmion *et al.* 2009, Brun *et al.* 2020, Liu & Liang 2020), adapted to our goals with feasibility as important selection criterion. We refer to sources of these both here and in the main text. We do not claim to be exhaustive in any form, many other techniques to determine weight could exist, within the same categories as used here or coming from different, *e.g.* a more mathematical angles

(Dormann *et al.* 2018). For example, it is less feasible to use more Bayesian oriented ensemble techniques from species distribution modelling – such as BIOMOD or GAM or similar applications or approaches; Thuiller *et al.* 2009; 2019–, in which most work has been conducted. Such models generally deal with binary data (presence vs non-presence) whereas ES uses continuous data; hence, available packages do not fulfil our goals. The development of suitable mathematical approaches and codes to do so is beyond our goal of providing feasible methods for a wider audience of policymakers and other stakeholders (Willcock *et al.* 2016).

All our ensemble calculations followed the same procedure, we show the flow among codes used in Figure SI-1-3; the ensembles are explained in SI-1-3b.

To generate uncertainty estimates allowing statistical comparison with the models and among ensembles we used a bagging approach (see Dormann *et al.* 2018) in which we jack-knifed with 50% of the spatial data polygons for 250 runs following such suggestions in Araújo & New (2007) and Refsgaard *et al.* (2014). To assure like-for-like comparisons, per run all model and ensemble calculations and their comparisons were calculated on the same set of jack-knifed spatial data points. Afterwards mean averages and standard deviations were calculated among runs. Table 1 (main text) presents our 10 ensemble types as equations, below they are textually explained (SI-1-3b). All mentioned codes are written in Matlab v7.0.14.739 with statistical and parallel toolboxes available. "*Modules*" are written code by ourselves, "*daughter functions*" were written within the modules and only available from within modules, "*tools*" refer to codes within the Matlab package and are not generated by the authors. For the latter full function descriptions are available within Matlab help and online via mathworks.com.

SI-1-3a. Code flow

In our GitHub account https://github.com/EnsemblesTypes/EnsemblesTypes, the steering module is *Ensembles_Analyses*, which needs to be called with its target validation set, mode and maximum numbers or runs. The full flow of codes is depicted in Figure SI-1-3, with arrows the module call flow. After parameter and definition fixations, the actual runs are performed through main module *TheRuns*. Fixed parameter values, such as thresholds, iterations, and model and validator sets are stored in *DefintionSet* called upon by the steering module. As well in *DefintionSet* the required model and validation data sets are loaded.

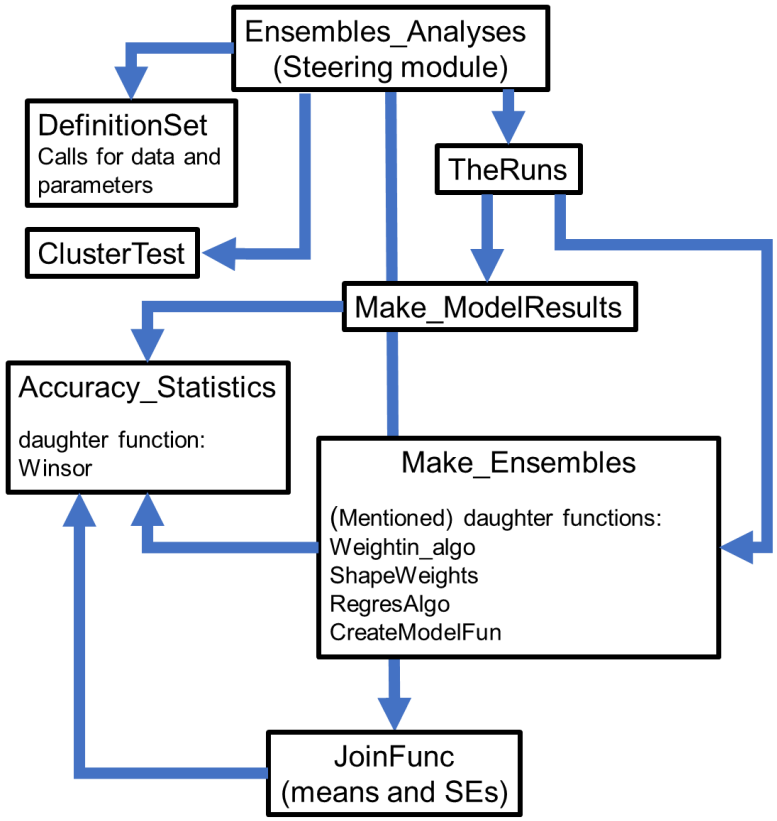

Figure SI-1-3. Module flow to guide to the GitHub deposited codes in https://github.com/EnsemblesTypes/EnsemblesTypes. We cite actual module names. Arrows indicate the direction of call, *e.g.* the steering module calls *DefinitionSet*, *ClusterTest* at the start of the model, *TheRuns* for performing all jack-knifed model and ensemble assessments and *JoinFunc* to join the runs into means and standard deviations among runs. Outputs flow the opposite directions. After selecting a jack-knife proportion of the dataset, set at 50% (100% for SI-3) in *TheRuns*. Model accuracy was run through the *Make_ModelResults* module; subsequently ensembles were generated and tested for accuracy in the *Make_Ensembles* module. Along with the steering information, both modules output accuracy measures, as well as the ensemble module outputting weights. Outputs were stored per run as labelled swap files to be collated by the *JoinFunc* module.

The accuracy calculations were run by the module *Accuracy_statistics*, which is called upon for all accuracy calculations as well as outputs ensemble polygon values for depictions (see SI-3). It includes a full normalisation to the 2.5% and 97.5% percentile prior to accuracy assessment for all models and ensembles each time the module is called upon (daughter function *Winsor*), hence all accuracy calculations were done over the full 0-1 range for both models, ensemble and the comparator. This also assures scales are identical among runs, as normalisation is done within runs. Missing values were reciprocally removed from both model and the comparator per combination – so the number of datapoints among models can be different due to missing information (Table SI-1-1); no missing values were present in the ensembles since that would imply none of the models contained valid values. The normalised ensemble output per validation polygon calculated generated in *Accuracy_statistics* was outputted for use in spatial mapping (SI-3).

After conducting all runs, they were collated through the *JoinFunc* module, generating among runs (mean) averages and standard deviations for accuracy, weights and especially bilateral differences, calculated as $\left[\frac{Accuracy_j}{Accuracy_i}\right]$, with *i* and *j* different ensemble types, or the mean among models (see below). Bilateral differences were as well calculated per run and averaged afterwards providing a mean and standard deviation of bilateral differences. These standard deviations among runs are the error bars used

in Figure 2 (main text, with *i* above the mean ensemble) and Figure SI-2-1 and Figure SI-2-2 (with *i* above the median ensemble).

SI-1-3b. Calculations per run, including all ensembles

1) We selected 50% of data points, in case of unequal data points (water) this was the ceiling. This selection would remain for all model and ensemble calculations. The remaining 50% of data-points was stored to be used in the trained ensembles procedures (*TheRuns*: lines 7:12)
2) For the models, accuracy was calculated over this 50% of datapoints between model output (see SI-1-1) and comparator (validation data sets, see SI-1-2) and stored. This was run through the *Make_ModelResults* module. "*A randomly chosen individual model*" (see main text) was represented by the mean accuracy among models per run. This mean value was afterwards averaged over all runs (see *JoinFunc* module).
3) Progressing to the *Make_Ensembles* module, we firstly calculated the **unweighted ensembles** by taking per data-point the mean (*En-1* in Table 1 main text) respectively the median (*En-2* in Table 1 main text), *i.e.* a datapoint represents one validation polygon. Accuracy of these ensembles were assessed against the comparator validation data set in the *Accuracy_statistics* module. See for unweighted method discussions *e.g.* Marmion *et al.* (2009), Grenouillet *et al.* (2011), Refsgaard *et al.* 2014 and Willcock *et al.* (2020).
4) Subsequently, we calculated the **untrained weighted ensembles** simulating the situation without any or reliable validation data present. Ensembles are of the general form:

$$E_{(x)} = \sum_i^n \left(\frac{\omega_i}{\sum_i^n \omega_i} \times Y_i \right)_{(x)} \qquad \text{Eq. SI-1-2}$$

with positive weights ω_i for model *i* of validation polygon *x*, weights ω_i are normalised to sum to 1, *Y* the modelled values for *i* per polygon (step 3), and *n* the total number of models per service.

This implies every model has one weight assigned which will be used to multiply all its containing datapoints with. The difference in approaches is how the weights are generated. After assessing weights Eq. SI-1-2 is performed using the daughter function *Weightin_algo*, including normalisation of weights to sum to 1 (daughter function *ShapeWeights*). The resulting ensemble values are tested against the comparator validation data set in the *Accuracy_statistics* module. Weights per runs are stored and averaged in *JoinFunc* for use in spatial mapping (SI-3).

We calculated the following untrained weighted ensembles (numbering follows Table 1 main text):

En-3 PCA as consensus axis is a deterministic consensus method. The suggestion of using PCA's comes from Marmion *et al.* (2009) and Grenouillet *et al.* (2011). Deterministic here means that the result is an inherent property of the dataset, *i.e.* the statistical outcome is identical given the same dataset. Principal components were calculated using the *princomp*-tool, the weights per model *i* outputted to Eq. SI-1-2 were the loadings to the first –main– pca axis. So models with the better correlation to the consensus axis are assigned higher weights.

En-4 The correlation coefficient method is our second deterministic consensus method. Here we calculated the full [model × model] correlation matrix using the *corrcoef*-tool. Following the weight per model was the mean correlation of that individual model with all other models, not including itself. Hence the higher general correlation to the other models, the more weight a model has. This technique was developed to have a second deterministic approach using a consensus axis different than under (En-3) and can be seen as further way to minimise variance among models (Dormann *et al.* 2018).

En-5 Regression to the median is our first iterative consensus method using log-likelihood regression (Dormann *et al.* 2018). Using multivariate regression we assess weights such that the summed results maximises the explanation of an comparator. The resulting regression coefficients are used as weights. In this case the comparator is the median ensemble (En-2), asking which contribution of models would be most closely result to the median – note this

can't be done with the mean, since on definition means are generated by equal weighting –. The regression contains no constant, hence it can be represented as: [$E_{ii} \sim \omega_1 Y_1 + \omega_2 Y_2 \ldots + \omega_n Y_n$].
Full regression equations can be found in the daughter function *RegresAlgo* and her linked daughter *CreateModelFun*.

This method is iterative, parameter space is step-wise systematically explored improving the maximum log-likelihood until convergence is reached, *i.e.* no better solutions is found. Theoretically, different outcomes would be possible by redoing the calculation. However, the used Matlab tools are of such quality, including multiple replications, that noticeable difference are absent – which could be caused by local but not absolute maximum log-likelihood in parameter space. Multi-variate regression to the median was done using the *nlmefit*-tool, maximising log-likelihood with 200 iterations: repeating the regression 200 times), an output tolerance of 1.0000e^{-4} and naïve priors (all $\omega_i = \frac{1}{n}$ at the start). We point to the tool help function (mathworks.com/help/stats/nlmefit.html), its called tools and explanations there for the exact algorithm. The resulting regression coefficients (ω_i) per model were the weights that were used in Eq. SI-1-2.

En-6 Exhaustive leave-one-out cross-validation is our second iterative consensus method following direct recommendation in Dormann *et al.* (2018), with the difference of not omitting sets of data points but entire models one-by-one. As for (En-5) this is done using a no constant multi-variation regression with the same *nlmefit*-tool, with the same settings and naïve priors. However, in this method we loop through the model outputs. One-by-one, a regression is performed using a single model output as comparator and the remaining model outputs as explanatory variables. For model 1 such would be the regression representation [$Y_1 \sim \omega_2 Y_2 + \omega_3 Y_3 \ldots + \omega_n Y_n$]. The regression coefficients (ω_i) are stored as consensus weights. After looping through all models – 9 water models or 10 carbon models, or less for our African comparison –, the mean is taken of all regression coefficients per model as weights (excluding itself), *i.e.* this represents the weights that would generate the highest mean consensus with all models. The resulting mean consensus per model were the weights that were used in Eq. SI-1-2.

En-7 Models that are generated on smaller scales (*i.e.* with smaller grid cells) could be more accurate since the information per cell could better represent the local situation whereas larger grid cells could be more averaged across larger areas (Willcock *et al.* 2019; 2020). To include this we generated an ensemble in which we penalised model outputs that are generated at coarser spatial resolutions (Willcock *et al.* 2016). The weights taken were: $\omega_i = \frac{1}{\log_{10}(\text{spatial resolution}_i)}$, for which the resulting weights were normalised afterwards to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$. These weights were the weights that were used in Eq. SI-1-2. Spatial resolution assessment was done in the *DefenitionSet* module at the start as these weights are not run specific. Due to jack-knifing data-points, the resulting accuracy was different among runs to some small extent.

En-8 One might *a priori* place value on particular model characteristics and use these to create weights (Masson & Knutti 2011; Englund *et al.* 2017; Willcock *et al.* 2019). For example, up- or down-weighting more distinct model types emphasising models that may contain processes not captured in others, or by penalising those models that go against the convention (Grenouillet *et al.* 2011). Attribute weighting could be done for many attributes, which are largely arbitrary in use. To not focus on one attribute but many at the same them we choose an overall attribute assessment into groups based on 17 categories per model. The grouping statistic used is a pairwise Spearman's rank correlation among binary classifications (SI-1-4). The goal was to generate 4 or 5 groups of different amounts of models with similar

attributes. The attributes, the resulting trees and model grouping are explained and depicted in SI-1-4. The output variable we included in our weighting is the *distinctiveness factor*, representing how proportional representation of a group of models among all models (see Table 1 main text). By upweighting distinctiveness, models in minority groups (g) are assigned a higher weight compared to majority groups as $\omega_i = \left(\frac{n^g}{n}\right)$ when upweighted with $n^g = i \in$ group g and n the number of models. Alternatively, consensus is sought so distinctiveness is downweighted, assigning higher weights to majority groups as $\omega_i = \left(\frac{n}{n^g}\right)$. In all cases resulting weights are normalised afterwards to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$. These weights were the weights that were used in Eq. SI-1-2. Grouping is done in the *ClusterTest* module, called at the start from the steering module as these weights are no run specific. Due to jack-knifing data-points the resulting accuracy is different among runs though.

5) Subsequently, we calculated **the trained weighted ensembles** simulating the situation in which validation data are partly present or present for a very similar area. So this approximates the standard Species distribution modelling techniques were data is split in a training and a testing set following Marmion *et al.* (2009), Thuiller *et al.* (2009; 2019) and Djengdoh *et al.* (2020). In these ensembles the 50% data points selected in step 1 and their accompanying validation data for the same points is to train the model. Subsequently, the resulting weights are multiplied with the second set of 50% of the data following Eq. SI-1-2, so the part of data that was not used for training. This ensemble is than tested for accuracy against the validation comparator belonging to this second set of data-points. By doing this in a jack-knife loop, a good representation of all possible combinations of selected data is provided and accuracy by chance is avoided.

En-9 Accuracy weighted ensembles are based on the model accuracies as calculated in step 2 over the 50% of datapoints assigned to all ensembles. See the supplementary materials of Willcock *et al.* (2019) for a preliminary try with this with ES data. This weighting is similar to AUC weighting as suggested in Marmion *et al.* (2009), Crossman *et al.* (2012), Grenouillet *et al.* (2011) and Dormann *et al.* (2018). These "trained" accuracies per model (either $D^{\downarrow}$ or Spearman ρ) are, after normalisation to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$, used as weights ($\omega_i$) in Eq. SI-1-2, with as Y_i the second set of datapoints not used in the training. Accuracy is assessed against the corresponding set of comparator validation data points not used in the training.

En-10 Log-likelihood regression is identical to (En-5), with the difference that models are not regressed against their median but against their corresponding validation data points represented as: $[V \sim \omega_1 Y_1 + \omega_2 Y_2 \ldots + \omega_n Y_n]$, with V the validator. The resulting regression coefficients (ω_i) per model are used, after normalisation to sum to 1 $\left(\frac{\omega_i}{\sum_i^n \omega_i}\right)$, as weights in Eq. SI-1-2 with as *Yi* the second set of datapoints not used in the training. Accuracy is assessed against the corresponding set of comparator validation data points not used in the training.

SI-1-4 Model attribute weighting (distinctiveness)

Attribute weighting could be done for many attributes following Marmion *et al.* (2009), Masson & Knutti (2011), Englund *et al.* (2017), Willcock *et al.* (2019) and Brun *et al.* (2020). To not focus on one attribute but many at the same time we made an overall attribute assessment into groups based on 17 categories per model with attributes either being true or false. The goal was to generate 4 or 5 groups of different amounts of models with similar attributes. Below we first show first the classification tables and following the resulting trees with the used model groupings.

Table. SI-1-5. Binary classification (1 = true, 0= false) of model outputs for water supply in these categories used in pairwise Spearman's ranked correlation distance among binary classifications (Matlab *pdist*-tool), so *e.g.* InVest water yield model is parameterized as being *true* for LCM2015 (Rowland *et al.* 2017) and so false for hydrological units (NOAH: Coxon *et al.* 2019a; HRU's: Bell *et al.* 2018a), contains measured instead of modelled Climate data (Tanguy *et al.* 2019 vs. Fick *et al.* 2017), is specific to the UK with inputs, has a below 1-hectare grid, contains no time-series or within year units (such as months) and is process based. †Contains no LCM; ‡No climate data. Model terminology follows SI-1-1.

		InVest Water Yield	Growth Days†	WaterWorld	$-Transfer‡	Grid-to-Grid	Aqueduct	DECIPHeR	LPJ-GUESS	LUCI
Land cover map employed.	LCM2015	1	0	0	1	0	0	0	0	1
	Corine	0	0	0	0	0	0	0	0	0
	Modis	0	0	1	0	0	0	0	0	0
	Mixed covers	0	0	0	0	0	0	0	1	0
	NOAH v. 3.3 land surface model	0	0	0	0	0	1	0	0	0
	Hydrological Research Units	0	0	0	0	1	0	1	0	0
Climate	Measured	1	0	0	0	1	1	1	1	1
	Modelled (WorldClim)	0	1	1	0	0	0	0	0	0
Model specificity (World = both 0)	Europe	0	0	0	0	0	0	0	0	0
	UK	1	0	0	1	1	0	1	0	1
Output Grid (>1km = all 0)	Below 1-Hectare	1	0	0	1	0	0	1	0	1
	Per hectare	0	0	0	0	0	0	0	0	0
	Per 1-km^2	0	1	1	0	1	0	0	0	0
Time Series over years	Yes/no	0	0	0	0	1	0	1	1	0
Within year units (months/ days, 0 = annual sum)	Yes/no	0	1	1	0	1	0	1	0	0
Model Type (Look-up = both 0)	Process based	1	0	1	0	1	0	1	1	1
	Deterministic	0	1	0	0	0	1	0	0	0

Table. SI-1-6. Binary classification (1 = true, 0= false) of model outputs for above ground carbon stocks in 17 categories used in pairwise Spearman's ranked correlation distance among binary classifications (Matlab *pdist*-tool), so *e.g.* InVest carbon model is parameterized being *true* for LCM2015 (Rowland *et al.* 2017) and so false for hydrological units (NOAH: Coxon *et al.* 2019a; HRU's: Bell *et al.* 2018a), contains no Climate data (both 0), has no UK/Europe specific inputs (being CDIAC based: Ruesch & Gibbs 2008), has a below 1-hectare grid, contains no time-series or within year units (such as months) and is look-up table based. Note specifically for carbon, being standing stocks, not all categories are employed as not being present in any of the models (*e.g.* hydrological research units, Within year units) but maintained to allow synchronising with Table SI-1-1, similarly for look-up table based models several categories are redundant (*e.g.* time-series) but no for process models. Model terminology follows SI-1-1.

		InVest Carbon	**$-Transfer**	**ARIES**	**Henrys**	**Barredo**	**Kinder-mann**	**Density**	**FC-Inventory**	**LPJ-GUESS**	**LUCI**
Land cover map employed.	LCM2015	1	1	0	1	0	0	0	1	0	1
	Corine	0	0	1	0	1	0	1	0	0	0
	Modis	0	0	0	0	0	1	0	0	0	0
	Mixed covers	0	0	0	0	0	0	0	0	1	0
	NOAH v. 3.3 land surface model	0	0	0	0	0	0	0	0	0	0
	Hydrological Research Units	0	0	0	0	0	0	0	0	0	0
Climate	Measured	0	0	0	0	0	0	0	0	1	0
	Modelled (WorldClim)	0	0	0	0	0	0	0	0	0	0
Model specificity (World = both 0)	Europe	0	0	0	0	0	0	1	0	0	0
	UK	0	0	0	1	0	0	0	1	0	1
Output Grid (>1km = all 0)	Below 1-Hectare	1	1	0	0	0	0	1	1	0	1
	Per hectare	0	0	1	0	0	0	0	0	0	0
	Per 1-km^2	0	0	0	1	1	1	0	0	0	0
Time Series over years	Yes/no	0	0	0	0	0	0	0	0	1	0
Within year units (months/days)	Yes/no	0	0	0	0	0	0	0	0	0	0
Model Type (Look-up = both 0)	Process based	0	0	0	0	0	0	0	0	1	0
	Deterministic	0	0	0	0	0	1	1	0	0	0

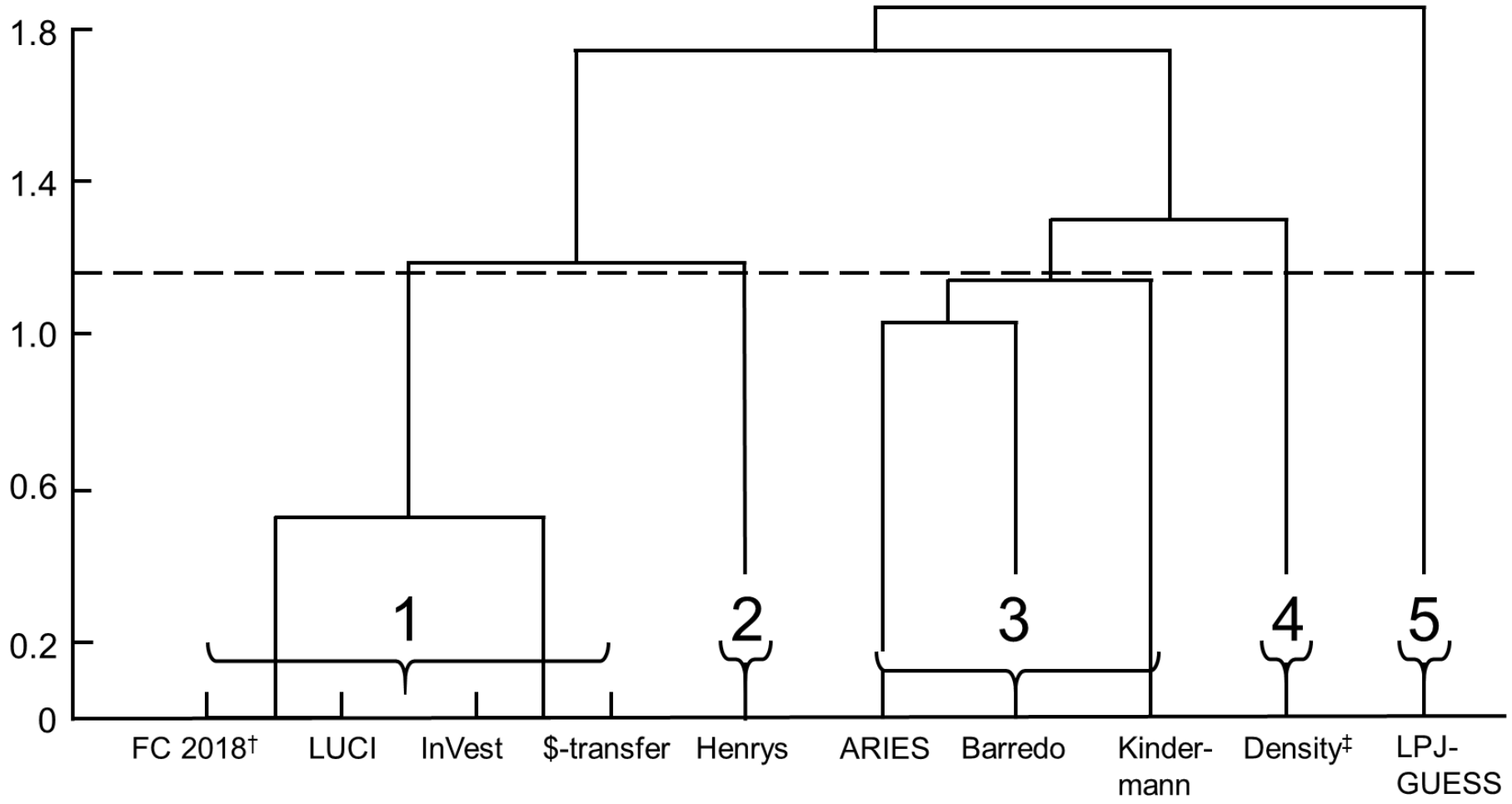

Figure SI-1-4. Distinctiveness in 5 groups of carbon stock models based on pairwise Spearman's rank correlations among binary classifications in 17 categories (Table SI-1-5) with manually set threshold (dotted line), without among category weighting (all differences weight equal). Model terminology follows SI-1.

Upweighting is done as $\beta_m = \frac{\#\, models}{\#\, models\ in\ group\ i}$, downweighting as [1./upweight value]. Both subsequently normalised as: $\beta'_m = \frac{\beta_m}{\sum_1^{\#\, models} \beta_m}$.

† FC 2018 = National Forest Inventory Woodland GB 2018; ‡ Density = Tree Cover Density 2015.

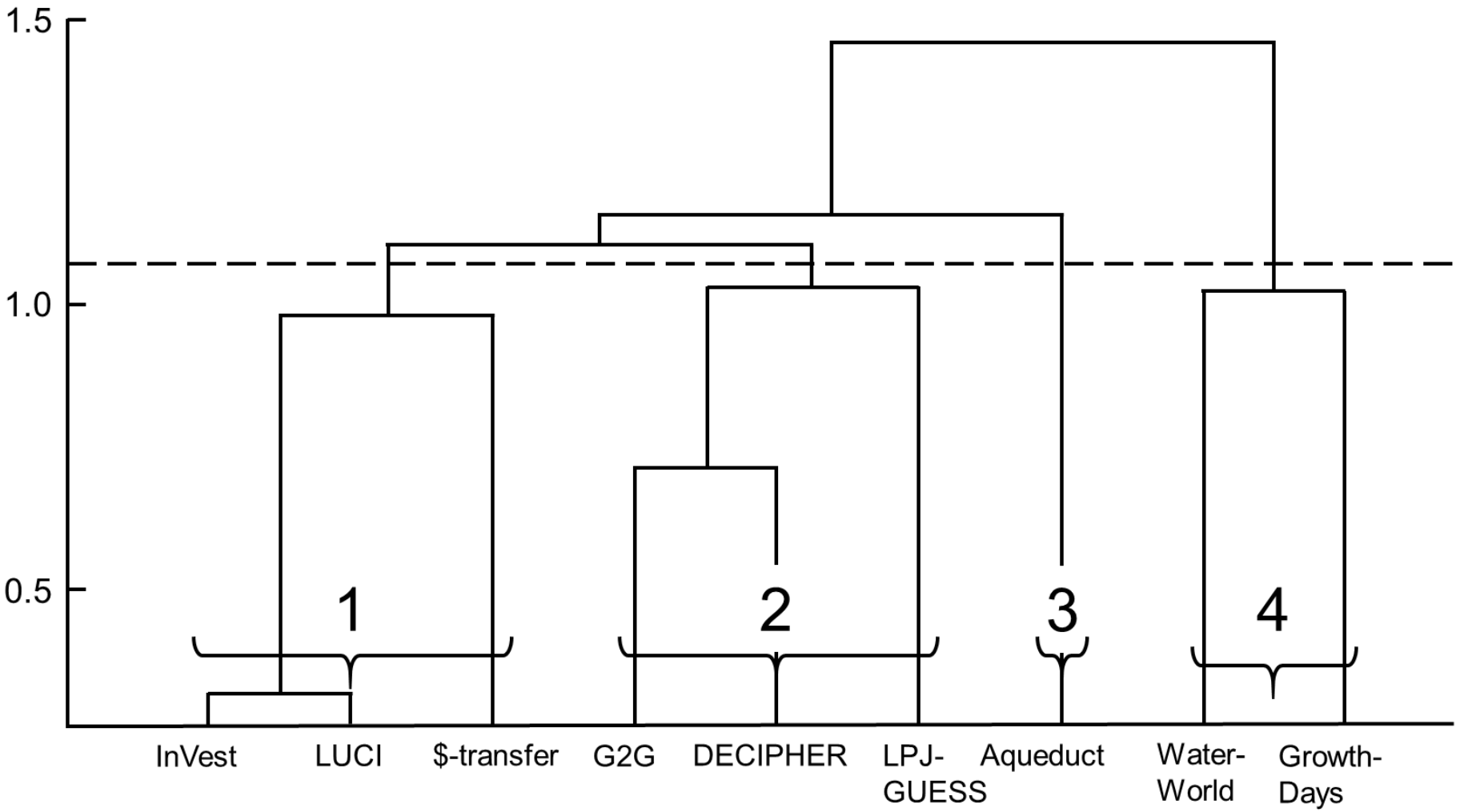

Figure SI-1-5. Distinctiveness in 4 groups of water supply models based on pairwise Spearman's rank correlations among binary classifications in 17 categories (Table SI-1-6) with manually set threshold (dotted line), without among category weighting (all differences weight equal). Model terminology follows SI-1.

Upweighting is done as $\beta_m = \frac{\#\ models}{\#\ models\ in\ group\ i}$, downweighting as [1./upweight value]. Both subsequently normalised as: $\beta'_m = \frac{\beta_m}{\sum_1^{\#\ models} \beta_m}$

SI-2 Additional analyses

Figure legends, figures below

Figure SI-2-1. Accuracy of above ground carbon stock ensembles (10 models) against inventories in Forest Research estates in England and Scotland (a and b), and of different water supply ensembles (9 models) against NRFA measurements 1996-2015 for water supply per hectare (c and d). For definitions and calculations of the different ensembles see main text Table 1. Shown is the average accuracy of 250 bootstrap runs with 50% of the dataset (N ≈ (1598/2=799) per bootstrap). Vertical dashed line indicates the reference unweighted median-averaged ensemble (black dot, 'median ensemble'). Error bars indicate standard deviation among runs in proportional difference to the median ensemble, calculated per bootstrap run as the difference in accuracy *with* the median ensemble divided by the accuracy *of* the median ensemble. The (not shown) Coefficient of Variation among bootstraps for the median carbon ensemble is 4% and 1%, for ρ and $D^{\downarrow}$ respectively, and 1 % and 2% for water. **Blue** coloured ensembles accuracies are significantly higher than the median ensemble (Bonferroni corrected $\alpha = (0.05/14)$) from the median ensemble; **Red** coloured bars are significantly lower with **Black** dashed bars not different from the median ensemble. This analyses mimics Figure 2 of the main text but with the median-averaged ensemble as reference.

Figure SI-2-2. Accuracy of different carbon stock ensembles of 4 models from Willcock *et al.* (2019) against ForestPlots.net plots (Avitabile *et al.* 2016) as above ground carbon stock per hectare (a and b), accuracy of different water supply ensembles of 6 models from Willcock *et al.* (2019) against GRDC recorded flows per weir (bafg.de/GRDC) as flow per ha catchment (c and d). For definitions and calculations of the different ensembles see main text Table 1. Shown is the average accuracy of 250 bootstrap runs with 50% of the dataset (N ≈ (147/2 = 74) and N ≈ (512/2 = 256) per bootstrap respectively for carbon and water. Vertical dashed line indicates the reference unweighted mean-averaged ensemble (black dot, 'mean ensemble'). Error bars indicate standard deviation among runs in proportional difference to the mean ensemble, calculated per bootstrap run as the difference in accuracy *with* the mean ensemble divided by the accuracy *of* the mean ensemble. The (not shown) Coefficient of Variation among bootstraps for the mean carbon ensemble is 15% and 3%, for ρ and $D^{\downarrow}$ respectively, and 3% and 1% for water. **Blue** coloured ensembles accuracies are significantly higher than the mean ensemble (Bonferroni corrected $\alpha = (0.05/14)$) from the mean ensemble; **Red** coloured bars are significantly lower with **Black** dashed bars not different from the mean ensemble.

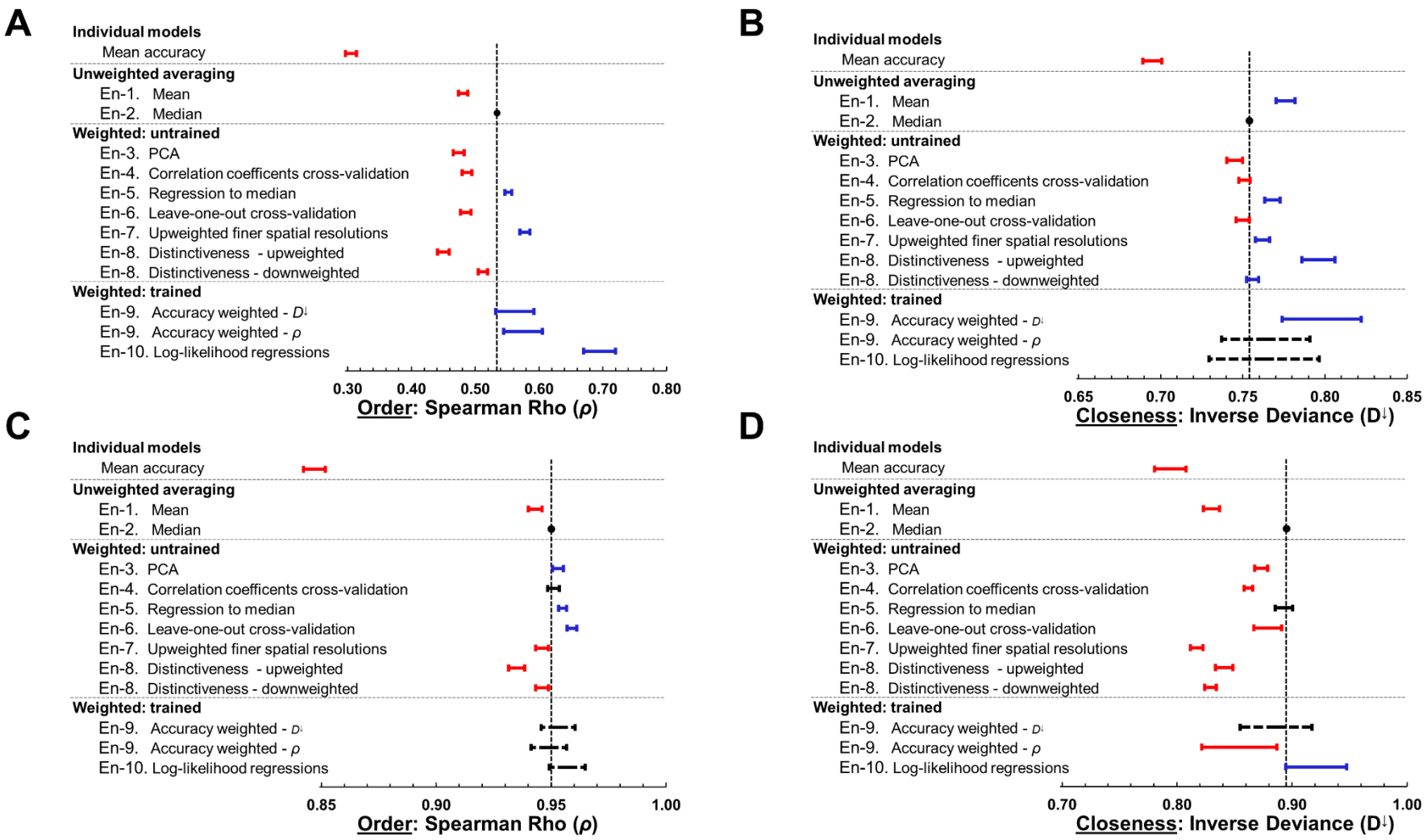
A
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
0.30 0.40 0.50 0.60 0.70 0.80
Order: Spearman Rho (ρ)
B
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
0.65 0.70 0.75 0.80 0.85
Closeness: Inverse Deviance (D↓)
C
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
0.85 0.90 0.95 1.00
Order: Spearman Rho (ρ)
D
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
0.70 0.80 0.90 1.00
Closeness: Inverse Deviance (D↓)

Figure SI-2-1

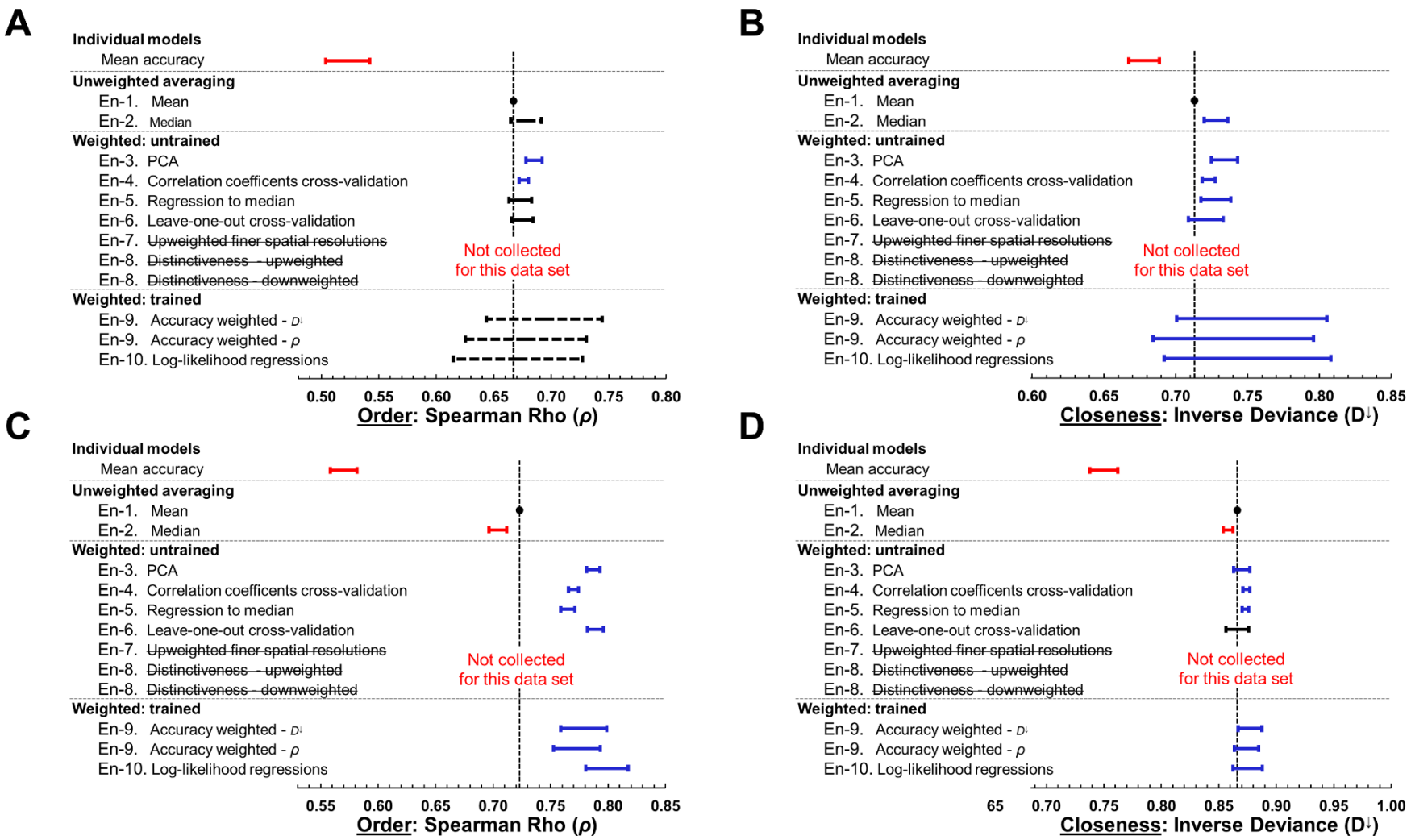
A
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
Not collected for this data set
0.50 0.55 0.60 0.65 0.70 0.75 0.80
Order: Spearman Rho (ρ)
B
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
Not collected for this data set
0.60 0.65 0.70 0.75 0.80 0.85
Closeness: Inverse Deviance (D↓)
C
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
Not collected for this data set
0.55 0.60 0.65 0.70 0.75 0.80 0.85
Order: Spearman Rho (ρ)
D
Individual models
Mean accuracy
Unweighted averaging
En-1. Mean
En-2. Median
Weighted: untrained
En-3. PCA
En-4. Correlation coefficents cross-validation
En-5. Regression to median
En-6. Leave-one-out cross-validation
En-7. Upweighted finer spatial resolutions
En-8. Distinctiveness - upweighted
En-8. Distinctiveness - downweighted
Weighted: trained
En-9. Accuracy weighted - D↓
En-9. Accuracy weighted - ρ
En-10. Log-likelihood regressions
Not collected for this data set
65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Closeness: Inverse Deviance (D↓)

Figure SI-2-2

SI-3 All spatial ensemble maps

To aid decision making we mapped all our ES ensembles for the UK. These spatial maps are available as gridded tiff-files through https://eidc.ac.uk/ (https://doi.org/10.5285/a9ae773d-b742-4d42-ae42-2b594bae5d38). All calculations mentioned here are performed using ArcGIS 10.7.111.

For all the water supply ensembles, the mean normalised ensemble predictions per ensemble method were mapped as catchment polygons (step 5, N = 519). For all carbon stocks ensembles we mapped these for the full UK as 1-km^2 grid cells –253,802 cells that (partially) contain non-sea land cover–, transferring the per model weights calculated for the forest polygons to the full area. We transferred the weights calculated for the forests since running cross-validation approaches on over 250K data points would not be computable. In this way we include areas beyond the small fraction of the UK used as validation polygons.

Following are depictions of the ensembles (see Table 1 and SI-1-3 for descriptions). Most apparent are the differences in spread of values among ensemble approaches: some ensembles tend more towards the extremes (red and blue colours), whereas others are more regressing the median values. Discussions of these among ensemble spatial differences are beyond the scope of our work with the exception of the variation coefficient among untrained ensembles (see SI-4).

For mapping, we followed the following procedure.

1) We conducted full runs with all data points (519 for water and 2078 forest polygons) to calculate the weights per ensemble approach using the identical codes as for jack-knifed runs (SI-1-3, https://github.com/EnsemblesTypes/EnsemblesTypes).
2) For water, the calculated ensemble values per validation polygon (SI-1-3) were one-to-one copied to ArcGIS for depiction per polygon. Uncertainty among ensembles and models as reported in SI-4 is directly calculated over all catchment polygons, *i.e.* with variation per polygon. Since not all model output cover the full area (see SI-1), for uncertainty we corrected for this by using a Standard Error of Means as $(\frac{\sigma_{(x)}}{\sqrt{n}_{(x)}})$, instead of Standard Deviation (σ), with n the number of models per grid cell x.

For carbon, extrapolating to the full UK area we used an ArcGIS approach:

3) All model outputs, with grid size above 100 × 100 meters (Henrys, Barredo, LPJ-GUESS), were resampled to 100 × 100 meters with their carbon indicator or proxy as per hectare. Smaller grid sizes were left untouched to avoid information loss. By adding no data values all models were extended to the spatial tile exactly including the full UK area and afterwards clipped to an UK non-sea outline polygon, *i.e.* full sea area is set as no data values and all 0-values would be land-based true zeros.
4) All model outputs were normalised to their 0.95% percentile based on the average (μ) and standard deviation (SE) of their non-sea and data area into Y, the normalised model output value, using [$T = \mu + 1.65 \times SE$] followed by [$Y = (\text{if}(Value_{(x)} > T, 1, Value_{(x)}/T))$] for all x = grid cells and *Value* the model output value at that grid cell. We used the ArcGIS Raster Calculator (spatial analyst toolbox) for this.
5) The weights as calculated in step 1 were recorded and transferred to ArcGIS Raster Calculator (spatial analyst toolbox) and multiplied with the respective resampled normalised model outputs into a 25 × 25 meter grid size (ArcGIS environment setting, using mean values when upscaling) into Ensemble values (E) at grid cell x as $E_{(x)} = \sum_i^n \left(\frac{\omega_i}{\sum_i^n \omega_i} \times Y_i\right)_{(x)}$, with Y the normalised model output value of model i at grid cell x, ω_i the weight of model i.
6) The resulting layer at 25-meter resolution was clipped once again to the UK non-sea outline. This results in exact identical sizes among all ensemble maps.

7) Identical to step 4, the ensemble layers are each individually normalised to their 95% percentile.
8) Using the ArcGIS Aggregate tool (spatial analyst toolbox) we aggregated to a 1000 × 1000 meter grid size using the mean across all containing values, setting processing extent and grid size to that of the first reference ensemble generated (the mean ensemble); creating all equal maps of 656 columns and 1212 rows of each grid cell being exact 1000 × 1000 meters.
9) To show spatial uncertainty among ensembles as well to test for drivers of this spatial uncertainty among models and among ensembles (SI-4), we calculated the variation among models and among untrained ensembles for 25-meter grid cell using the standard deviation (Cell Statistics, spatial analyst toolbox), with 25-meter environment settings. Noting that not all model output cover the full area (see SI-1), for uncertainty we corrected for this by using a Standard Error of Means as $(\frac{\sigma_{(x)}}{\sqrt{n}_{(x)}})$, with Standard Deviation (σ), with n the number of models per grid cell x. Afterwards we followed the same procedure as steps 6-8, The reported levels of uncertainty reported in main paper and in SI-4 are of the aggregated 1000 × 1000 meter resolution uncertainty maps.

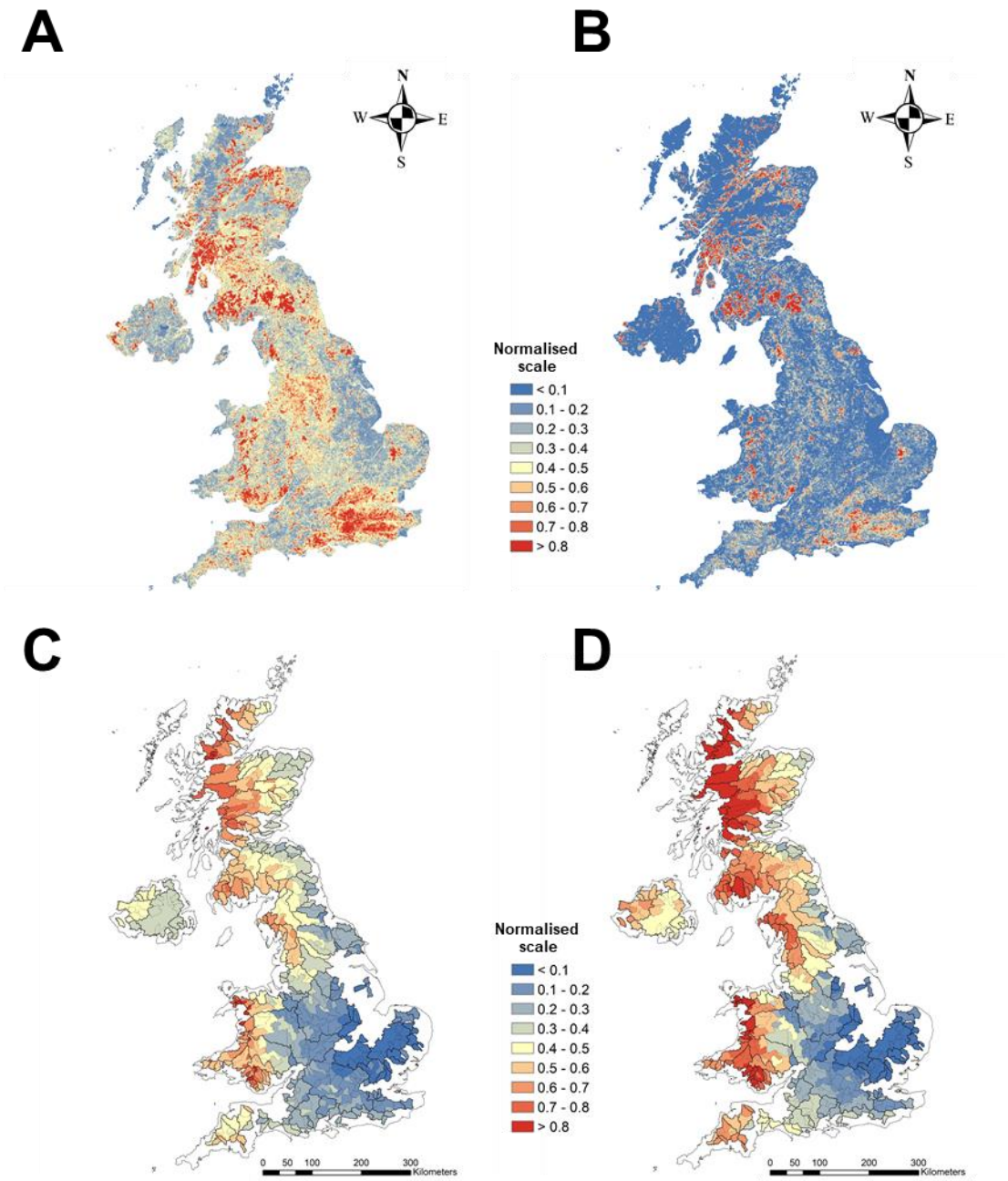

Figure SI-3-1. Unweighted averaging ensembles; **a** carbon mean ensemble (En-1); **b** carbon median ensemble (En-2); **c** water mean ensemble (En-1); **d** water median ensemble (En-2)

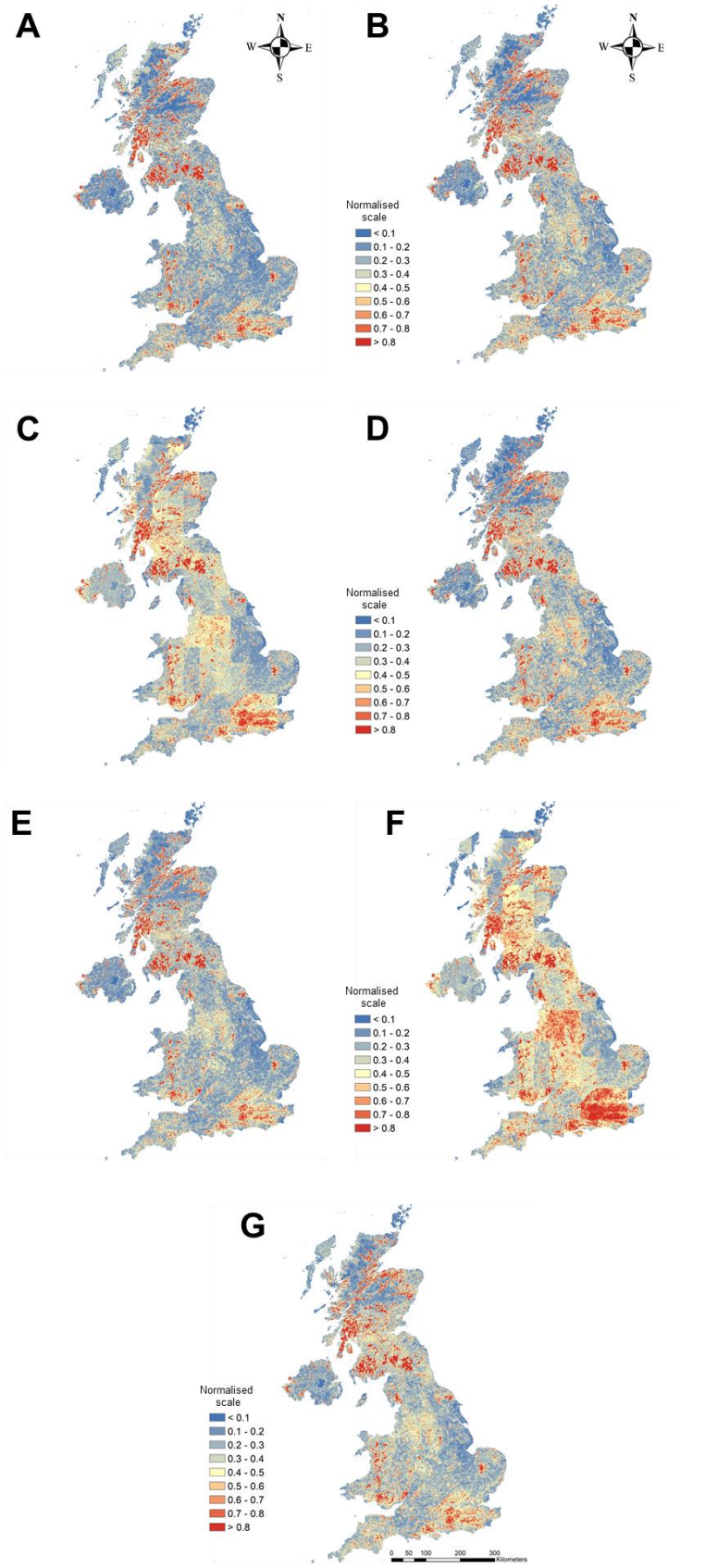

Figure SI-3-2 . Untrained weighted ensembles <u>carbon</u>: **a** PCA (En-3); **b** correlation coefficient (En-4); **c** regression to the median (En-5); **d** Leave-one-out cross-validation (En-6); **e** upweighted finer spatial resolutions (En-7); **f** upweighted distinctiveness (En-8) ; **g** downweighted distinctiveness (En-8).

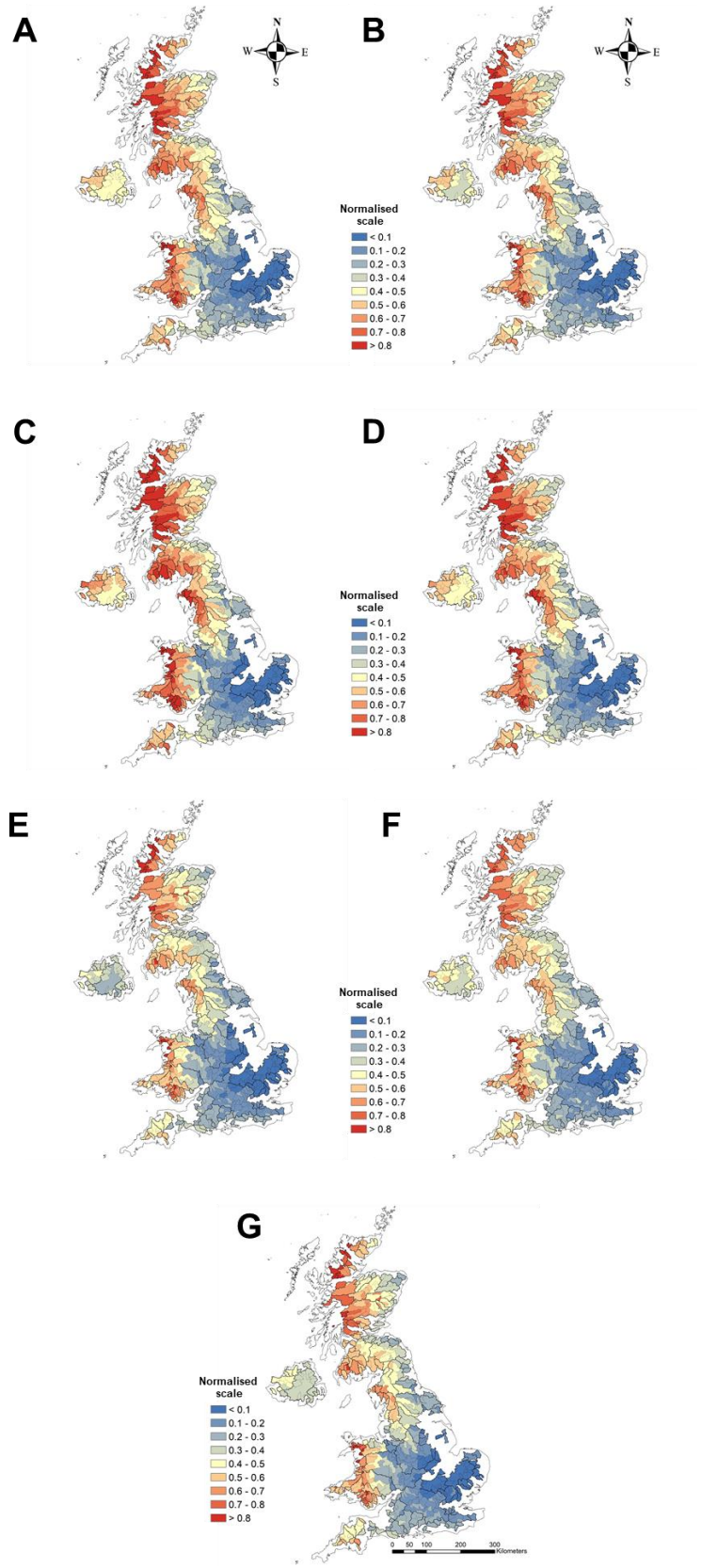

Figure SI-3-3. Untrained weighted ensembles <u>water</u>: **a** PCA (En-3); **b** correlation coefficient (En-4); **c** regression to the median (En-5); **d** Leave-one-out cross-validation(En-6); **e** upweighted finer spatial resolutions (En-7); **f** upweighted distinctiveness (En-8) ; **g** downweighted distinctiveness (En-8).

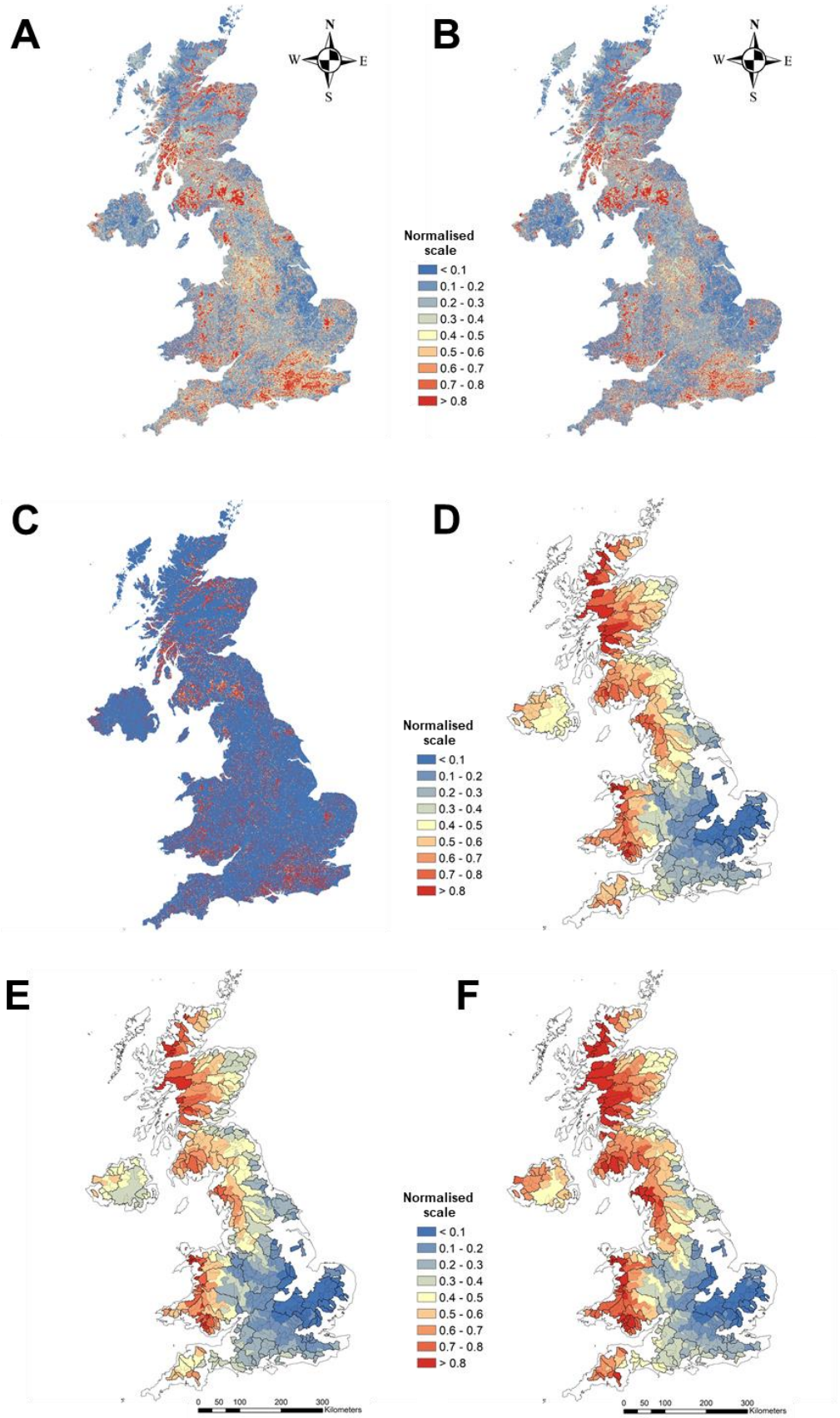

Figure SI-3-4. Trained weighted ensembles; <u>carbon</u>; **a** accuracy weighted inversed deviance carbon ($D^{\downarrow}$, En-9; **b** accuracy weighted Spearman ρ carbon (En-9) **c** log-likelihood regressions carbon (En-10); <u>water</u>; **d** accuracy weighted inversed deviance water ($D^{\downarrow}$, En-9); **e** accuracy weighted Spearman ρ water (En-9) **f** log-likelihood regressions water (En-10).

SI-4 Spatial patterns of uncertainty

Introduction

Decision-makers require information on a wide range of ES, across a variety of temporal and spatial scales, and show both capacity and willingness to engage with the uncertainty associated with such information should these data be made available (McKenzie *et al.* 2014; Willcock *et al.* 2016). For example, in a survey of stakeholders within sub-Saharan Africa, technical experts indicated that ES models with up to a 10% uncertainty were useful to support decision-making (Willcock *et al.* 2016). Whilst environmental management policy can be sensitive to the inclusion of uncertainty (Polasky *et al.* 2011), stakeholders understand that uncertainty is unavoidable and acceptable, provided it is expressed transparently Evans *et al.* 2009; Hamel & Bryant 2017). Furthermore, uncertainty in ES ensembles is a reasonable proxy for ensemble accuracy (Willcock *et al.* 2020), and so communicating where ensemble variation is higher enables decision-makers to understand which areas decisions based on model outputs might be less robust.

Since ecosystems differ in many ways – such as in land-use, in topography and in climate (Schirpke er al. 2013) and ES models process this variation differently, uncertainty in modelled predictions may be increased by such spatial attributes. Many studies have been conducted into the relationships between ES and climate (Prather *et al.* 2013; Nelson er al. 2013) – mostly focusing at change but this would be equally valid for gradients–, similarly land use differences are steering ES production (Lawler *et al.* 2014) as are elevation differences in mountainous terrain (Lavorel et al. 2011), as especially Scotland and Wales contain mountain ranges. Whilst our ES ensembles are UK specific, but containing substantial gradients such as for rainfall (Tanguy *et al.* 2019), such geographic drivers of uncertainty may be transferable to other regions as they may indicate systematic biases within currently available ES models.

Here, we tested the degree of uncertainty in both the individual model and ensemble predictions (as the standard error of the mean of predictions at each location) against 15 putative drivers to identify the causes of this uncertainty (Table SI-4-1). We corrected for spatial autocorrelation with all drivers normalised (Dormann *et al.* 2007; Willcock *et al.* 2019).

Methods

We generated UK-scale maps of spatial variation in the differences among individual models, in terms of the standard error of the mean (SEM) among model outputs (SI-3 step 9; Figure SI-4-1). We did the same to map differences among the untrained ensemble approaches (SI-3 step 9; Figure SI-4-1). For the water supply untrained ensembles, the mean across jack-knifed ensemble predictions per run were mapped as polygons (N = 519). For carbon stocks maps these were mapped as 1-km^2 grid cells. In total this carbon dataset had 253,802 cells that contained non-sea land cover. See SI-3 for all model and ensemble mapping details.

The causes of spatial variation in uncertainty were assessed using 15 putative drivers representative for climate, land use and topography (Table SI-4-1) Drivers included human population size from WorldPop (2018), UK climatic data for annual precipitation (Tanguy *et al.* 2019) and potential evapotranspiration (Zomer *et al.* 2006) as well as modelled climatic data from WorldClim (Fick *et al.* 2017) for seasonality in precipitation, annual mean temperature, temperature coldest month, seasonality in temperature, and temperature range. Furthermore we correlated uncertainty against average elevation from the Copernicus EU-DEM (v1, land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1) at a 25 × 25 meter scale and the ruggedness of the terrain. The latter was estimated as the average slope per 1-km^2 cell employing the ArcGIS 10.7.111 slope-tool (spatial analyst toolbox). We completed the set of putative drivers with land use km^{-2} as proportion of 25 × 25 meter cells from the leading UK land cover map, LCM2015 (Rowland *et al.* 2017; 1600 cells km^{-2}). We calculated per 1-km^2 the proportional cover of (sub)urban, forest, peatlands, agricultural land including improved grassland, and the proportion natural grassland.

We made sure environmental layers and spatial variation among models and ensembles were aggregated to exact 1000 × 1000 meter (1-km^2), with the layers overlapping fully and all sized as 656 cells wide and 1212 cells high, identical to all ensemble layers (see SI-3). For comparison to water supply, the mean of these cells was taken per polygon, with the exception of population size, which was summed. After exporting the layers to Matlab v7.14.0.739 we correlated putative drivers after normalising the driver data following the same procedure as above, one-by-one (*Driver*) with both the variation among models and the variation among untrained ensembles using a SS-type I model with the Matlab tools *LinearModel.fit* and *Anova*:

$[T_{(x)} \sim \beta_0 + \beta_1 Auto_{(x)} + \beta_2 Driver_{(x)} + \varepsilon]$ in which $T_{(x)}$ is either variation among models or ensembles for spatial cell x, with effect sizes β.

We incorporated a correction for potential spatial autocorrelation through inclusion of a covariate (*Auto*) prior to estimating the correlation of the driver of interest, describing relatedness between individual predictions in *T* with the Euclidean distances among centroids of grid cells (Dormann *et al.* 2007; Willcock *et al.* 2020). For water supply, correlations were performed using all 519 data-points at once. For carbon, to avoid spurious findings of significance through having over 250 thousand replicates, we assessed correlations using bootstrapped tranches of N = 519 each for 10,000 runs, from a total cover of 253,802 km^2-cells. We used the median sum of squares across the runs to generate F statistics using the residual error (ε), P-values were calculated based on this median F and median degrees of freedom based on the F-probability density function (*fpdf* tool). Since we independently performed the same statistical test for 15 drivers, we employed a full Bonferroni correction as ($\alpha = 0.05/15$). Our codes for these correlations are provided at https://github.com/EnsemblesTypes/DriverRegressions.

Results

For carbon, population density and proportion of urban area per 1 km^2-cell explained these spatial differences in ensemble certainty (Bonferroni corrected $P < 0.001/15$). Between the least and most densely-populated areas there was an estimated 43% increase in uncertainty among ensembles (calculated as [effect size/μ]), and a 81% uncertainty increase between the lowest and highest proportions of urban area. This population density correlation did not explain variation among individual models themselves, although the maps show similar patterns (Figure SI-4-1, Table SI-4-1). The proportion of woodland and peatland per cell (*i.e.* high carbon areas above- and below ground respectively) explained uncertainty among both models and ensemble approaches ($P < 0.001/15$). However, as our weights were estimated using data from forest/woodland locations only, increased accuracy in woodland areas is to be expected. None of the other tested drivers were significant predictors of uncertainty ($P > 0.05/15$, Table 1).

For water, the gradient both East-West and North-South in precipitation in the UK (Tanguy *et al.* 2019) is clearly seen as a gradient in water supply per hectare (Figure SI-4-1). The uncertainty both among water models and among ensemble approaches was highest in high rainfall areas (Figure SI-4-1). After removing spatial autocorrelation, the amount of precipitation and its seasonality were significantly positively correlated with this uncertainty (Bonferroni corrected $P < 0.001/15$; Table SI-4-1) with an estimated difference of 53% increase in among-ensemble uncertainty between the areas of lowest and highest precipitation (calculated as [effect size/μ]). Furthermore, among model and among ensemble variation were greater with higher variation in elevation change per cell (*i.e.* ruggedness) and lower in areas of greater land use without permanent cover (agriculture; Table SI-4-1) vs permanent cover (woodlands and grassland areas; Table SI-4-1).

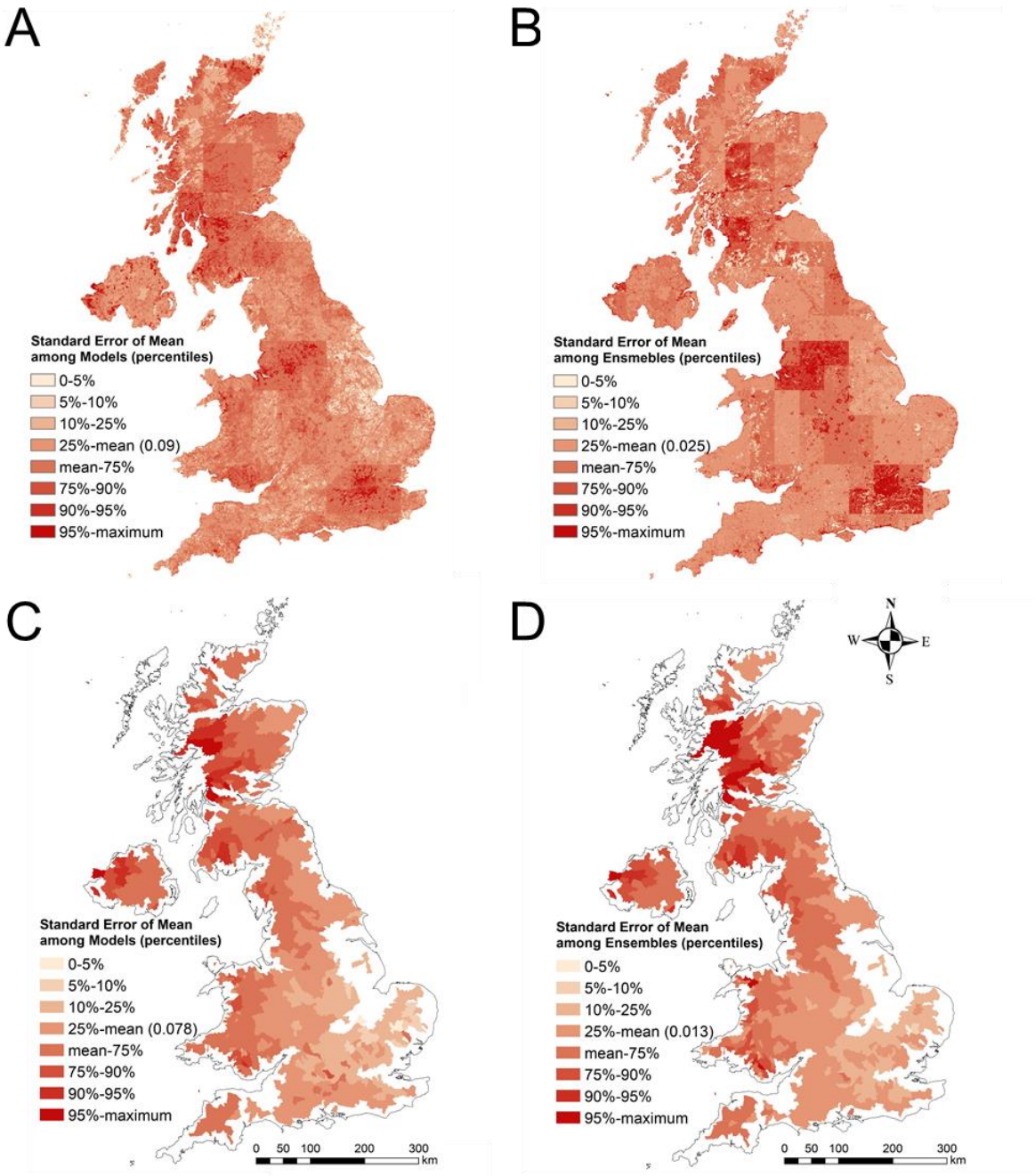

Figure SI-4-1. Spatial variation in the differences among ES models per hectare, represented as standard error of the mean in percentiles. (a) uncertainty among 10 carbon models (μ = 0.09 with σ 0.028); (b) uncertainty among all untrained carbon ensembles (μ = 0.025 with σ 0.011); N = 253,802 cell of 1-km^2; (c) uncertainty among 9 water models (μ = 0.078 with σ 0.034); (d) uncertainty among all untrained water ensembles (μ = 0.013 with σ 0.007).

Table. SI-4-1. Explanatory drivers of variation correlated to uncertainty, as standard error of mean, among models and among all untrained ensembles per grid cell/catchment. For carbon the estimated weights as calculated for the validation data locations are transferred to the full map; water remains on a per catchment base since non-accumulated flow, accumulated flow and outlet flow per catchment cannot be combined in one gridded map. We show linear F-values with significance (direction). We employed a SS-type I model: $[T_{(x)} \sim \beta_0 + \beta_1 Auto_{(x)} + \beta_2 Driver_{(x)} + \varepsilon]$ in which $T_{(x)}$ is either variation among models or ensembles for spatial cell *x*, *Driver* is the tested driver, with effect sizes β. Since we perform the same statistical test separately for 15 independent drivers, we employ a full Bonferroni correction.

	Uncertainty for			
	Carbon		**Water**	
	Individual Models	Ensembles	Individual Models	Ensembles
Spatial Auto Correlation (*Auto*)	18.9***(+)	80.5***(+)	1118***(+)	753***(+)
Explanatory Driver (*Driver*)				
Population density†	2.47	44.7***(+)	0.07	3.26
Annual precipitation‡	0.95	0.32	116***(+)	111***(+)
Potential evapotranspiration¶	0.80	0.86	1.66	2.65
Seasonality in precipitation§	2.70	0.80	90.8***(+)	63.8***(+)
Annual mean temperature§	3.76	0.66	0.21	0.11
Temperature coldest month§	2.28	0.45	1.38	0.20
Seasonality in temperature§	4.05	0.39	2.73	0.21
Temperature range§	3.82	0.40	5.32	1.48
Mean elevation©	2.97	0.33	21.5***(+)	15.7**(+)
Ruggedness (mean slopes)††	1.20	1.08	74.0***(+)	61.0***(+)
% Agriculture‡‡	58.5***(-)	7.70**(-)	32.9***(-)	19.5***(-)
% Grasslands‡‡	1.79	0.63	59.4***(+)	65.0***(+)
% Peatlands‡‡	1.57	9.87*(+)	1.12	9.24*(-)
% (sub)Urban‡‡	67.0***(+)	119***(+)	0.12	0.41
% Forests‡‡	71.4***(+)	111***(-)	25.0***(+)	21.0***(+)

Following: †(WorldPop 2018); ‡(Rowland *et al.* 2017); ¶(Zomer *et al.* 2006); §(Fick *et al.* 2017); © Copernicus EU-DEM v1 25m; †† derived from © using ArcGis Spatial Analyst Slope tool; ‡‡ % per 1-km^2 for LCM2015 (Rowland *et al.* 2017); * P < (0.05/15); ** P < (0.01/15); *** P < (0.001/15).

Discussion

Whilst our ES ensembles are UK specific, the geographic drivers of ensemble uncertainty identified here may be transferable to other regions. Broadly, our results show that our ES ensembles are less accurate in urban areas. For example we showed population dense areas to contain a higher uncertainty for carbon – forest trees do not equal urban trees with regard to ES (McHale *et al.* 2009). Most ES models are derived from a natural science perspective and focus on biophysical capacity in rural areas (Egoh *et al.* 2012; Martínez-Harms & Balvanera 2012; Wong *et al.* 2014). This is not to say that urban ES are not of high importance, but rather that urban residents may use different services (Larondell & Haase 2013; Haase *et al.* 2014) or the same services with different flow structures. For instance, it has been argued that rural inhabitants are more dependent on their local environment but urban inhabitants instead capitalise on ES flows from distant ecosystems (Cumming *et al.* 2014). Differences such as these have yet to be incorporated into many ES modelling platforms. For example, none of the InVest modules focus on urban ES, although this is currently being addressed by the UK National Capital Project. Thus, ES models developed for rural areas might not reliably characterise ES within a city ward; indeed, Co$ting Nature avoids spurious estimates by masking out ES in urban areas (Mulligan *et al.* 2010; Mulligan 2015). Despite this, these models are being applied in landscapes containing substantial urban and peri-urban areas (Pataki *et al.* 2011; Haase *et al.* 2014; Lee *et al.* 2015) and for ES comparisons under scenarios of increasing urbanisation (Bagstad *et al.* 2013; Zank *et al.* 2016), with potentially detrimental consequences for model accuracy.

We also find that ensembles for water are less accurate in areas of high rainfall, seasonality and rugosity. This could be because extreme processes are less well captured in models, requiring additional input data, *i.e.* it is likely models have an tendency to regress to the mean and not capture extreme events (Willcock *et al.* 2019).

Thus, evaluation of the accuracy of ES models and ensembles of ES models should become standard practice within the scientific community, although the feasibility of this is dependent on the availability of suitable validation data (Bryant *et al.* 2018). We advocate for the need to collect primary data on ES supply, use, perceptions, and well-being contributions over large regions (*e.g.* through national censuses) partly to incorporate into, but also to independently validate, ES models. Extra caution should be taken when using model predictions to support decision-making in areas with more among ensemble uncertainty.

SI-5 Advantages of winsorisation protocol for this dataset

In this work, 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 2.5% and 97.5% percentiles of the number of data points 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 values but have skewed and are corrected for by compressing the variance tails rather than trimming these (Kesselman *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. An even distribution is defined as having symmetric two sized tails – with ideally the lower and higher 2.5% of data points each covering 2.5% of the data-range –, with both the distribution mean and median values close to 0.5.

In our case, the skew is partly driven by the per area correction. Although we use lower size thresholds (2-ha for 'forests' and 25km2 for catchments), this effect still skews the dataset. On one side smaller areas outlier values could have a much stronger effect on the area corrected value per polygon due to the sampling effect of having a very high or low value per chance included. On the other hand, polygons are unequal in size and larger ones might contain lower value areas (especially for carbon).

Here, we show the effects of two normalisation protocols on our validation datasets in Figure SI-5-1A (carbon) and SI-5-1B (water supply). See SI-1-2 for the datasets. For the comparison 'standard normalisation', we divided the dataset by the absolute maximum.

For carbon, under standard normalisation there is a long one-sided higher end flat tail, without a lower-tail present (Figure SI-5-1A). The upper valued 2.5% of datapoints covers 40% of the data-range (calculated as maximum minus minimum value). This can be seen by the upper drawn dashed bar, representing the 97.5% percentile of number of datapoints, crossing the standard normalisation (brown) bars at the value of 0.62. This indicates there is an skew towards the dataset containing too many relatively low valued points and too few high valued points. The winsorising protocol compresses this upper 40% of the data-range that is covered with these few high valued points, spreading the whole data spectrum. As an effect of winsorisation the distribution is much less skewed and better spread out over the full 0-1 range. The mean value becomes closer to the expected 0.5. For the standard normalisation this was relatively low 0.31 (+/- 0.18 std) with a median of 0.32, for the winsorising protocol the mean is 0.47 (+/- 0.26 std) with an median value of 0.48, *i.e.*, the latter the value at 50% of data-points. By spreading out the data, the standard deviation increases accordingly.

For water supply, under standard normalisation, this dataset is already more balanced than carbon, having two tails under standard normalisation. However, there is longer lower end tail, with a small upper tail effect. For standard normalisation the lower 2.5% of points take up about 25% of the data-range: there are relatively too few low valued points. This can be seen by the upper drawn dashed bar, representing the 2.5% percentile of number of datapoints, crossing the standard normalisation (brown) bars at the value of 0.25. The winsorising protocol generates more relatively lower values, spreading the whole data spectrum. The upper tail of 10% of the range for 2.5% of the data-range is as well spread out more. The mean value becomes closer to an expected 0.5 value, in this case the mean value comes down because of removal of the lower tail range – *i.e.* there were too many high values compared to low values. For the standard normalisation this was a skewed low 0.62 (+/- 0.18 std), with a median of 0.63, for the winsorising protocol the mean is 0.54 (+/- 0.27 std) with a median of 0.56.

In conclusion, although the extremes are likely potentially ecologically relevant, the step of expressing our data as values per unit area has a skewing effect on the data distribution creating under standard normalisation, creating relatively long one-sided data range tails. The double winsorising protocol partly corrects for this skew, without removing data-points. This is most profound for carbon. With this double-

sided winsorising protocol, we generated a more statistically balanced dataset with mean values more close to the expected 0.5-value.

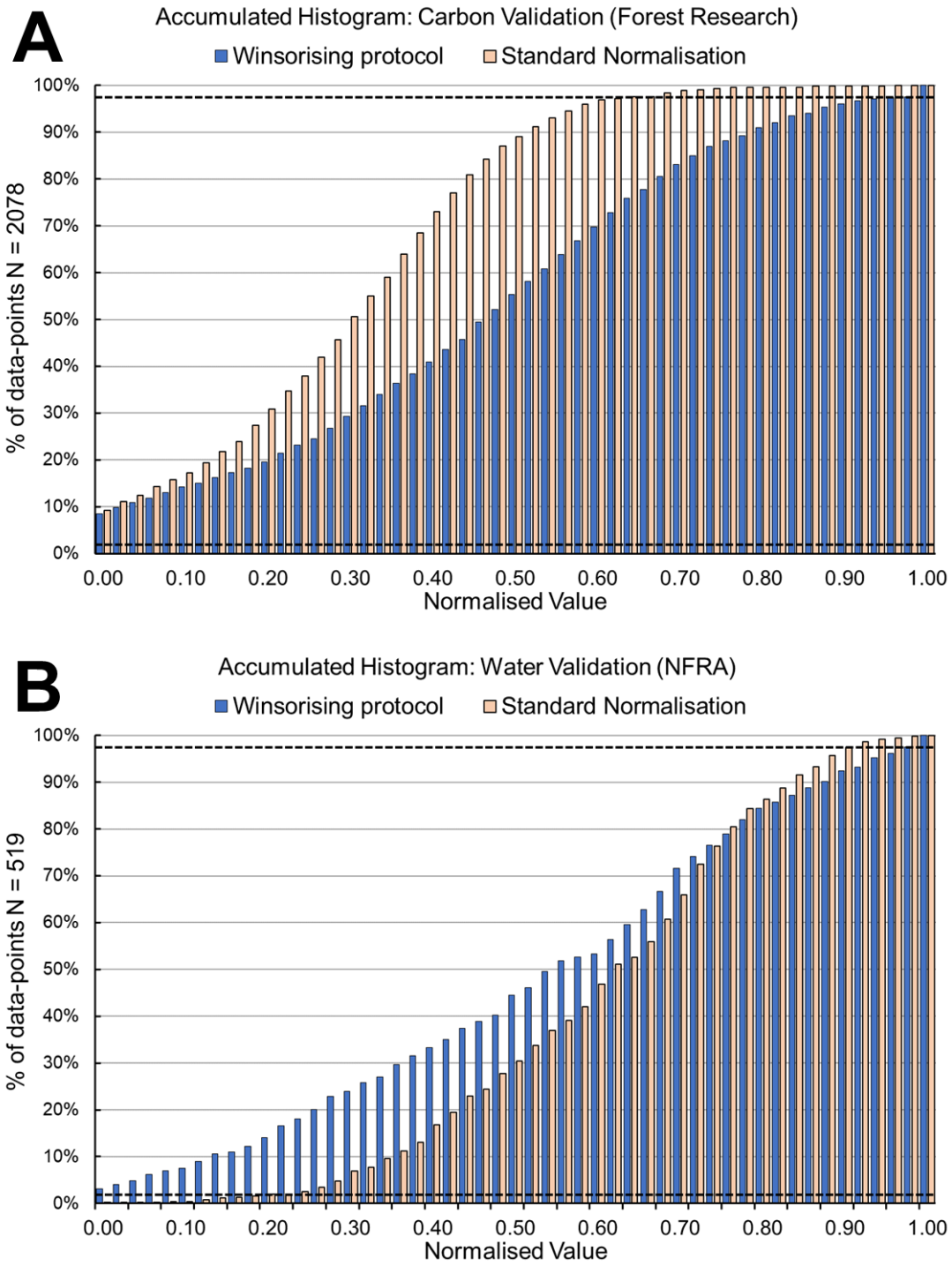

Figure SI-5-1. Effect of two different normalisation protocols on the distribution of the validation data-sets from this study (SI-1-2) after correction by area. Standard normalisation divides the full data-range by the absolute maximum value whereas the Winsorising protocol uses the lower and upper 2.5% percentile of the number of datapoints.

Supplementary Information References

Ahlström, A. *et al.* (2015). Carbon cycle. The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink. *Science* **348**, 895–899. https://doi.org/10.1126/science.aaa1668

Araújo, M.B. & New, M. (2007). Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**, 42–47. https://doi.org/10.1016/j.tree.2006.09.010.

Avitabile, V. *et al.* (2016). An integrated pantropical biomass map using multiple reference datasets. *Glob. Chang. Biol* **22**, 1406–1420. doi.org/10.1111/gcb.13139

Avitabile, V. & Camia, A. (2018). An assessment of forest biomass maps in Europe using harmonized national statistics and inventory plots. *For. Ecol. Manag.* **409**, 489–498. doi.org/10.1016/j.foreco.2017.11.047

Bagstad, K.J., Semmens, D.J. & Winthrop, R. (2013). Comparing approaches to spatially explicit ecosystem service modeling: A case study from the San Pedro River, Arizona. *Ecosyst. Serv.* **5**, 40–50. doi.org/10.1016/J.ECOSER.2013.07.007

Barredo, J.I., San Miguel, J., Caudullo, G. & Busetto, L. (2012). *A European map of living forest biomass and carbon stock.* (European Commission, Joint Research Centre). op.europa.eu/en/publication-detail/-/publication/b9345574-a96f-4417-87ed-1a85d2252834/language-en

Bell, V.A. *et al.* (2009). Use of soil data in a grid-based hydrological model to estimate spatial variation in changing flood risk across the UK. *J. Hydrol.* **377**, 335-350. doi.org/10.1016/j.jhydrol.2009.08.031

Bell, V.A. *et al.* (2018a). The MaRIUS-G2G datasets: Grid-to-Grid model estimates of flow and soil moisture for Great Britain using observed and climate model driving data. *Geosci. Data J.* **5**, 63–72. https://doi.org/10.1002/gdj3.55

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]*. (NERC Environmental Information Data Centre). https://doi.org/10.5285/f52f012d-9f2e-42cc-b628-9cdea4fa3ba0

Bryant, B.P. *et al.* (2018). Transparent and feasible uncertainty assessment adds value to applied ecosystem services modeling. *Ecosyst.Serv.* **33**, 103–109. doi.org/10.1016/j.ecoser.2018.09.001

Bugmann, H.A (2001). Review of Forest Gap Models. *Clim Change* **51**, 259–305. doi.org/10.1023/A:1012525626267

Brun, P. *et al.* (2020). Model complexity affects species distribution projections under climate change. *J. Biogeogr.* **47**, 130–142. https://doi.org/10.1111/jbi.13734

Costanza, R. *et al.* (2014). Changes in the global value of ecosystem services. *Glob. Environ. Change* **26**, 152–158. doi.org/10.1016/j.gloenvcha.2014.04.002

Coxon, G. *et al.* (2019a). DECIPHeR v1: Dynamic fluxEs and ConnectIvity for Predictions of HydRology. *Geosci. Model Dev.* **12**, 2285–2306. doi.org/10.5194/gmd-12-2285-2019

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

Crossman, N.D. *et al.* (2012). Identifying priority areas for reducing species vulnerability to climate change. *Divers. Distrib.* **18**, 60–72. doi.org/10.1111/j.1472-4642.2011.00851.x

Cumming, G.S. *et al.* (2014). Implications of agricultural transitions and urbanization for ecosystem services. *Nature* **515**, 50–57. doi.org/10.1038/nature13945

Ding, H. & Bullock, J.M. (2018). *A Guide to Selecting Ecosystem Service Models for Decision-Making: Lessons from Sub-Saharan Africa*. (World Resources Institute). wri.org/publication/guide-selecting-ecosystem-service

Dixon, H. *et al.* (2013). The effective management of national hydrometric data: experiences from the United Kingdom. *Hydrol Sci J.* **58**, 1383–1399. doi.org/10.1080/02626667.2013.787486

Diengdoh, V.L. *et al.* (2020). A validated ensemble method for multinomial land-cover classification. *Ecol. Inform.* **56**, 101065. https://doi.org/10.1016/j.ecoinf.2020.101065

Dormann, C.F. *et al.* (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**, 609-628. doi.org/10.1111/j.2007.0906-7590.05171.x

Dormann, C.F. *et al.* (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecol. Monogr.* **88**, 485–504. https://doi.org/10.1002/ecm.1309

EEA (1993). *CORINE Land Cover - Technical Guide*. (Office for Official Publications of European Communities). eea.europa.eu/publications/COR0-landcover

EEA (2000). *CORINE land cover technical guide – Addendum 2000*. European Environment Agency). eea.europa.eu/publications/tech40add

Egoh, B. *et al.* (2012). *Indicators for mapping ecosystem services: a review*. (Publications Office of the European Union). ec.europa.eu/jrc/en/publication/ecosystem-services-review

Englund, O. *et al.* (2017). How to analyse ecosystem services in landscapes–A systematic review. *Ecol. Indic.* **73**, 492–504. https://doi.org/10.1016/j.ecolind.2016.10.009

Erceg-Hurn, D.M. & Mirosevich, V.M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *Am. Psychol.* **63**, 591–601. http://dx.doi.org/10.1037/0003-066X.63.7.591.

Evans, L.R. *et al.* (2009). Surrogate decision-makers' perspectives on discussing prognosis in the face of uncertainty. *Am. J. Respir. Crit. Care Med.* **179**, 48–53. doi.org/10.1164/rccm.200806-969OC

FAO (2001). *Global Ecological Zoning for the Global Forest Resources Assessment 2000 - Final Report*. (Food and Agriculture Organization of the United Nations). fao.org/3/ad652e/ad652e00.htm

FAO (2010). *Global Forest Resources Assessment 2010. Main Report*. (Food and Agriculture Organization of the United Nations). fao.org/3/i1757e/i1757e00.htm

Fick, S.E. & Hijmans, R.J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302-4315. doi.org/10.1002/joc.5086.

Friedlingstein, P. *et al.* (2019). Global carbon budget 2019. *Earth Syst. Sci. Data* **11**, 1783-1838. doi.org/10.5194/essd-11-1783-2019

Gassert, F. *et al.* (2015). *Aqueduct Global Maps 2.1*. (World Resources Institute). www.wri.org/resources/data-sets/aqueduct-global-maps-21-data

Grenouillet, G. *et al.* (2011). Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography* **34**, 9–17. https://doi.org/10.1111/j.1600-0587.2010.06152.x

Goldstein, J.H *et al.* (2012). Integrating ecosystem-service trade-offs into land-use decisions. *Proc. Natl. Acad. Sci. USA* **109**, 7565–7570. ncbi.nlm.nih.gov/pubmed/22529388.

de Groot, R. *et al.* (2012). Global estimates of the value of ecosystems and their services in monetary units. *Ecosyst. Serv.* **1**, 50–61. doi.org/10.1016/j.ecoser.2012.07.005

Haase, D. *et al.* (2014). A Quantitative Review of Urban Ecosystem Service Assessments: Concepts, Models, and Implementation. Ambio **43**, 413–433. doi.org/10.1007/s13280-014-0504-0

Hamel, P. & Bryant, B.P. (2017). Uncertainty assessment in ecosystem services analyses: Seven challenges and practical responses. *Ecosyst. Serv.* **24**, 1–15. doi.org/10.1016/J.ECOSER.2016.12.008

Harris, I. *et al.* (2014). Updated high-resolution grids of monthly climatic observations– the CRU TS3.10 Dataset, *Int. J. Climatol*. **34**, 623–642. doi.org/10.1002/joc.3711

Harvey, C.L. *et al.* (2012) An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. *Hydrol. Res.* **43**, 618–636. doi.org/10.2166/nh.2012.110

Henrys, P.A. *et al.* (2016). *Model estimates of aboveground carbon for Great Britain*. (NERC Environmental Information Data Centre). doi.org/10.5285/9be652e7-d5ce-44c1-a5fc-8349f76f5f5c

Hickler, T. *et al.* (2012). Projecting the future distribution of European potential naturalvegetation zones with a generalized, tree species-based dynamic vegetation model. *Glob. Ecol. Biogeogr.* **21**, 50–63. doi.org/10.1111/j.1466-8238.2010.00613.x

IPCC (2006). *2006 IPCC Guidelines for National Greenhouse Gas Inventories (Chapter 4 Forest Land)*. (IGES). ipcc.ch/report/2006-ipcc-guidelines-for-national-greenhouse-gas-inventories/

Kareiva, P. *et al.* (2011). *Natural Capital: Theory and Practice of Mapping Ecosystem Services*. (Oxford University Press). oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199588992.001.0001

Keselman, H. J. *et al.* (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychol. Methods* **13**, 110–129. https://doi.apa.org/doi/10.1037/1082-989X.13.2.110

Kindermann, G.E. *et al.* (2008). A global forest growing stock, biomass and carbon map based on FAO statistics. *Silva Fennica* **42**, 397–396. http://pure.iiasa.ac.at/id/eprint/8616/

Knorr, W. *et al.* (2016). A.Climate, CO_2 and human population impacts on global wildfire emissions. *Biogeosciences* **13**, 267–282. doi.org/10.5194/bg-13-267-2016

Knutti, R. *et al.* (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.* **40**, 1194–1199. https://doi.org/10.1002/grl.50256

Kobayashi, S. *et al.* (2015). The JRA-55 Reanalysis: General Specifications and Basic Characteristics, *J. Meteorol. Soc. Jpn.* **93**, 5–48. doi.org/10.2151/jmsj.2015-001.

Larondelle, N. & Haase, D. (2013). Urban ecosystem services assessment along a rural–urban gradient: A cross-analysis of European cities. *Ecol. Indic.* **29**, 179–190. doi.org/10.1016/J.ECOLIND.2012.12.022

Lavorel, S. *et al.* (2011). Using plant functional traits to understand the landscape distribution of multiple ecosystem services. *J. Ecol.* **99**, 135-147. doi.org/10.1111/j.1365-2745.2010.01753.x

Lawler, J.J. *et al.* (2014). Projected land-use change impacts on ecosystem services in the United States. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 7492-7497. doi.org/10.1073/pnas.1405557111

Lee, Y.C. *et al.* (2015). Ecosystem services in peri-urban landscapes: The effects of agricultural landscape change on ecosystem services in Taiwan's western coastal plain. *Landsc. Urban Plan.* **139**, 137–148. doi.org/10.1016/J.LANDURBPLAN.2015.02.023

Liu, D. *et al.* (2020). An integrated approach towards modeling ranked weights. *Comput. Ind. Eng*. **147**, 106629. https://doi.org/10.1016/j.cie.2020.106629

Marmion, M. *et al.* (2009). Evaluation of consensus methods in predictive species distribution modelling. *Divers. Distrib.* **15**, 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x

Martínez-Harms, M.J. & Balvanera, P. (2012). Methods for mapping ecosystem service supply: a review. *Int. J. Biodivers. Sci. Ecosyst. Serv. Manag.* **8**, 17–25. doi.org/10.1080/21513732.2012.663792

Martínez-López, J. *et al.* (2019). Towards globally customizable ecosystem service models. *Sci. Total Environ.* **650**, 2325–2336. doi.org/10.1016/j.scitotenv.2018.09.371

Maskell, L.C. *et al.* (2008). Countryside Survey Technical Report No.01/07: Field Mapping Handbook. (Centre for Ecology & Hydrology). countrysidesurvey.org.uk/content/all-technical-reports

Masson, D. & Knutti, R. (2011). Climate model genealogy. *Geophys. Res. Lett.* **38**. L08703. https://doi.org/10.1029/2011GL046864

McHale, M.R. *et al.* (2009). Urban forest biomass estimates: is it important to use allometric relationships developed specifically for urban trees?. *Urban Ecosyst.* **12**, 95–113. doi.org/10.1007/s11252-009-0081-3

McKenzie, E. *et al.* (2014). Understanding the use of ecosystem service knowledge in decision making: lessons from international experiences of spatial planning. *Environ. Plan C Politics Space.* **32**, 320–340. doi.org/10.1068%2Fc12292j

Milne, R. & Brown, T.A. (1997). Carbon in the vegetation and soils of Great Britain. *J. Environ. Manage*. **49**, 413-433. homepages.ed.ac.uk/shs/Climatechange/Carbon%20sequestration/UK%20soilcarbon.pdf

Monteith, J.L. (1965). Evaporation and environment. in *Symposia of the society for experimental biology, vol. 19*, 205-234 (Cambridge University Press). repository.rothamsted.ac.uk/item/8v5v7/evaporation-and-environment

Morton, R.D. *et al.* (2014). *Land Cover Map 2007 (1km dominant target class, GB) v1.2*. (NERC Environmental Information Data Centre). doi.org/10.5285/6cffd348-dad7-46f9-9c5b-8d904dd5b2a2

Mulligan M. (2013). WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally. *Hydrol. Res.* **44**, 748–69. doi.org/10.2166/nh.2012.217

Mulligan, M., (2015). Trading off agriculture with nature's other benefits, spatially. in *Impact of Climate Change on Water Resources in Agriculture* (eds. Zolin, C. & de Rodrigues, R.A.R) 184–204 (CRC Press). 10.1201/b18652-10

Mulligan, M. *et al.* (2010). Capturing and quantifying the flow of ecosystem services, in *Framing the Flow: Innovative Approaches to Understand, Protect and Value Ecosystem Services Across Linked Habitats* (eds. Silvestri, S. & Kershaw, F) 26–33 (UNEP World Conservation Monitoring Centre, 2010). unenvironment.org/es/node/11920

Mulligan, M. & Wainwright, J. (2013). Modelling and model building. in *Environmental modelling: Finding simplicity in complexity* (eds. Wainwight, J. & Mulligan, M.) 7-23 (Wiley).

Nelson, E.J. *et al.* (2013). Climate change's impact on key ecosystem services and the human well-being they support in the US. *Front. Ecol. Environ.* **11**, 483-893. doi.org/10.1890/120312

Pataki, D.E. *et al.* (2011). Coupling biogeochemical cycles in urban environments: ecosystem services, green solutions, and misconceptions. *Front. Ecol. Environ.* **9**, 27–36. doi.org/10.1890/090220

Polasky, S. *et al.* (2011). Decision-making under great uncertainty: environmental management in an era of global change. *Trends Ecol. Evol.* **26**, 398–404. doi.org/10.1016/j.tree.2011.04.007

Prather, C.M. *et al.* (2013). Invertebrates, ecosystem services and climate change. *Biological Reviews* **88**, 327-348. doi.org/10.1111/brv.12002

Redhead, J.W. *et al.* (2016). Empirical validation of the InVEST water yield ecosystem service model at a national scale. *Sci. Total Environ.* **569**, 1418–1426. doi.org/10.1016/j.scitotenv.2016.06.227

Refsgaard, J.C. *et al.* (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

Robinson, E.L. *et al.* (2017).*Climate Hydrology and Ecology Research Support System Meteorology Dataset for Great Britain (1961-2012) [CHESS-met]*, (NERC Environmental Information Data Centre). doi.org/10.5285/b745e7b1-626c-4ccc-ac27-56582e77b900

Rowland, C.S. (2017). *et al. Land Cover Map 2015.* (NERC Environmental Information Data Centre). doi.org/10.5285/6c6c9203-7333-4d96-88ab-78925e7a4e73.

Ruesch, A. & Gibbs, H.K. (2008). *New IPCC Tier-1 Global Biomass Carbon Map For the Year 2000*. (Carbon Dioxide Information Analysis Center). cdiac.ess-dive.lbl.gov/epubs/ndp/global_carbon/tables.html#

Schirpke, U. *et al.* (2013). Multiple ecosystem services of a changing Alpine landscape: past, present and future. *Int. J. Biodivers. Sci. Ecosyst. Serv. Manag.* **9**, 123-135. doi.org/10.1080/21513732.2012.751936

Scholes, R.J. (1998). *The South African 1: 250 000 maps of areas of homogeneous grazing potential.* (CSIR, South Africa). No internet reference

Sharps, K. *et al.* (2017). Comparing strengths and weaknesses of three ecosystem services modelling tools in a diverse UK river catchment. *Sci. Total Environ.* **584**, 118–130. doi.org/10.1016/j.scitotenv.2016.12.160

Smith, B. *et al.* (2001). Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. *Glob. Ecol. Biogeogr.* **10**, 621–37. jstor.org/stable/3182691

Smith, B. *et al.* (2014). Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. *Biogeosciences* **11**, 2027–2054. hdl.handle.net/11858/00-001M-0000-0019-8928-C

Tanguy, M. *et al.* (2019). *Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2017) [CEH-GEAR].* (NERC Environmental Information Data Centre). doi.org/10.5285/ee9ab43d-a4fe-4e73-afd5-cd4fc4c82556

Tebaldi, C. & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 365, 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Thomas, A. *et al.* (2020). Fragmentation and thresholds in hydrological flow-based ecosystem services. *Ecol. Appl.* **30**, e02046. doi.org/10.1002/eap.2046

Thuiller, W. *et al.* (2009). BIOMOD–a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369-373. doi.org/10.1111/j.1600-0587.2008.05742.x

Thuiller, W. *et al.* (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nat. Commun.* **10,** 1-9. doi.org/10.1038/s41467-019-09519-w

Trodahl, M.I. *et al.* (2017). Investigating trade-offs between water quality and agricultural productivity using the Land Utilisation and Capability Indicator (LUCI)—a New Zealand application. *Ecosyst. Serv.* **26**, 388–399.

Verhagen, W. *et al.* (2017). Use of demand for and spatial flow of ecosystem services to identify priority areas. *Conserv. Biol.* **31**, 860–871. https://doi.org/10.1111/cobi.12872

Villa, F. *et al.* (2014). A methodology for adaptable and robust ecosystem services assessment. *PloS One* **9**, e91001. doi.org/10.1371/journal.pone.0091001

Viovy, N. (2009) *CRUNCEP data set*, available at: ftp://nacp.ornl.gov/synthesis/2009/frescati/temp/land_use_change/original/readme.htm, last access: June 2016.

Willcock, S., *et al.* (2014). Quantifying and understanding carbon storage and sequestration within the Eastern Arc Mountains of Tanzania, a tropical biodiversity hotspot. *Carbon Balance Manag.* **9**, 2. doi.org/10.1186/1750-0680-9-2

Willcock, S. *et al.* (2016). Do ecosystem service maps and models meet stakeholders' needs? A preliminary survey across sub-Saharan Africa. *Ecosyst. Serv.* **18**, 110–117. doi.org/10.1016/j.ecoser.2016.02.038

Willcock, S. *et al.* (2019). A Continental-Scale Validation of Ecosystem Service Models. *Ecosystems* **22**, 1902–1917. doi.org/10.1007/s10021-019-00380-y

Willcock, S. *et al.* (2020). Ensembles of ecosystem service models can improve accuracy and indicate uncertainty. *Sci. Total Environ*. 141006. doi.org/10.1016/j.scitotenv.2020.141006

Wong, C.P. *et al.* (2014). Linking ecosystem characteristics to final ecosystem services for public policy. *Ecol. Lett.* **18**, 108–118. doi.org/10.1111/ele.12389

WorldPop & CIESIN Columbia University (2018). *Global High Resolution Population Denominators Project*. dx.doi.org/10.5258/SOTON/WP00645

Zomer, R. *et al.* (2006). *Carbon, land and water: A global analysis of the hydrologic dimensions of climate change mitigation through afforestation/reforestation* (IWMI). cgiarcsi.community/zomer-et-al-2007

Zank, B. *et al. (2016).* Modeling the effects of urban expansion on natural capital stocks and ecosystem service flows: A case study in the Puget Sound, Washington, USA. *Landsc. Urban Plan.* **149**, 31–42. doi.org/10.1016/J.LANDURBPLAN.2016.01.004