Towards an ensemble-based evaluation of land surface models in light of uncertain forcings and observations

Vivek. K. Arora¹, Christian Seiler², Libo Wang², and Sian Kou-Giesbrecht¹

¹Canadian Centre for Climate Modelling and Analysis, Climate Research Division, Environment Canada, Victoria, BC, Canada

²Climate Processes Section, Climate Research Division, Environment and Climate Change Canada, Toronto, ON,
 Canada

2 3 4

Correspondence to: Vivek K. Arora (vivek.arora@ec.gc.ca)

5 Abstract

6

7 Quantification of uncertainty in fluxes of energy, water, and CO₂ simulated by land surface 8 models (LSMs) remains a challenge. LSMs are typically driven with, and tuned for, a specified 9 meteorological forcing data set and a specified set of geophysical fields. Here, using two data sets 10 each for meteorological forcing and land cover representation (in which the increase in crop area over the historical period is implemented in the same way), as well as two model structures (with 11 and without coupling of carbon and nitrogen cycles), the uncertainty in simulated results over 12 13 the historical period is quantified for the Canadian Land Surface Scheme Including 14 Biogeochemical Cycles (CLASSIC) model. The resulting eight $(2 \times 2 \times 2)$ model simulations are evaluated using an in-house model evaluation framework that uses multiple observations-based 15 data sets for a range of quantities. The simulated area burned, fire CO₂ emissions, soil carbon 16 mass, vegetation carbon mass, runoff, heterotrophic respiration, gross primary productivity, and 17 sensible heat flux show the largest spread across the eight simulations relative to their global 18 ensemble mean values. Simulated net atmosphere-land CO₂ flux, a critical determinant of the 19 20 performance of LSMs, is found to be largely independent of the simulated pre-industrial 21 vegetation and soil carbon mass although our framework represents the historical increase in crop area in the same way in both land cover representations. This indicates that models can 22 provide reliable estimates of the strength of the land carbon sink despite some biases in carbon 23 stocks. Results show that evaluating an ensemble of model results against multiple observations 24 disentangles model deficiencies from uncertainties in model inputs, observation-based data, and 25 model configuration. 26

28 **1. Introduction**

The current generation land surface models (LSMs) explicitly simulate the fluxes of 29 30 energy, water, momentum, and trace gases (including CO₂, CH₄, and N₂O) between the 31 atmosphere and the land surface. These models have become an essential tool in understanding 32 what role the land surface plays in the global climate system under current and projected future changes in environmental conditions, including atmospheric CO₂ concentration (Bonan and 33 Doney, 2018). LSMs are also an essential component of climate and Earth system models (ESMs), 34 35 together with their ocean and atmosphere components. Within the framework of ESMs, LSMs 36 are coupled interactively to their atmospheric components through the fluxes of energy, 37 momentum, and matter.

The complexity of LSMs has increased over time as more physical and biogeochemical 38 39 processes have been included in their framework (Fisher and Koven, 2020; Kyker-Snowman et al., 2022). This increased complexity combined with the uncertainty in our understanding of 40 physical and biogeochemical processes implies that different models respond differently even 41 when driven with the same external forcings. One estimate of the uncertainty in our 42 understanding of land surface physical and biogeochemical processes is obtained by evaluating 43 the inter-model spread in a given quantity when models are forced in the same manner. Other 44 than the uncertainty among models due to differences in their model structures and 45 parameterizations of various processes, uncertainty also exists due to at least three other 46

reasons. These include uncertainty 1) in parameter values¹ of represented processes, 2) in driving 47 48 meteorological data, and 3) in the specification of the geophysical fields. LSMs are typically driven with meteorological data consisting of seven primary variables (incoming long and shortwave 49 radiation, temperature, precipitation, specific humidity, wind speed, and pressure). In addition, 50 51 the geophysical fields of land cover, soil texture, and soil permeable depth are also required. Driving data for LSMs also consist of atmospheric CO₂ concentration and other model-specific 52 external forcings such as nitrogen deposition and fertilizer application rates for models that 53 54 include a representation of the terrestrial nitrogen cycle, and lightning, population density, and 55 gross domestic product (GDP) for models that simulate wildfires.

56 Every year more than 15 land surface modelling groups participate in the TRENDY (trends 57 in net land-atmosphere carbon exchanges) project where they perform a set of simulations that are driven with specified external forcings. The simulations are performed from the year 1700 to 58 59 the present day. These simulations contribute to the annual Global Carbon Project's (GCP) 60 analysis of the land carbon sink together with its analysis of anthropogenic CO₂ emissions and the ocean carbon sink (Friedlingstein et al., 2019). The external forcings used to drive LSMs in the 61 TRENDY intercomparison include, 1) six hourly meteorological data from 1901 to the present day 62 (the most recent 2020 TRENDY intercomparison used the CRU-JRA forcing obtained by blending 63 the climate research unit (CRU) monthly data and the Japanese reanalysis (JRA)); 2) atmospheric 64 65 CO_2 concentration; and 3) information about changes in crop area and other land use changes (LUC) from the land use harmonization (LUH) product (Hurtt et al., 2020a). The information about 66

¹ Changes in parameter values do not constitute different parameterizations. For example, two models may use the same parameterization, say y=mx+b, but different values of its parameters m and b. However, y=mx+b and $y = mx^2$ are considered to be two different parameterizations.

67 changes in crop area and other LUC is used by land surface modelling groups to reconstruct 68 historical land cover from the year 1700 to the present day consistent with the number of the 69 plant functional types (PFTs) a given model represents. The protocol also provides nitrogen 70 deposition and fertilization application rates for models including nitrogen cycling.

Models participating in the TRENDY simulations are thus driven with common 71 meteorological and LUC forcings as part of its protocol. The resulting spread across models 72 participating in the TRENDY project thus provides a measure of inter-model uncertainty, as 73 74 mentioned earlier. Traditionally the uncertainty associated with model structure has gained the 75 most attention and the scientific community has responded to this by performing model 76 intercomparison projects (MIPs) where models are driven according to a common protocol. The 77 coupled model intercomparison project (CMIP) in the climate community together with its various sub-projects (Eyring et al., 2016) is another prominent example. MIPs now routinely form 78 79 the basis of evaluating models against observations and multi-model means of various quantities. Multi-model means are also considered the best estimate for a given quantity (Tebaldi and 80 Knutti, 2007). 81

The modelling community has been long aware of the uncertainty associated with parameter values, since a large fraction of physical and biogeochemical model processes are parameterized, and such uncertainty analysis dates back to the early hydrological models (e.g. Hornberger and Spear, 1981; Beven and Binley, 1992). More recent examples of parameter value uncertainty in the context of a given LSM include Poulter et al. (2010), Booth et al. (2012), and Li et al. (2018a). The land surface modelling community, however, has only recently begun to address and quantify uncertainty related with driving meteorological data. Wu et al. (2017), for

example, illustrate the uncertainty in gross primary productivity (GPP) simulated by the Lund-89 90 Potsdam-Jena General Ecosystem Simulator (LPJ-GUESS) model when driven by six different meteorological data sets. Bonan et al. (2019) analyze the uncertainty in simulated carbon cycle 91 related variables using three versions of the Community Land Model (CLM) when driven with two 92 93 meteorological data sets over the historical period. Slevin et al. (2017) assess the uncertainty in simulated GPP by the JULES land model when driven by three different meteorological data sets. 94 Studies that evaluate the effect of different land cover representations on model performance 95 96 are even fewer. Tian et al. (2004) and Lawrence and Chase (2007) study the effect of new land 97 surface boundary conditions, including leaf area index and fractional vegetation cover, based on the MODIS satellite data as implemented in CLM2 in the Community Atmosphere Model (CAM2) 98 99 and CLM3 in the Community Climate System Model (CCSM 3.0), respectively.

100 Here, we drive the Canadian Land Surface Scheme Including Biogeochemical Cycles 101 (CLASSIC) with two sets of historical meteorological forcings and also two land cover representations to quantify the uncertainty associated with both these forcings. Other than 102 these, we also use two versions of the CLASSIC model: one that represents the interactions 103 104 between the carbon (C) and nitrogen (N) cycles and the other in which these interactions are turned off. CLASSIC has contributed to the simulations for the TRENDY intercomparison, and the 105 GCP, since 2016 (formerly under the CLASS-CTEM name). Seiler et al. (2021a) have evaluated how 106 107 well the CLASSIC model performs when forced with three different meteorological data sets using the model version without the N cycle. Using the two meteorological forcing data sets, two 108 representations of land cover, and two versions of the model we perform eight simulations over 109 the historical period since 1700. All of these simulations may be considered equally likely 110

representations of the modelled state of the land surface over the historical period. Yet, they all 111 112 have their own distinct biases since simulated land surface states and fluxes are different. We use these simulations to illustrate the uncertainty associated with meteorological forcing and the 113 two different representations of land cover that are used to drive the model. We also use an in-114 115 house open-source benchmarking system (see code/data availability section) to evaluate these different simulations against observations-based data sets: AMBER (Automated Benchmarking R 116 Package) (Seiler et al., 2021b) uses gridded and in-situ observation-based estimates of 19 energy, 117 118 water, and C cycle related variables to evaluate LSMs.

119 Section 2 of this paper describes the framework of the CLASSIC land model and the forcing 120 data that are required to drive the model. Section 3 describes the two meteorological data sets, 121 the two representations of land cover that are used to drive the model, and the simulations performed for this study. Section 4 analyses the results from the simulations to illustrate their 122 123 different states and reports results from the AMBER benchmarking exercise. Finally, the discussion and conclusions are presented in Section 5. The use of more than one meteorological 124 forcing data sets and land cover representation yields a conundrum since tuning model 125 parameters for a given forcing data set is not a useful exercise anymore. We also report a new 126 finding that despite different land C states (characterized in terms of vegetation and soil C mass) 127 in the eight simulations considered here, the net atmosphere-land CO₂ flux over the historical 128 129 period in these simulations is consistent with estimates from the GCP. This and the discussion about the broader question of model tuning are also presented in Section 5. 130

131 **2. The CLASSIC land modelling framework**

132 **2.1** The physical and carbon biogeochemical processes

133 The CLASSIC land model is the successor to, and based on, the coupled Canadian Land 134 Surface Scheme (CLASS; (Verseghy, 1991; Verseghy et al., 1993)) and the Canadian Terrestrial 135 Ecosystem Model (CTEM; (Arora and Boer, 2005; Melton and Arora, 2016b). CLASSIC also serves as the land component in the family of Canadian Earth System Models (Arora et al., 2009, 2011; 136 Swart et al., 2019). Melton et al. (2019) provide an overview of the CLASSIC land model and 137 launched it as a community model. The basis of the modelling of physical and biogeochemical 138 139 processes in CLASSIC comes from CLASS and CTEM, respectively, both of which have a long 140 history of development. CLASSIC simulates land-atmosphere fluxes of water, energy, and 141 momentum based on its physics, and fluxes of CO₂, CH₄, N₂O, NO_x, and NH₃ based on its biogeochemical process. The representation of the terrestrial N cycle is a new addition to CLASSIC 142 (Asaadi and Arora, 2021; Kou-Giesbrecht and Arora, 2022) and allows for the simulation of the 143 144 interactions between the C and N cycles explicitly.

The CLASSIC model simulations can be performed over a spatial domain, which may be global or regional, using gridded data or at a point scale, e.g. using meteorological and geophysical data from a FluxNet site. The primary physical and biogeochemical processes of CLASSIC are briefly summarized in the next two sections.

149 **2.1.1 Physical processes**

The calculations for physical processes in CLASSIC are performed over vegetated, snow, and bare fractions at a time step of 30 minutes. In the version used here, the fractional coverage of the four plant functional types (PFTs) (needleleaf trees, broadleaf trees, crops, and grasses)

characterizes vegetation for each grid cell. The fractional coverage of these four PFTs is specified 153 154 over the historical period in this study. The structure of vegetation is characterized by leaf area index (LAI), vegetation height, canopy mass, and rooting distribution through the soil layers all of 155 which are dynamically simulated by the biogeochemical module of CLASSIC. Twenty ground 156 157 layers represent the soil profile, starting with 10 layers of 0.1 m thickness. The thickness of layers gradually increases to 30 m for a total ground depth of over 61 m. The depth of permeable soil 158 159 layers and thus the depth to bedrock varies geographically and is specified based on the 160 SoilGrids250m data set (Hengl et al., 2017). Liquid and frozen soil moisture contents, and soil 161 temperature, are determined prognostically for permeable soil layers. The temperature, albedo, 162 mass, and density of a single-layer snow pack (when the climate permits snow to exist) are also prognostically modelled. The result of physics calculations yields fluxes of energy (primarily net 163 164 radiation, ground heat flux, and latent and sensible heat fluxes) and water (primarily 165 evapotranspiration and runoff) at the land-atmosphere boundary.

166 **2.1.2 Biogeochemical processes**

The biogeochemical processes in CLASSIC, based on CTEM, are described in detail in the appendix of Melton and Arora (2016). The biogeochemical processes simulate the landatmosphere exchange of CO_2 and as a result simulate vegetation as a dynamic component depending on the environmental conditions.

The biogeochemical module of CLASSIC prognostically calculates the amount of C in the model's three live (leaves, stem, and root) and two dead (litter and soil) C pools for each PFT. The live vegetation pools are separated into their structural and non-structural components. The C

amount in these pools is represented per unit land area (kg C/m^2). The amount of C in the live 174 175 and dead C pools and all terrestrial ecosystem processes in the biogeochemical module in this study are modelled for nine PFTs that map directly onto the four base PFTs used in the physics 176 module of CLASSIC. Needleleaf trees are divided into their deciduous and evergreen phenotypes, 177 178 broadleaf trees are divided into cold deciduous, drought deciduous, and evergreen phenotypes, and crops and grasses are divided based on their photosynthetic pathways into C₃ and C₄ 179 180 versions. The physical process in CLASSIC are less sensitive to this sub-division of PFTs which is 181 essential for modelling biogeochemical processes. For instance, simulating the onset and offset 182 of leaves is different between evergreen and deciduous phenotypes of needleleaf and broadleaf 183 trees and this is simulated in the biogeochemical module of the model. However, once the leaf 184 area index (LAI) is known, the physical processes in CLASSIC do not need information about the underlying deciduous or evergreen nature of leaf phenology. For example, the interception of 185 186 rain and snow by canopy leaves (that is typically modelled as a function of LAI and a PFTdependent parameter that accounts for leaf orientation and shape) does not depend on the 187 188 underlying evergreen or deciduous nature of the leaf phenology. In general, biogeochemical processes benefit more in terms of realism than physical processes when the number of PFTs is 189 190 increased. For example, in offline CLASSIC simulations, large changes in leaf area index (LAI) do 191 not change total latent heat flux considerably since the partitioning of evapotranspiration into its 192 sub-components (transpiration, soil evaporation, and evaporation/sublimation of intercepted rain/snow) adjusts. A decrease in transpiration and evaporation of intercepted precipitation, due 193 194 to a decrease in LAI, is compensated by an increase in soil evaporation. This is expected since 195 water and energy fluxes are determined largely by available energy and precipitation.

The litter and soil C pools are tracked for each soil layer but the movement of C between 196 197 the soil layers is not yet modelled. Other than photosynthesis and leaf respiration which are modelled at a time step of 30 minutes all other biogeochemical processes are modelled at a daily 198 199 time step. These include: 1) allocation of C from leaves to stem and roots, 2) autotrophic 200 respiration from the live C pools and heterotrophic respirations from the dead C pools, 3) leaf phenology, 4) turnover of live vegetation components that generates litter, 5) mortality, 6) LUC, 201 and 7) fire (Arora and Melton, 2018). Competition between PFTs for space is not modelled in this 202 203 study and fractional coverage of the nine PFTs is specified based on the representation of the 204 land cover as explained in the next section.

205 When the N cycle is turned on, land-atmosphere fluxes of N₂O, NO_x, and NH₃, and N 206 leaching are also modelled in response to biological N fixation, N fertilizer inputs, and N 207 deposition from the atmosphere. In particular, when the N cycle interacts with the C cycle, the 208 maximum photosynthetic capacities of model PFTs (V_{c,max}) are determined prognostically as a 209 function of their leaf N content (Asaadi and Arora, 2021; Kou-Giesbrecht and Arora, 2022). When the N cycle is turned off, prescribed PFT-specific V_{c,max} rates are used (Melton and Arora, 2016a) 210 and an empirical downregulation parameterization is used to emulate the effect of nutrient 211 constraints as atmospheric CO₂ increases (Arora et al., 2009). N in all model components (leaves, 212 213 stem, roots, litter, and soil organic matter) is prognostically tracked, and therefore C:N ratio of 214 all components is prognostically modelled except for soil organic matter for which a C:N ratio of 13 is specified. In addition, N in the soil mineral pools of nitrate (NO₃⁻) and ammonium (NH₄⁺) is 215 216 also prognostically modelled.

3. Driving data for CLASSIC and model simulations

218 **3.1 Land cover**

Land cover is one of the most important geophysical fields that is required by LSMs and 219 220 at its most basic level provides information about fractional vegetation cover in each grid cell for 221 a given regional or global domain. Vegetation in LSMs is typically represented in terms of PFTs. 222 Models may choose to represent a basic set of a few PFTs (trees, grasses, shrubs, and crops) or a more elaborate set that distinguishes PFTs based on their stature (trees, grasses, or shrubs), leaf 223 form (needleleaf or broadleaf), leaf phenology (evergreen or deciduous), photosynthetic 224 225 pathway (C_3 or C_4), and geographical location (tropical, temperate, or boreal). The version of 226 CLASSIC in this study uses a somewhat smaller set of nine PFTs for biogeochemical processes as 227 described in the previous section. The fractional coverage of PFTs in a model may be dynamically 228 simulated based on competition between PFTs or prescribed based on observation-based land 229 cover information. While CLASSIC does have a parameterization of competition between its PFTs 230 (Arora and Boer, 2006; Melton and Arora, 2016b), for the historical simulations considered here and for the simulations that contribute to the TRENDY ensemble, prescribed fractional coverage 231 of PFTs is used. 232

For the process of generating a historical reconstruction of land cover, consisting of time-233 varying fractional coverage of a model's PFTs, two types of observation-based data sets are used. 234 235 The first is a remotely-sensed land cover product that represents the geographical distribution of land cover for the present day for a short period. Examples of this include the GLC 2000 land 236 237 cover product which represents November 1999 to December 2000 period 238 (https://forobs.jrc.ec.europa.eu/products/glc2000/glc2000.php) and the more recent European Space Agency (ESA) Climate Change Initiative (CCI) land cover product for the period 1992-2018 239

(ESA, 2017). The second type of data set required to reconstruct historical land cover is that of a
spatially and temporally varying cropland (and pasture) area for a much longer period, which in
this case is based on the data set provided by the land use harmonization (LUH) product as part
of the TRENDY protocol for the period 850-2018. The LUH product is comprehensive (Hurtt et al.,
2020b). For example, not all models use the pasture area and other information provided in the
LUH product.

The process of generating land cover for a given model's PFTs is a three-step process. 246 247 First, the fractional coverage of model PFTs is obtained from a remotely sensed land cover 248 product that represents the snapshot of land cover for the present day. This requires typically mapping 20 – 40 land cover classes that exist in a remotely-sensed land cover product to a given 249 250 model's PFTs. This step introduces the largest uncertainty in the entire process. The original land 251 cover in the CLASSIC model is based on the GLC 2000 land cover product. Table 2 of Wang et al. 252 (2006) summarizes the mapping/reclassification of the 22 GLC 2000 land cover categories to the nine PFTs used in CLASSIC. Each land cover class was split into one or more of the nine CLASSIC 253 PFTs based on the class description and knowledge of global biomes. For example, the discrete 254 "broadleaf deciduous open tree cover" category of the GLC 2000 product is assumed to consist 255 256 of 60% broadleaf deciduous trees, 20% grasses, and 20% bare ground. This first step yields a 257 snapshot of land cover expressed in terms of the fractional coverage of CLASSIC's nine PFTs 258 corresponding to the time associated with the land cover product (e.g. for year 2000 for the GLC 2000 land cover product). The second step of generating fractional coverage of PFTs for a given 259 snapshot in time requires replacing the fractional area of crop categories with values from the 260 261 LUH data set for the same year. For example, when using the GLC 2000 land cover product, the

area of C₃ and C₄ crops from the LUH data set for the year 2000 are used, and the fractional 262 263 coverage of the other seven non-crop CLASSIC PFTs is adjusted such that the total vegetation fraction in each grid cell stays the same. Finally, in the last step, the temporally varying crop area 264 from the LUH product is used to go backward in time to 1700 from the year 2000 with typically 265 266 decreasing crop area while the area of other non-crop PFTs is adjusted in proportion to their existing fractional coverage such that the total vegetation fraction in each grid cell stays the 267 same. Similarly, the area of C_3 and C_4 crops from the LUH product is used from the year 2000 268 onwards to the present day with changing crop area and the area of non-crop PFTs is adjusted 269 270 such that the total vegetation fraction in each grid cell stays the same. All these steps yield a 271 reconstruction of historical land cover, expressed in terms of fractional coverage of CLASSIC's 272 nine PFTs (as interpreted from the GLC 2000 land cover product), from 1700 to 2018, in which crop area changes spatially and temporally according to the LUH product. 273

274 GLC 2000 is an older land cover product and more recent land cover products are now available. Here, in addition to the GLC 2000 based land cover, we also use the European Space 275 Agency (ESA) Climate Change Initiative (CCI) land cover product. The ESA CCI land cover product 276 277 is available at 300 m spatial resolution for the period 1992-2018 and contains 37 land cover categories (ESA, 2017). We use the land cover from the year 1992 to create a snapshot of CLASSIC 278 279 PFTs for the present day. Although there is some interannual variability overall the total 280 vegetated area doesn't change substantially from 1992-2018 in the ESA-CCI land cover. A default mapping/reclassification table for converting the ESA CCI classes into PFTs is provided in its user 281 guide (ESA, 2017). However, it overestimates tree cover along the taiga-tundra transition zone 282 283 and underestimates it elsewhere in Canada (Wang et al., 2018, 2019). Wang et al. (2022) have

developed a new reclassification table for converting the 37 ESA CCI land cover categories to CLASSIC's nine PFTs which is used in this study. A high-resolution land cover map over Canada and a tree cover fraction data at 30 m resolution are used to compute the sub-pixel fractional composition of each class in the ESA CCI dataset, which is then used to inform the cross-walking reclassification procedure (Wang et al., 2022).



292

Figure 1: Comparison of zonally summed areas of total vegetation (a), grass (b), and tree (c) cover used in the CLASSIC model based on GLC 2000 (blue line) and ESA CCI (dark red line) land cover products to each other, to selected other models that participated in the 2020 TRENDY intercomparison (grey lines) for which land cover information was available, and to Li et al. (2018) (dotted black line) who analyzed the ESA CCI data. All data correspond to the 1992-2018 period. CLASSIC does not yet explicitly represents shrub PFTs. Tall shrubs are merged into tree PFTs in CLASSIC. For the Li et al. (2018) data plotted here, the shrub PFTs are combined with the tree PFTs for a consistent comparison to CLASSIC. 302 The above process yields two representations of land cover in which the geographical 303 distribution of CLASSIC PFTs is based on GLC 2000 and ESA CCI land cover products. Both these 304 representations include the same reconstruction of crop area over the historical period. Figure 1 illustrates the uncertainty in land cover by comparing zonally summed areas of total vegetation, 305 306 tree, and grass cover in CLASSIC, averaged over the period 1992-2018, when model land cover is based on the GLC 2000 (blue line) and ESA CCI (dark red line) land cover products. These two 307 308 estimates are also compared to selected other models that participated in the 2020 TRENDY 309 intercomparison (grey lines), also for the period 1992-2018, for which land cover information was 310 available, and to Li et al. (2018b) (dotted black line) who analyzed the ESA CCI data based on the 311 default reclassification table from the ESA CCI user guide. Figure 1 shows while there is relatively good agreement across TRENDY models in terms of total vegetation cover there's a much larger 312 313 uncertainty in its split between tree and grass PFTs. There are two reasons for the spread in total vegetated, treed, and grassed areas across TRENDY models. First, modelling groups use different 314 remotely sensed land cover products for obtaining fractional cover of their model PFTs. Second, 315 the current process of mapping/reclassifying 20-40 land cover classes of a land cover product to 316 a model's PFTs is mainly based on the class description and expert judgment which introduces 317 318 subjectiveness in the process. Compared to the GLC 2000 based land cover in the CLASSIC model, 319 the newer ESA CCI based land cover yields a somewhat higher total vegetation cover, a higher grass cover, and a somewhat lower tree cover area. Unlike the older GLC 2000 based land cover 320 used in CLASSIC, the newer ESA CCI based grass and tree cover area are within the range of the 321 322 TRENDY models reported here. Finally, Figure 1 also allows us to compare the results from the

analysis of Li et al. (2018b) for the ESA CCI land cover (dotted black line) to the ESA CCI reclassification for CLASSIC (dark red line) by (Wang et al., 2022). Li et al. (2018b) used the default mapping/reclassification table for converting the ESA CCI classes into PFTs. This comparison illustrates that the remapping of the ESA CCI land cover classes to CLASSIC's PFTs yields total vegetation, tree, and grass coverage that is broadly comparable to Li et al. (2018b) although some differences remain for the grasses.

Our framework accounts for the uncertainty in land cover representation. However, since 329 330 both land cover representations in our study account for the increase in crop area over the 331 historical period in the same way by adjusting the area of non-crop PFTs in proportion to their 332 existing coverage using the LUH product, our framework is unable to account for the uncertainty 333 associated with the implementation of LUC. Di Vittorio et al. (2018) quantify this uncertainty by 334 implementing several approaches to account for the increase in crop area over the historical 335 period in the framework of an integrated assessment model: by preferentially converting grasses and shrubs, by preferentially converting forests, and by proportionally adjusting areas of non-336 crop PFTs in a way similar to ours. LUC emissions are higher if the increase in crop area is 337 338 preferentially obtained by converting forests. A similar uncertainty analysis for LUC emissions is performed by Peng et al. (2017) using the ORCHIDEE land model who analyze the effect of using 339 340 different rules to incorporate the changes in crop and pasture area over the historical period. The 341 uncertainty related to incorporating LUC information to modify a model's land cover is further illustrated in Di Vittorio et al. (2014) and Meiyappan and Jain (2012). 342

343

344 3.2 Meteorological data

As a land surface component of an ESM, CLASSIC requires meteorological forcing at a sub-345 346 daily temporal resolution. In the offline simulations reported here, the model is run with halfhourly values of meteorological data (incoming long and shortwave radiation, temperature, 347 precipitation, specific humidity, wind speed, and pressure). The first meteorological data set used 348 349 to drive CLASSIC is from the TRENDY protocol for the year 2020, CRU-JRA v2.1.5, which provides 6 hourly values of the seven variables from the Japanese reanalysis (JRA) with monthly values 350 351 adjusted to the climate research unit's data (CRU, https://crudata.uea.ac.uk/cru/data/hrg/). This 352 yields a blended product from year January 1901 to December 2018 with the 6-hourly temporal 353 resolution of a reanalysis but without the biases that may be present in reanalysis data (Harris, 354 2020). The second meteorological data set used here to drive CLASSIC is from the Global Soil Wetness Project 3 (GSWP3). The GSWP3 forcing data are based on a dynamical downscaling of 355 the 20th century reanalysis (Compo et al., 2011) using a Global Spectral Model (GSM) run at about 356 50 km resolution. GSM is nudged towards the vertical structures of 20th century (20CR) zonal and 357 meridional air temperature and winds so that the synoptic features are retained at their higher 358 spatial resolution. Additional bias corrections are also performed as explained in van den Hurk et 359 al. (2016). The GSWP3 forcing is available for the 1901-2016 period. The 6-hourly values from 360 both the CRU-JRA and GSWP3 forcings are further disaggregated to half-hourly values for use by 361 CLASSIC. 362

Figure A1 compares the two meteorological forcings data sets, over the 1997-2016 period, to illustrate that although these two data sets are very similar there are differences between the two. Mean annual global precipitation over land (excluding Greenland and Antarctica) in the GSWP3 data set (71.4 mm/month, 857 mm/year) is somewhat higher than in

the CRU-JRA data set (68.3 mm/month, 820 mm/year). The global near-surface air temperature 367 368 over land (excluding Greenland and Antarctica) is also slightly higher in the GSWP3 data set (14.22 °C) compared to the CRU-JRA data set (14.08 °C). The largest temperature difference occurs 369 370 between the two data sets over the northern tropics (panel h) where the GSWP3 data set is about 371 0.93 °C warmer than the CRU-JRA data set. The geographical distribution of mean annual temperature is very similar between the two data sets but there are some differences in the 372 373 geographical distribution of precipitation (not shown). Despite very similar total precipitation 374 amounts and their seasonality over large global regions in the two data sets, differences exist in 375 the frequency distribution of precipitation. Figure A2 illustrates this over three broad regions, the 376 Amazon, the Sahel, and the Midwest United States, which shows the frequency distribution of 377 daily precipitation amounts (mm/day) over the 1997-2016 period from the two data sets. Figure A2 shows that the frequency of precipitation events greater than about 5-10 mm/day is higher 378 379 in the GSWP3 data set compared to the CRU-JRA data set for the Amazonian, the Sahel, and the 380 Midwest United States regions.

381 3.3 Other forcings

Other than the land cover and meteorological forcings CLASSIC requires globally averaged atmospheric CO₂ concentration, geographically varying time-invariant soil texture and soil permeable depth, population density, monthly climatological lightning, and geographically and time-varying N fertilizer application rates and atmospheric N deposition rates. The atmospheric CO₂ concentration values are provided by the TRENDY protocol. The soil texture information consists of the percentage of sand, clay, and organic matter and is derived from Shangguan et al. (2014). N fertilizer is specified according to the TRENDY protocol and based on Lu and Tian

- 389 (2017). N deposition is also specified according to the TRENDY protocol and based on model
- 390 forcings provided for the sixth phase of CMIP (CMIP6) through input4MIPs (Hegglin et al.,
- 2016). N deposition for the historical (1850-2014) period is used as is provided while that for
- the period 2015-2018 is specified based on N deposition from the SSP5-85 scenario. For the
- 393 period 1700-1849, N deposition values from the year 1850 are used.
- 394
- Table 1: Summary of simulations performed with two representations of the historical land
- cover, two sets of meteorological data, and two versions of the CLASSIC land model.
- 397

Simulation	Land cover reconstruction	Meteorological forcing	N cycle interactions with the C cycle
1	based on GLC 2000	CRU-JRA v2.1.5	On
2	based on GLC 2000	GSWP3	On
3	based on GLC 2000	CRU-JRA v2.1.5	Off
4	based on GLC 2000	GSWP3	Off
5	based on ESA CCI	CRU-JRA v2.1.5	On
6	based on ESA CCI	GSWP3	On
7	based on ESA CCI	CRU-JRA v2.1.5	Off
8	based on ESA CCI	GSWP3	Off

- 398
- 399
- 400

401 **3.4 Model simulations**

Using the two representations of the historical land cover (based on the GLC 2000 and ESA CCI land cover products), the two sets of meteorological data (CRU-JRA and GSWP3), and the two versions of the CLASSIC model (with and without interactions between the C and N cycles) we perform eight sets of pre-industrial and historical simulations as summarized in Table 1. Preindustrial simulations that correspond to the year 1700 are required before doing the historical simulations (from which we analyze the model results) so that model pools can be spun up to near equilibrium for each combination of land cover, meteorological forcing, and model

version. The pre-industrial simulations use 1901-1925 meteorological data repeatedly since this 409 410 period shows little trends in meteorological variables. Global thresholds of net atmosphereland C flux of 0.05 Pg/yr and net atmosphere-land N flux of 0.5 Tg N/yr, in simulations with the 411 412 N cycle turned on, are used to ensure the model pools have reached equilibrium. Each historical 413 simulation is then initialized from its corresponding pre-industrial simulation after it has reached equilibrium. Simulations driven with the CRU-JRA meteorological data are performed 414 for the period 1701-2018, and the period 1701-2016 for simulations driven with the GSWP3 415 416 meteorological data, although results are reported for the period 1997-2016 which is common 417 to both simulations. Similar to the pre-industrial simulations, meteorological data from 1901-1925 is used repeatedly for the period 1701-1900. The global model simulations are performed 418 419 at a spatial resolution of about 2.81° (about 312 km at the equator) and the size of the spatial 420 longitude-latitude grid is 128×64 grid cells. All model forcings are regridded to this common spatial resolution. The model is run over about 1900 land grid cells at this resolution excluding 421 422 glacial cells in Greenland and Antarctica.

423

424 **3.5 Automated benchmarking**

The results from the eight CLASSIC simulations reported here are evaluated using an inhouse model benchmarking system called the Automated Model Benchmarking R package (AMBER) (Seiler et al., 2021b). AMBER is based on a skill score system originally developed by (Collier et al., 2018) which is used to quantify model performance and explained in detail in the appendix. Five scores are used that assess a model's bias (S_{bias}), root-mean-square error (S_{rmse}), seasonality (S_{phase}), interannual variability (S_{iav}), and spatial distribution (S_{dist}) against globally

gridded and in-situ data set(s) of observation-based estimates for a given quantity. A score is
computed by first calculating a dimensionless statistical metric, which is then scaled onto a unit
interval, and finally calculating its spatial mean. Scores range from 0 to 1 and are dimensionless.
Higher values indicate better performance. Finally, an overall score S_{overall} is calculated as follows
by giving twice as much weight to S_{rmse}

436
$$S_{overall} = \frac{S_{bias} + 2S_{rmse} + S_{phase} + S_{iav} + S_{dist}}{1 + 2 + 1 + 1 + 1}.$$
 (1)

437 The decision to give extra weight to S_{rmse} is entirely subjective but follows Collier et al. (2018).

438 The scores are calculated by comparing gridded and in-situ observation-based estimates, 439 referred to as reference data sets in Seiler et al. (2021b), for 19 energy (surface albedo, net shortwave and longwave radiation, total net radiation, latent heat flux, sensible heat flux, ground 440 heat flux), water (soil moisture, snow, and runoff), and C cycle (GPP, ecosystem respiration, net 441 ecosystem exchange, net biome productivity, aboveground biomass, soil C, LAI, area burnt, and 442 443 fire CO₂ emissions) related variables to model simulated quantities. Table 2 summarizes the 444 source of these observation-based data sets. The resulting model scores express to what extent simulated and observation-based data agree. A low score does not necessarily indicate poor 445 model performance. Uncertainties in the meteorological forcing data and geophysical fields used 446 to drive the model, and/or in the observation-based data itself are possible reasons for the lack 447 of agreement. One way to assess uncertainties in observation-based data sets is to quantify the 448 449 skill score by comparing two independently-derived observation-based data sets (Seiler et al., 2022). The resulting scores are referred to as benchmark scores and quantify the level of 450 agreement among the observation-based data sets themselves provided, of course, there are at 451

452 Table 2: Observation-based data sets used for model evaluation in AM
--

			-
Globally gridded variable(s)	Source	Approach used	Reference
Leaf area index	AVHRR	Artificial neural network	Claverie et al. (2016)
Net biome productivity	CAMS	Atmospheric inversion	Agustí-Panareda et al. (2019)
Net biome productivity	Carboscope	Atmospheric inversion	Rödenbeck et al. (2018)
Surface albedo, net shortwave and	CERES	Radiative transfer model	Kato et al. (2013)
longwave radiation, net radiation			
Net radiation, latent and sensible heat	CLASSr	Blended product	Hobeichi et al. (2019)
flux, ground heat flux, runoff			
Leaf area index	Copernicus	Artificial neural network	Verger et al. (2014)
Net biome productivity	CT2019	Atmospheric inversion	Jacobson et al. (2020)
Fire CO ₂ emissions	CT2019	Atmospheric inversion	Jacobson et al. (2020)
Snow amount	ECCC	Blended product	Mudryk (2020)
Liquid soil moisture	ESA	Land surface model	Liu et al. (2011)
Area burnt	ESA CCI	Burned area mapping	Chuvieco et al. (2018)
Latent and sensible heat flux, gross	FLUXCOM	Machine learning	Jung et al. (2019, 2020)
primary productivity			
Above ground biomass	GEOCARBON	Machine learning	Avitabile et al. (2016); Santoro
			et al. (2015)
Surface albedo, net shortwave and	GEWEXSRB	Radiative transfer model	Stackhouse et al. (2011)
longwave radiation, net radiation			
Area burnt	GFED 4s	Burned area mapping	Giglio et al. (2010)
Gross primary productivity	GOSIF	Statistical model	Li and Xiao (2019)
Soil carbon	HWSD	Soil inventory	Wieder (2014); Todd-Brown et
			al. (2013)
Surface albedo	MODIS	Bidirectional Reflectance	Strahler et al. (1999)
		Distribution Function	
Gross primary productivity	MODIS	Light use efficiency model	Zhang et al. (2017)
Leaf area index	MODIS	Radiative transfer model	Myneni et al. (2002)
Soil carbon	SGS250m	Machine learning	Hengl et al. (2017)
Above ground biomass	Zhang	Data fusion	Zhang and Liang (2020)
In situ variable(s)	Source	Approach used (number of	Reference
		sites)	
Leaf area index	CEOS	Transfer function (141)	Garrigues et al. (2008)
Latent, sensible, and ground heat flux,	FLUXNET 2015	Eddy covariance (204)	Pastorello et al. (2020)
gross primary productivity, ecosystem			
respiration, net ecosystem exchange			
Above ground biomass	FOS	Allometry (274)	Schepaschenko et al. (2019)
Runoff	GRDC	Gauge records (50)	Dai and Trenberth (2002)
Snow amount	Mortimer	Gravimetry (3271)	Mortimer et al. (2020)
Above ground biomass	Xue	Allometry (1974)	Xue et al. (2017)
2			· · · · ·

454 least two sets of observation-based data for a given quantity. The comparison of model scores

against benchmark scores then shows how well a model-simulated quantity compares to the

456 reference data sets relative to the agreement between the observation-based data sets457 themselves.

458

459 **4. Results**

Figures 2 through 9 show the time series and/or zonally-averaged values of annual values 460 461 of a variable of interest when averaged across four ensemble members each according to whether the N cycle is turned on or not, whether the GLC 2000 or ESA CCI based land cover is 462 used, and whether model simulations are driven by the CRU-JRA or GSWP3 meteorological data. 463 Figures A3, A4, A6, A7, A9, and A11 in the appendix, which are complementary to the above-464 mentioned figures, show the physical and biogeochemical states of the land surface and primary 465 physical fluxes of water and energy, and primary biogeochemical fluxes of CO₂ simulated by 466 467 CLASSIC at the land-atmosphere boundary for all the eight simulations considered here. While the figures in the appendix illustrate the range in simulated physical and biogeochemical states 468 and fluxes across the eight simulations, Figures 2 through 9 evaluate the effect of model 469 470 structure, meteorological forcing, and land cover on a given quantity. We also quantify the spread across the eight simulations in Table 3 using the coefficient of variation (cv= standard 471 472 deviation/mean) calculated using annual global values for a given quantity averaged over the 1997-2016 20-year period of each simulation. This time period is also used for other reported 473 results. 474

475 **4.1 Physical land surface state and fluxes**

476



Figure 2: Time series of annual globally-averaged soil moisture in the top 1m averaged over the four ensemble members that are driven with and without an interactive N cycle (panel a), driven with the GLC 2000 and ESA CCI based land cover representations (panel b), and driven with the GSWP3 and CRU-JRA meteorological data (panel c). The thin lines show the individual years and the thick lines show their 11-year moving average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, for each ensemble averaged over its set of four simulations are also shown.

Figure A3, panels a and b, shows the globally-averaged simulated soil moisture and temperature in the top 1 m soil layer. While simulated soil temperature in the top 1 m is fairly similar across the eight simulations, the simulated soil moisture is distinctly separated into two groups. The separation into these two groups is caused by the driving meteorological data as shown in Figure 2. The coefficient of variation for soil moisture and temperature values averaged over the 1997-2016 period of each simulation are 0.02 and 0.004, respectively, indicating that overall the variation in these quantities is relatively small compared to their means. The use of the GSWP3 meteorological dataset yields slightly higher (~4%) globally-averaged soil moisture
compared to the CRU-JRA meteorological data set (236.5 mm vs. 227.1 mm, Figure 2c) for the
1997-2016 period.

494 Figure A3, panels c and d, shows the simulated fluxes of global evapotranspiration and runoff across the eight simulations. Similar to soil moisture, evapotranspiration and runoff also 495 fall broadly into two groups and the reason for this again is the driving meteorological data. 496 Figure 3 shows that while the biggest factor that affects evapotranspiration and runoff is the 497 498 difference in driving meteorological data the interactive N cycle also affects these water fluxes. 499 Neither evapotranspiration nor runoff is significantly affected by the choice of land cover. The 500 reason an interactive N cycle affects evapotranspiration is that the N cycle in CLASSIC affects the rate of photosynthesis through the prognostic determination of leaf N content. Photosynthesis 501 502 in turn affects canopy conductance, which affects transpiration through the canopy leaves. 503 Average evapotranspiration over the 1997-2016 period of the simulations driven with GSWP3 meteorological data is about 9% lower than in simulations driven with CRU-JRA meteorological 504 data (65.8 vs. 72.1 ×1000 km³/year, Figure 3, panel e). An interactive N cycle reduces 505 506 evapotranspiration by about 2% due to lower photosynthesis rates as shown later (Figure 3, panel a). Average runoff is about 27% higher in simulations driven with GSWP3 compared to 507 simulations driven with CRU-JRA meteorological data (52.6 vs 41.3 ×1000 km³/year, Figure 3, 508 509 panel f). This is due to slightly high precipitation in the GSWP3 meteorological data set (Figure A1) but is more so due to the simulated lower evapotranspiration when using the GSWP3 data 510 (Figure 3, panel e). The coefficient of variation for evapotranspiration and runoff values averaged 511 over the last 20 years of each simulation are 0.05 and 0.13, respectively. 512



Figure 3: Time series of annual global evapotranspiration and runoff (over all land area excluding 514 515 Greenland and Antarctica) averaged over the four ensemble members that are driven with and without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land 516 cover (panels c, d), and driven with the GSWP3 and CRU-JRA meteorological data (panels e, f). 517 The thin lines show the individual years and the thick lines show their 11-year moving average. 518 Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time 519 periods, and their difference, for each ensemble averaged over its set of four simulations are also 520 521 shown.







Figure A4 shows the primary energy fluxes from the eight simulations. These include net 530 531 downward shortwave and longwave radiation, and latent and sensible heat fluxes. Incoming shortwave and longwave radiation are part of the driving meteorological data. Similar to water 532 fluxes, the differences in energy fluxes in CLASSIC are also primarily driven by differences in 533 534 meteorological data (Figure A4, A5, and Figure 4). Net shortwave radiation (Figure A4, panel a) is equal to incoming shortwave radiation minus the fraction that is reflected back. Net longwave 535 radiation (Figure A4, panel b) is equal to incoming longwave radiation minus the longwave 536 537 radiation emitted by the land based on its surface temperature following the Stefan-Boltzmann law. The difference in net shortwave radiation is also affected by simulated vegetation biomass 538 and leaf area index. The latter affects surface albedo which determines what fraction of incoming 539 shortwave radiation is reflected. This is the reason why an interactive N cycle affects net 540 shortwave radiation since the N cycle affects photosynthesis, and in turn, simulated vegetation 541 542 biomass and leaf area index (Figure A5, panel b). Latent heat flux is affected primarily by meteorological data (Figure 4) but also if the N cycle is interactive or not since it is essentially 543 evapotranspiration but in energy units. Finally, differences in sensible heat fluxes are strongly 544 affected by differences in driving meteorological data (Figure 4). Globally-averaged sensible heat 545 flux in the simulations driven with GSWP3 data is ~14% higher compared to CRU-JRA driven 546 simulations (40.2 vs. 35.0 W/m²). The coefficient of variation for latent and sensible heat flux 547 values averaged over the last 20 years of each simulation are 0.05 and 0.07, respectively. Net 548 shortwave (cv=0.006) and longwave (cv=0.03) radiative fluxes vary little across the eight 549 simulations. 550



Figure 5: Time series of annual global vegetation C mass (over all land area excluding Greenland 552 and Antarctica) (panels a, c, and e) and zonally-averaged values of vegetation C mass over land 553 (panels b, d, and f) averaged over the four ensemble members, for the period 1997-2016, that 554 555 are driven with and without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land cover (panels c, d), and driven with GSWP3 and CRU-JRA meteorological 556 557 data (panels e, f). The thin lines for the time series show the individual years and the thick lines show their 11-year moving average. Model values averaged over the pre-industrial (1851-1860) 558 559 and present-day (1997-2016) time periods, and their difference, are also shown in panels a, c, 560 and e.

562 4.2 Biogeochemical land surface state and fluxes

563 4.2.1 Primary CO₂ fluxes and C pools

Figure A6 shows the simulated C state of the land surface expressed in terms of vegetation 564 and soil C pools. Panels a and b show the annual time series of global vegetation and soil C mass 565 from the eight simulations, and panels c and d show their zonally-averaged distributions 566 averaged over the last 20 years of each simulation. The biggest difference in the time series of 567 568 global vegetation (cv=0.16) and soil (cv=0.21) C mass compared to soil moisture and temperature, which characterized the physical land surface state, is the large spread across the 569 570 eight simulations as indicated by their high cv values. The zonally-averaged values further provide 571 insight into the reasons for this spread and show that the largest differences between simulated vegetation and soil C occur at northern high latitudes (north of about 40°N). Panels c and d of 572 Figure A6 also show observation-based zonally-averaged values of vegetation and soil C mass 573 574 based on the Reusch and Gibbs (2008) and the Harmonized World Soils Database (v1.2) (Fischer et al., 2008), respectively, to provide a reference. A more thorough comparison with observations 575 is provided in Section 4.3. 576

577 Differences in vegetation C mass are caused primarily when the N cycle is interactive or 578 not (Figure 5). Both land cover and the driving meteorological data play a smaller role in the 579 simulated spread of vegetation C mass (Figure 5). The ESA CCI based land cover has a larger 580 vegetated area but most of this increase comes from an increase in the area of grasses that do 581 not store a lot of C in their vegetation C mass. The spread in simulated soil C is caused due to the



583

Figure 6: Time series of annual global soil carbon mass (over all land area excluding Greenland 584 585 and Antarctica) (panels a, c, and e) and zonally-averaged values of soil carbon mass over land (panels b, d, and f) averaged over the four ensemble members, for the period 1997-2016, that 586 are driven with and without an interactive N cycle (panels a, b), driven with the GLC 2000 and 587 588 ESA CCI based land cover (panels c, d), and driven with GSWP3 and CRU-JRA meteorological data (panels e, f). The thin lines for the time series show the individual years and the thick lines show 589 590 their 11-year moving average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown in panels a, c, and e. 591



592 Figure 7: Time series of annual global gross primary productivity (over all land area excluding Greenland and Antarctica) (panels a, c, and e) and zonally-averaged values of gross primary 593 productivity over land (panels b, d, and f) averaged over the four ensemble members, for the 594 period 1997-2016, that are driven with and without an interactive N cycle (panels a, b), driven 595 with the GLC 2000 and ESA CCI based land cover (panels c, d), and driven with GSWP3 and CRU-596 JRA meteorological data (panels e, f). The thin lines for the time series show the individual years 597 598 and the thick lines show their 11-year moving average. Model values averaged over the preindustrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also 599 600 shown in panels a, c, and e.

N cycle but also the choice of land cover (Figure 6). Since CLASSIC assumes that litter from grasses 601 602 is more recalcitrant than that from trees, the choice of ESA CCI based land cover leads to a higher soil C mass because it has a higher grass area than the GLC 2000 based land cover (Figure 6, 603 panels c and d). The choice of meteorological data does not affect the magnitude of simulated 604 605 globally-summed soil C mass significantly but does affect its change over the historical period. In Figure 6 (panel c) the decrease in soil C mass from the 1851-1860 period to the 1997-2016 period 606 is higher when using the GSWP3 (29.9 Pg C) compared to when using the CRU-JRA (14.8 Pg C) 607 608 meteorological data.

609 The reason why an interactive N cycle in CLASSIC affects vegetation C and soil C mass, and 610 why the ESA CCI based land cover yields high soil C, is seen in Figures A7 and 7. Figure A7 shows the spread of primary C fluxes including gross primary productivity (GPP) (cv=0.07), and 611 612 autotrophic (cv=0.04) and heterotrophic (cv=0.10) respiratory fluxes, across the eight 613 simulations. Since GPP is lower in the runs with the N cycle, both vegetation (Figure 5a) and soil C mass (Figure 6a) are also lower. The lower GPP in the runs with the N cycle is due primarily to 614 lower GPP at high latitudes (Figure 7b) which yields low vegetation C mass at high latitudes 615 (Figure 5b). Low GPP at high latitudes translates to even larger relative differences in soil C given 616 the longer turnover time scales of soil C at high latitudes (Figure 6b). The use of the ESA CCI based 617 618 land cover which has a higher grass area than the GLC 2000 based land cover leads to higher GPP 619 (Figure 7d) and therefore higher soil C at all latitudes (Figure 6d). In Figure A8, global heterotrophic and autotrophic respiratory fluxes are most affected by land cover and the 620 inclusion or absence of an interactive N cycle but not as much by the driving meteorological data. 621



Figure 8: Time series of annual area burned (over all land area excluding Greenland and 623 Antarctica) (panels a, c, and e) and zonally-averaged values of area burned (panels b, d, and f) 624 averaged over the four ensemble members, for the period 1997-2016, that are driven with and 625 without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land 626 cover (panels c, d), and driven with GSWP3 and CRU-JRA meteorological data (panels e, f). The 627 628 thin lines for the time series show the individual years and the thick lines show their 11-year moving average in panels (a), (c), and (e). Model values averaged over the pre-industrial (1851-629 630 1860) and present-day (1997-2016) time periods, and their difference, are also shown for 631 panels (a), (c), and (e). 632

The transient behaviour of heterotrophic respiration over the historical period is not affected by
 meteorological data, although the effect of meteorological data on autrotrophic respiration
 varies over time.

636

4.2.2 Area burned and fire CO₂ emissions

Figure A9 shows the time series of global area burned and global fire CO₂ emissions, and 637 their zonally-averaged values. We chose the area burned (cv=0.24) and fire CO₂ emissions 638 639 (cv=0.21) in addition to the primary biogeochemical fluxes since fire shows large variability both 640 in space and in time, and both these variables yield the largest spread across the eight 641 simulations, among all the fluxes and simulated quantities considered here. Figures A9 (panels c 642 and d) also show observation-based estimates for area burned and fire CO₂ emissions based on 643 GFED 4s (Giglio et al., 2013) to provide an observation-based context. Figures 8 and A10 help us understand which factors contribute to this large variability. The variability in the area burned is 644 caused primarily by the choice of land cover and meteorological data and the variability is higher 645 in the southern hemisphere (Figure 8, panels d and f). An interactive N cycle does not affect the 646 zonal distribution of area burned and fire CO₂ emissions (Figures 8 and A10) as much. The reason 647 both area burned and fire CO₂ emissions are affected by the choice of land cover is because the 648 ESA CCI land cover has higher grass area and, as a result, it yields higher area burned and fire CO₂ 649 650 emissions since a larger area is burned for grasses than for trees in the model. The choice of driving meteorological data is a factor in the area burned and our simulations show that the use 651 652 of GSWP3 meteorological forcing yields a higher area burned than the CRU-JRA data. In particular wind speed, which determines the rate of spread of fire in CLASSIC, is much higher in the GWSP3 653 than in the CRU-JRA meteorological data. Globally-averaged land wind speed (excluding 654
Greenland and Antarctica) in GSWP3 data is 6.1 m/s compared to 3.4 m/s in the CRU-JRA data

656 for the period 2000-2016.

Table 3: Simulated energy, water, and carbon cycle quantities considered in this study sorted

according to their coefficient of variation. The quantities are listed from the most variable at

the top to the least variable at the bottom. The coefficient of variation is based on annual

values averaged over the 1997-2016 period across the eight simulations. The last column shows
 the dominant source of variability for each model simulated quantity.

662

Energy, water, or carbon cycle quantities	Coefficient	Dominant source of variability
	of variation	
Area burned (million km ²)	0.24	Land cover
Fire CO ₂ emissions (Pg C/year)	0.21	Land cover
Soil carbon mass (Pg C)	0.21	The inclusion or the absence of
Vegetation carbon mass (Pg C)	0.16	The inclusion or the absence of the N cycle
Runoff (1000 km³/year)	0.13	Meteorological forcing
Leaf area index (m ² /m ²)	0.11	The inclusion or the absence of
		the N cycle
Heterotrophic respiration (Pg C/year)	0.10	Land cover
Gross primary productivity (Pg C/year)	0.07	Land cover
Sensible heat flux (W/m ²)	0.07	Meteorological forcing
Autotrophic respiration (Pg C/year)	0.04	Land cover
Latent heat flux (W/m ²) / Evapotranspiration	0.05	Meteorological forcing
(1000 km ³ /year)		
Net longwave radiation (W/m ²)	0.03	Meteorological forcing
Soil moisture in the top 1m soil layer (mm)	0.02	Meteorological forcing
Albedo for shortwave radiation (fraction)	0.008	The inclusion or the absence of
		the N cycle
Net shortwave radiation (W/m ²)	0.006	Meteorological forcing
Soil temperature in the top 1m soil layer (°C)	0.004	Meteorological forcing

663

664 **4.2.3 Coefficient of variation summary**

Table 3 shows the energy, water, and C-related quantities considered so far but also leaf

area index and albedo and lists them from the most variable at the top to the least variable at

the bottom according to their coefficient of variation. The area burned is found to be the most 667 668 variable quantity and soil temperature is the least variable quantity. Table 3 also shows the most dominant source of variability for each simulated quantity: land cover, meteorological forcings, 669 670 or the inclusion or absence of an interactive N cycle. Net atmosphere-land CO₂ flux (or net biome 671 productivity), net ecosystem exchange, and ground heat flux are not included in Table 3 because these fluxes are calculated as the difference of larger fluxes and as a result, their values are closer 672 to zero which yields a large value of the coefficient of variation. Net surface radiation is the sum 673 674 of net shortwave and longwave radiation and both of them exhibit low coefficient of variability 675 across the eight simulations (Table 3).

676 **4.2.4 Model tuning**

Overall, the results presented so far illustrate that different model simulated quantities 677 are sensitive to different forcings and model versions. The use of more than one meteorological 678 forcing data sets and land cover representation, and the use of two model versions (with and 679 680 without N cycle), yields a dilemma since it is no longer possible to tune model parameters without 681 choosing a preferred meteorological data set, land cover representation, and model version. As 682 such it seems logical that rather than tuning the model for a preferred forcing or model version, model results from an ensemble of simulations be compared against an ensemble of 683 observations in so long as it is possible. This is the approach taken in Section 4.3 with automated 684 685 benchmarking.

686 **4.2.5 Net biome productivity**



Figure 9: Time series of global net atmosphere-land CO₂ flux (over all land area excluding 687 Greenland and Antarctica) (panels a, c, and e) and its zonally-averaged values (panels b, d, and f) 688 689 averaged over the four ensemble members, for the period 1997-2016, that are driven with and 690 without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land 691 cover (panels c, d), and driven with GSWP3 and CRU-JRA meteorological data (panels e, f). The thin lines for the time series show the individual years and the thick lines show their 11-year 692 693 moving average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown for panels (a), (c), and (e). 694

Figure A11 shows the spread in the time series of annual global net atmosphere-land CO₂ flux and their zonally-averaged values across the eight simulations averaged over the 1997-2016 period from each simulation. The global net atmosphere-land CO₂ flux or net biome productivity (NBP) is considered a critical determinant of the performance of LSMs, and is treated as such by TRENDY, because this flux ultimately affects the changes in the atmospheric CO₂ burden. TRENDY requires that LSMs simulate a terrestrial C sink for the decades of the 1990s to the present to be considered for inclusion in the TRENDY ensemble.

702 Figure A11 also shows the estimates of global net atmosphere-land CO₂ flux from the 703 participating TRENDY models in grey boxes with mean and shaded ranges for the decades from the 1960s to 2010s from the Global Carbon Project (Friedlingstein et al., 2022). Positive values in 704 705 Figure A11 indicate a C sink over land and negative values a C source to the atmosphere. In Figure 706 A11a, all eight simulations reported here would qualify for inclusion in the TRENDY ensemble 707 since they all simulate a terrestrial C sink from the 1990s to the present day. Before 1960, since 708 the atmospheric CO_2 concentration is not high enough, the model yields both a land C sink and source in response to interannual variability in meteorological data. In addition, the time series 709 710 of global NBP from all eight simulations lie within the uncertainty range of reported estimates from the Global Carbon Project. Figure A11a suggests that based on global NBP, at least, it is not 711 712 possible to exclude any of the eight simulations. In Figure A11b, zonally-averaged NBP averaged 713 over the 1997-2016 period from each of the eight simulations mostly lie within the range of NBP simulated by models that participated in TRENDY 2020. CLASSIC simulates a C sink at northern 714 high latitudes consistent with TRENDY models but it simulates a C sink on the stronger side of 715 TRENDY models in the southern tropics (0° - 20°S). This is likely because CLASSIC is known to 716

simulate low C emissions associated with LUC most of which are generated in tropical regions(Asaadi and Arora, 2021).

719 Figure 9 provides additional insights into the effect of different forcings on the simulated 720 NBP. In Figure 9, averaged over the 1997-2016 period, an interactive N cycle leads to a somewhat 721 weaker C sink (panel a, 0.98 vs. 1.11 Pg C/yr), the choice of the ESA CCI based land cover leads to a somewhat stronger C sink (panel c, 1.14 vs 0.94 Pg C/yr), and the choice of the GSWP3 722 meteorological data leads to a much weaker C sink (panel e, 0.74 vs 1.33 Pg C/yr) than the CRU-723 724 JRA meteorological data. In Figure 9, panels a and b, the largest difference between the model 725 versions with and without the N cycle occurs in the tropics (~ 5°N - 20°S) where an interactive N cycle leads to a weaker C sink. There are differences in zonally-averaged NBP with and without 726 727 the N cycle south of 45°S but the land area below this latitude is small so the averages are 728 calculated over only a few grid cells. The choice of the land cover (Figure 9, panels c and d) does 729 not substantially change the distribution of the zonally-averaged values of NBP although, as noted above, the choice of ESA CCI based land cover leads to a somewhat stronger C sink. Finally, 730 the choice of the GSWP3 meteorological forcing leads to a weaker C sink at most latitudes (Figure 731 9, panels e and f). 732

733 4.3 Automated benchmarking

Figure 10 plots the overall score, S_{overall}, against benchmark scores for 16 of the 19 energy, water, and C cycle related variables using which AMBER calculated model and benchmark scores. AMBER does not yet evaluate N cycle related variables for which observations are more scarce than for C cycle related variables. The range in model scores comes from the eight simulations,

and the range in benchmark scores comes from the different observation-based data sets. The 738 739 whiskers show the range in the overall score both for the benchmark and model scores. The vertical whiskers show the range of eight model scores when a given variable from all eight model 740 simulations is compared to an observation-based data set. The horizontal whiskers show the 741 742 range when three or more observation-based datasets are compared to each other. When only two observation-based data sets are compared to each other there is only one benchmark score, 743 and therefore there is no range. In Figure 10, three quantities are missing: soil moisture, 744 745 ecosystem respiration, and fire CO₂ emissions since there is only one observation-based 746 reference data used for these variables and therefore a benchmark score cannot be calculated. Figure 10 shows that typically as the benchmark scores increase so do the overall model scores 747 748 or a given quantity. This indicates that uncertainty in observation-based estimates themselves leads to a poor agreement between observations and model-simulated quantities. 749

750 For energy and water fluxes scores (panels a and b) the model overall scores lie around the 1:1 line indicating that model scores are generally as good as the benchmark scores, except 751 for surface albedo (ALBS), runoff (MRRO), ground heat flux (HFG), and comparison against one 752 753 observation-based estimate of snow water equivalent which lie below the 1:1 line. For C cycle 754 related variables most scores lie somewhat below the 1:1 line indicating that simulated quantities 755 do not agree as well with observations as observations agree among themselves. The lower 756 benchmark score for soil C (panel c) is because the SoilGrids250m (SG250m) data and the Harmonized World Soil Database (HWSD) do not agree well amongst themselves because the 757 SG250m soil C data includes peatlands and permafrost C at high latitudes while the HWSD data 758 759 does not (see Figure 11b). Since the version of CLASSIC used here does not represent peatlands

and permafrost C it compares better with the HWSD data than with the SG250m data. In the case of soil C, the choice of HSWD data for comparison against model values is obvious. However, for other variables, it may not always be obvious which observation-based estimate is more appropriate or better for comparison against model results. The uncertainty in forcing data sets and in observation-based estimates, against which model results are evaluated, implies that even a perfect model cannot be evaluated to its fullest extent.



Figure 10: Comparison of benchmark scores with model overall scores for a range of energy-, water-, and carbon-related quantities. The whiskers indicate the range for benchmark scores across different observation-based data sets and the range across the eight model simulations for the overall model scores. The quantities in panel (a) are ALBS (surface albedo), RSS (net shortwave radiation), RLS (net longwave radiation), and RNS (net radiation). Quantities in panel (b) are HFLS (latent heat flux), HFSS (sensible heat flux), HFG (ground heat flux), MRRO (runoff), and SNW (snow water equivalent). Quantities in panel (c) are GPP (gross primary productivity),

NEE (net ecosystem exchange), NBP (net biome productivity), AGB (aboveground biomass), CSOIL
 (soil carbon mass), BURNT (area burned), and LAI (leaf area index).

775

Figure 11 shows the zonal distribution of vegetation C mass, LAI, area burnt, GPP, and fire 776 777 CO₂ emissions (which constitute standard output from AMBER) and illustrates how AMBER compares the spread across the simulations indicated by 50%, 80%, and 100% shading against 778 779 observation-based estimates. The black and shades of grey indicate the model mean and the spread across the eight model simulations, respectively, and the thick lines in other colours show 780 the mean values of observation-based estimates. The time period over which observations and 781 782 model quantities are averaged is chosen to be the same. In Figure 11a, for aboveground biomass, 783 the GEOCARBON data set uses one product for the extratropics and another for the tropics to create a global aboveground biomass product. The Zhang product (Zhang and Liang, 2020) is 784 785 based on the fusion of multiple gridded biomass datasets for generating a global product. Both products are described in detail in Seiler et al. (2022). The model results generally compare better 786 787 with the Zhang product outside the 10°N to 10°S region but with the GEOCARBON product within this region. The values to the south of 40°S are generally less reliable because of the little 788 789 vegetated land area below this latitude. In Figure 11b, the model simulated values for soil organic 790 C compare better with the HWSD dataset compared to the SG250m data for reasons mentioned 791 in the previous paragraph. Simulated leaf area index (Figure 11c) and gross primary productivity (Figure 11e) generally compare well their observation-based estimates. The simulated area 792 burned (Figure 11d) and fire emissions (Figure 11f) also compare well with observation-based 793 794 estimates except that the model is not able to capture the small area burned and emissions at 795 northern high latitudes between around 50°N to 70°N. Figures A12 and A13 compare zonally

averaged values of other simulated quantities with observation-based estimates used in the
AMBER framework. Together Figures 11, A12, and A13 illustrate that the model is overall able
to capture the latitudinal distribution of most land surface quantities.

799



Figure 11: Zonally-averaged values of aboveground biomass (a), soil carbon mass (b), leaf area index (c), fractional area burnt (d), gross primary productivity (e), and fire CO₂ emissions (f) from the eight simulations summarized in Table 1. The model results are shown as their mean (black) and the spread across the eight simulations indicated by 50%, 80%, and 100% ranges in different shades of grey. The observation-based estimates used in AMBER to calculate scores are shown in coloured lines.





Figure 12: Summary of difference in overall scores for model simulated quantities and combinations for which the differences are statistically significant. The scores in parentheses for each quantity are the average scores across the eight simulations and provide context. The error bars denote the 95% confidence interval as explained in the text.

812

Since overall scores are available for all eight simulations for model quantities that are 813 814 compared to observations it is possible to evaluate how an interactive N cycle, and the choice of meteorological data and land cover data affect model performance. Figure 12 summarizes the 815 difference in overall scores for model quantities and combinations for which the differences are 816 statistically significant at the 5% level based on Tukey's test (Tukey, 1977). The score indicated in 817 parentheses for each quantity is the average score across the eight simulations and provides 818 context. For example, when evaluating the effect of change in land cover for NEE the use of the 819 820 GLC 2000 based land cover, compared to the use of the ESA CCI based land cover, degrades the average score for net ecosystem exchange by about 0.02 given that the average score for net 821

ecosystem exchange in 0.53. The error bars on the value 0.02 denote the 95% confidence interval 822 823 and in this case are calculated by differencing four simulations that use the GLC 2000 based land cover versus four simulations that use the ESA CCI based land cover. The use of the GLC 2000 824 based land cover on the other hand slightly improves scores for ecosystem respiration and liquid 825 826 soil moisture. The use of GSWP3 data improves model scores for net shortwave, longwave, and total radiation, for sensible and ground heat flux but degrades the overall score for area burned, 827 828 soil moisture, and more so for snow water equivalent. Finally, an interactive N cycle slightly 829 improves model performance for area burned and fire CO₂ emissions (due to improved 830 aboveground biomass in the tropics) but degrades it for ecosystem respiration, GPP, and net 831 ecosystem exchange. The inclusion of an interactive N cycle changes V_{c,max} to a prognostic 832 variable for each PFT as opposed to being specified based on observations. This is analogous to 833 running an atmospheric model with a fully dynamic 3-dimensional ocean as opposed to using 834 specified sea surface temperatures (SST) and sea ice concentrations (SIC). Using a dynamic ocean allows future projections (since future SSTs and SICs are not known) but invariably degrades a 835 836 model's performance for the present day since simulated SSTs and SICs will have their biases. Similarly, using an interactive N cycle allows to project future changes in V_{c,max} (based on changes 837 in N availability) but also degrades CLASSIC's performance for the present day since simulated 838 V_{c,max} has its own biases. Overall, the model performance is most affected by the choice of the 839 840 driving meteorological data for water and energy fluxes, and by the inclusion or absence of an N cycle and by the choice of land cover for carbon-cycle related state variables and fluxes. 841

842

The response of the terrestrial biosphere over the historical period is driven primarily by 845 846 four global change drivers – increasing atmospheric CO₂, changing climate, LUC, and N deposition 847 and fertilizer application. Our framework allows us to evaluate how a land surface model responds to increasing atmospheric CO₂, changing climate, and anthropogenic N additions to the 848 coupled soil-vegetation system and how this response is dependent on two driving 849 meteorological data sets, two land cover representations, and the two model variations (with 850 851 and without an interactive N cycle). However, the framework used here does not quantify the 852 uncertainty associated with LUC over the historical period since we use only one reconstruction 853 of increasing crop area over the historical period. These results help draw three primary 854 conclusions. First, even if the observations and models were perfect (including their structure 855 and their parameterizations) the uncertainty associated with driving meteorological data and 856 geophysical fields makes it difficult to evaluate LSMs. The uncertainty in global scale driving data 857 implies that a model can never be truly evaluated to its fullest extent. Model results can only be as good as the data that are used to force them and therefore even a perfect model cannot yield 858 perfect results. 859

Second, model tuning when driving the model with a single set of forcings and evaluating it against a single set of observations is likely not a fruitful exercise. Models should not be tuned to a single set of driving data and observation-based evaluation data. Rather their performance must be evaluated against a range of available observations in light of the uncertainty associated with driving data and the uncertainty associated with observations. A model's ability to reproduce a given single set of observations when driven with a single set of driving data is not a

true measure of its success. Here again, a perfect model driven by perfect forcing data cannot be
truly evaluated to its fullest extent since observations themselves have uncertainties.

868 Third, with the caveat that our framework uses only one reconstruction of increase in 869 crop area over the historical period, the response of a model expressed in terms of net 870 atmosphere-land CO₂ flux to perturbation in meteorological, CO₂, and LUC forcing over the historical period appears to be largely independent of its pre-industrial state as simulated here. 871 The pre-industrial soil and vegetation C mass for the eight simulations considered here vary 872 873 between 1035 ± 195 Pg C and 405 ± 58 Pg C (mean ± standard deviation), respectively. Both pre-874 industrial and present-day vegetation and soil C pools explain only about 2% to 7% of the variability in simulated net atmosphere-land CO₂ flux (Figure A11) over the 1997-2016 period of 875 876 each of the eight simulations. The net atmosphere-CO₂ flux from all eight simulations for the period the 1960s to 2000s is found to lie within the uncertainty range provided by the GCP 877 878 (Friedlingstein et al., 2022). Given the current uncertainty in net atmosphere-land CO₂ flux, it is therefore not possible to exclude any of the eight simulations at least on this basis. The finding 879 that a transient response of a model is independent of its pre-industrial state is also consistent 880 with land components of CMIP6 models. Arora et al. (2020) analyzed results from CMIP6 881 simulations in which atmospheric CO₂ increases at a rate of 1% per year from the year 1850 until 882 CO₂ quadruples from ~285 to ~1140 ppm. They found that the C-concentration and C-climate 883 884 feedback parameters for the land component of CMIP6 models do not depend on the absolute values of their vegetation and soil C pools but rather how a given model responds to changes in 885 atmospheric CO₂ and the associated change in temperature. This conclusion is perhaps 886 887 somewhat comforting in that while pre-industrial states of LSMs may be different from their true

observed states they still have the ability to reproduce net atmosphere-land CO₂ flux over the historical period that is consistent with current observation-based estimates. Clearly, this reasoning does not apply if pre-industrial vegetation or soil C mass are zero. One reason why present day net atmosphere-land CO₂ flux is independent of a LSM's pre-industrial state is because the model is first spun up to equilibrium conditions and then forced with time-variant forcings. However, successful reproduction of atmosphere-land CO₂ fluxes over the historical period is no guarantee that future projections from LSMs are reliable.

895 The ensemble-based approach used here also allows for the evaluation of the effect of a 896 given meteorological forcing and land cover, and the effect of an interactive N cycle on model 897 simulated quantities in a robust manner. Ensemble averages of simulations that use the CRU-JRA and GSWP3 meteorological forcing show that the use of the GSWP3 meteorological forcing yields 898 899 lower evapotranspiration (latent heat flux), higher runoff, higher sensible heat flux, a higher 900 burned area, and a weaker land C sink for the present day compared to when the CRU-JRA meteorological forcing is used. Possible reasons that explain these differences when using the 901 902 GSWP3 meteorological data are the higher frequency of high precipitation events (greater than ~5-10 mm/day) (Figure A2) and 0.93 °C higher temperature in the northern tropical region (Figure 903 A1h) in the GSWP3 compared to the CRU-JRA meteorological data. High precipitation intensity in 904 905 regions of high annual precipitation (e.g. the tropical regions) would lead to more surface runoff 906 since less precipitation infiltrates the top soil layer, further leading to less soil moisture, less 907 evapotranspiration, higher sensible heat flux, and more area burned. Higher temperatures in the northern tropical region in the GSWP3 meteorological data certainly contribute to all these 908 909 differences (except higher runoff). While, annual globally-averaged soil moisture is about 4%

higher in the simulations driven with the GSWP meteorological data (Figure 2c), in several parts 910 911 of the tropical regions annual simulated soil moisture is lower for GSWP3 simulations (not shown). The use of the ESA CCI land cover leads to higher soil C, higher GPP, and higher area 912 burned primarily because of the larger grass area when land cover is based on the ESA CCI 913 914 product compared to the GLC 2000 product. The use of the ESA CCI based land cover also leads to a slightly weaker land C sink for the present day. Finally, the comparison of simulations with 915 916 and without the N cycle averaged over all meteorological data and land cover combinations 917 allows us to identify the effect of the N cycle. Simulated vegetation C mass and GPP are lower in 918 the model version with the interactive N cycle. In particular, we found that the somewhat low 919 productivity at high latitudes, when the N cycle is turned on, leads to relatively large differences 920 in soil C at high latitudes regardless of the meteorological data or land cover being used to drive 921 the model. Although, this is not the reason for differences in net atmosphere-land CO_2 flux 922 between models with and without N cycling: as mentioned above present-day net atmosphereland CO₂ flux is independent of both the pre-industrial and present-day vegetation and soil C 923 924 pools. Given the knowledge about the effect of N cycling on model behaviour, the reasons can now be investigated to further improve the N cycle component of CLASSIC. 925

It is logical to assume that the results presented here are sensitive to the horizontal resolution of the model. Both forcing data that are used to drive the model, and observations against which model results are compared, are regridded to be consistent with the model's spatial resolution. For example, at the scale of a few meters, meteorological variables measured at a given site will indeed be less uncertain than their spatially-averaged values say for a 2.81° grid cell. Similarly, observations at a scale of a few meters for soil C and/or vegetation C mass will

also likely be more certain than their values at large spatial scales. This is one reason why AMBER
uses both gridded and in-situ observation-based estimates to calculate its scores. Fluxes of latent
and sensible heat, on the other hand, may not be any more certain at a given site than over large
spatial scales. This is because of the problems associated with energy budget closure (Mauder et
al., 2020) which, at the point scale, prevent the sum of annual latent and sensible heat flux to be
equal to net radiation (average of ground heat fluxes is close to zero at an annual time scale).

LSMs have become increasingly complex over the years and so has the requirement for 938 939 forcing data to drive these models. The evaluation of LSMs has also become complex as the 940 models now generate a multitude of variables that must be evaluated against their observationbased estimates. Estimates of observation-based data to evaluate models, and the availability of 941 942 forcing data, have also increased. Given the uncertainties associated with model inputs, model 943 structure, and observation-based data, it is unrealistic to expect LSMs to perfectly reproduce 944 observations for large-scale global simulations. It is not known a priori which model structure, forcing data sets, and observation data sets are better. Driving data including meteorological data 945 sets and land cover representations may be more realistic in some parts of world and less in 946 947 others. Observation-based data sets also have their limitations and attributes which may make them better or ill-suited for comparison with a given model. A more robust model evaluation 948 949 must therefore take into account the uncertainties both in the forcing and observation-based 950 data. A comprehensive and robust model evaluation can be performed by comparing multiple model realizations against multiple observation-based data sets. 951

952

953 References

954 Agustí-Panareda, A., Diamantakis, M., Massart, S., Chevallier, F., Muñoz-Sabater, J., Barré, J., Curcoll, R.,

955 Engelen, R., Langerock, B., Law, R. M., Loh, Z., Morguí, J. A., Parrington, M., Peuch, V.-H., Ramonet, M.,

956 Roehl, C., Vermeulen, A. T., Warneke, T., and Wunch, D.: Modelling CO₂ weather – why horizontal

957 resolution matters, Atmos Chem Phys, 19, 7347–7376, 2019.

Arora, V. K. and Boer, G. J.: A parameterization of leaf phenology for the terrestrial ecosystem
component of climate models, Glob. Change Biol., 11, 39–59, https://doi.org/10.1111/j.13652486.2004.00890.x, 2005.

- Arora, V. K. and Boer, G. J.: Simulating Competition and Coexistence between Plant Functional Types in a
 Dynamic Vegetation Model, Earth Interact., 10, 1–30, 2006.
- 963 Arora, V. K. and Melton, J. R.: Reduction in global area burned and wildfire emissions since 1930s
- 964 enhances carbon uptake by land, Nat. Commun., 9, 1326, https://doi.org/10.1038/s41467-018-03838-0,
 965 2018.
- 966 Arora, V. K., Boer, G. J., Christian, J. R., Curry, C. L., Denman, K. L., Zahariev, K., Flato, G. M., Scinocca, J.

967 F., Merryfield, W. J., and Lee, W. G.: The Effect of Terrestrial Photosynthesis Down Regulation on the

968 Twentieth-Century Carbon Budget Simulated with the CCCma Earth System Model, J. Clim., 22, 6066–

- 969 6088, https://doi.org/10.1175/2009JCLI3037.1, 2009.
- Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W.
- 971 G., and Merryfield, W. J.: Carbon emission limits required to satisfy future representative concentration
- pathways of greenhouse gases, Geophys. Res. Lett., 38, https://doi.org/10.1029/2010GL046270, 2011.
- 973 Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp,
- L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina,

975 T., Joetzjer, E., Kawamiya, M., Koven, C. D., Krasting, J. P., Law, R. M., Lawrence, D. M., Lenton, A.,

976 Lindsay, K., Pongratz, J., Raddatz, T., Séférian, R., Tachiiri, K., Tjiputra, J. F., Wiltshire, A., Wu, T., and

- 277 Ziehn, T.: Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison
- 978 to CMIP5 models, Biogeosciences, 17, 4173–4222, https://doi.org/10.5194/bg-17-4173-2020, 2020.
- Asaadi, A. and Arora, V. K.: Implementation of nitrogen cycle in the CLASSIC land model, Biogeosciences,
 18, 669–706, https://doi.org/10.5194/bg-18-669-2021, 2021.
- Avitabile, V., Herold, M., Heuvelink, G. B. M., and others: An integrated pan tropical biomass map usingmultiple reference datasets, Glob Chang Biol, 2016.
- Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction,
 Hydrol. Process., 6, 279–298, https://doi.org/10.1002/hyp.3360060305, 1992.
- Bonan, G. B. and Doney, S. C.: Climate, ecosystems, and planetary futures: The challenge to predict life
 in Earth system models, Science, 359, eaam8328, https://doi.org/10.1126/science.aam8328, 2018.
- Bonan, G. B., Lombardozzi, D. L., Wieder, W. R., Oleson, K. W., Lawrence, D. M., Hoffman, F. M., and
 Collier, N.: Model Structure and Climate Data Uncertainty in Historical Simulations of the Terrestrial

- 989 Carbon Cycle (1850–2014), Glob. Biogeochem. Cycles, 33, 1310–1326,
- 990 https://doi.org/10.1029/2019GB006175, 2019.

Booth, B. B. B., Jones, C. D., Collins, M., Totterdell, I. J., Cox, P. M., Sitch, S., Huntingford, C., Betts, R. A.,

- Harris, G. R., and Lloyd, J.: High sensitivity of future global warming to land carbon cycle processes,
- 993 Environ. Res. Lett., 7, 024002, https://doi.org/10.1088/1748-9326/7/2/024002, 2012.
- 994 Chuvieco, E., Lizundia-Loiola, J., Pettinari, M. L., Ramo, R., Padilla, M., Tansey, K., Mouillot, F., Laurent,
 995 P., Storm, T., Heil, A., and Others: Generation and analysis of a new global burned area product based on
 996 MODIS 250 m reflectance bands and thermal anomalies, Earth Syst. Sci. Data, 10, 2015–2031, 2018.
- Claverie, M., Matthews, J. L., Vermote, E. F., and Justice, C. O.: A 30+ Year AVHRR LAI and FAPAR Climate
 Data Record: Algorithm Description and Validation, Remote Sens., 8, 263, 2016.
- 999 Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and
- 1000 Randerson, J. T.: The International Land Model Benchmarking (ILAMB) System: Design, Theory, and
- 1001 Implementation, J. Adv. Model. Earth Syst., 10, 2731–2754, https://doi.org/10.1029/2018MS001354,
- 1002 2018.
- 1003 Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R.
- 1004 S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman,
- 1005 P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T.
- 1006 F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The Twentieth Century Reanalysis Project,
- 1007 Q. J. R. Meteorol. Soc., 137, 1–28, https://doi.org/10.1002/qj.776, 2011.
- Dai, A. and Trenberth, K. E.: Estimates of Freshwater Discharge from Continents: Latitudinal andSeasonal Variations, J Hydrometeorol, 3, 660–687, 2002.
- 1010 Di Vittorio, A. V., Chini, L. P., Bond-Lamberty, B., Mao, J., Shi, X., Truesdale, J., Craig, A., Calvin, K., Jones,
- 1011 A., Collins, W. D., Edmonds, J., Hurtt, G. C., Thornton, P., and Thomson, A.: From land use to land cover:
- 1012restoring the afforestation signal in a coupled integrated assessment–earth system model and the1013implications for CMIP5 RCP simulations, Biogeosciences, 11, 6435–6450, https://doi.org/10.5194/bg-11-10141015
- 1014 6435-2014, 2014.
- 1015 Di Vittorio, A. V., Mao, J., Shi, X., Chini, L., Hurtt, G., and Collins, W. D.: Quantifying the Effects of
- Historical Land Cover Conversion Uncertainty on Global Carbon and Climate Estimates, Geophys. Res.
 Lett., 45, 974–982, https://doi.org/10.1002/2017GL075124, 2018.
- 1018ESA: Land Cover CCI Product User Guide Version 2 Technical Report, European Space Agency. Available1019at http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf, 2017.
- 1020 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of
- 1021 the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization,
- 1022 Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.
- Fischer, G., Nachtergaele, F., Prieler, S., van Velthuizen, H. T., Verelst, L., and Wiberg, D.: Global Agroecological Zones Assessment for Agriculture (GAEZ 2008), IIASA and FAO, Laxenburg, Austria and Rome,
 Italy, 2008.

- Fisher, R. A. and Koven, C. D.: Perspectives on the Future of Land Surface Models and the Challenges of
 Representing Complex Terrestrial Systems, J Adv Model Earth Syst, 12, 2020.
- 1028 Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Hauck, J., Peters, G. P., Peters, W.,
- 1029 Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D. C. E., Canadell, J. G., Ciais, P., Jackson, R. B., Anthoni, P.,
- 1030 Barbero, L., Bastos, A., Bastrikov, V., Becker, M., Bopp, L., Buitenhuis, E., Chandra, N., Chevallier, F.,
- 1031 Chini, L. P., Currie, K. I., Feely, R. A., Gehlen, M., Gilfillan, D., Gkritzalis, T., Goll, D. S., Gruber, N.,
- 1032 Gutekunst, S., Harris, I., Haverd, V., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Joetzjer, E., Kaplan, J.
- 1033 O., Kato, E., Klein Goldewijk, K., Korsbakken, J. I., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lenton, A.,
- 1034 Lienert, S., Lombardozzi, D., Marland, G., McGuire, P. C., Melton, J. R., Metzl, N., Munro, D. R., Nabel, J.
- 1035 E. M. S., Nakaoka, S.-I., Neill, C., Omar, A. M., Ono, T., Peregon, A., Pierrot, D., Poulter, B., Rehder, G., 1036 Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P. P., Tian, H.,
- 1036 Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P. P., Tian, H.,
 1037 Tilbrook, B., Tubiello, F. N., van der Werf, G. R., Wiltshire, A. J., and Zaehle, S.: Global Carbon Budget
- 1038 2019, Earth Syst. Sci. Data, 11, 1783–1838, https://doi.org/10.5194/essd-11-1783-2019, 2019.
- 1039 Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., Le Quéré, C.,
- 1040 Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni,
- 1041 P., Bates, N. R., Becker, M., Bellouin, N., Bopp, L., Chau, T. T. T., Chevallier, F., Chini, L. P., Cronin, M.,
- 1042 Currie, K. I., Decharme, B., Djeutchouang, L. M., Dou, X., Evans, W., Feely, R. A., Feng, L., Gasser, T.,
- 1043 Gilfillan, D., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Houghton, R. A., Hurtt,
- 1044 G. C., lida, Y., Ilyina, T., Luijkx, I. T., Jain, A., Jones, S. D., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer,
- 1045 J., Korsbakken, J. I., Körtzinger, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lienert, S., Liu, J.,
- 1046 Marland, G., McGuire, P. C., Melton, J. R., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Niwa, Y., Ono,
- 1047 T., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M.,
- 1048 Schwinger, J., Schwingshackl, C., Séférian, R., Sutton, A. J., Sweeney, C., Tanhua, T., Tans, P. P., Tian, H.,
- 1049 Tilbrook, B., Tubiello, F., van der Werf, G. R., Vuichard, N., Wada, C., Wanninkhof, R., Watson, A. J.,
- 1050 Willis, D., Wiltshire, A. J., Yuan, W., Yue, C., Yue, X., Zaehle, S., and Zeng, J.: Global Carbon Budget 2021,
- 1051 Earth Syst. Sci. Data, 14, 1917–2005, https://doi.org/10.5194/essd-14-1917-2022, 2022.
- Garrigues, S., Lacaze, R., Baret, F., Morisette, J. T., Weiss, M., Nickeson, J. E., Fernandes, R., Plummer, S.,
 Shabanov, N. V., Myneni, R. B., and Others: Validation and intercomparison of global Leaf Area Index
 products derived from remote sensing data, J. Geophys. Res. Biogeosciences, 113, 2008.
- Giglio, L., Randerson, J. T., van der Werf, G. R., Kasibhatla, P. S., Collatz, G. J., Morton, D. C., and DeFries,
 R. S.: Assessing variability and long-term trends in burned area by merging multiple satellite fire
 preducts 7, 1171, 1186, 2010
- 1057 products, 7, 1171–1186, 2010.
- Giglio, L., Randerson, J. T., and van der Werf, G. R.: Analysis of daily, monthly, and annual burned area
 using the fourth-generation global fire emissions database (GFED4), J. Geophys. Res. Biogeosciences,
 118, 317–328, https://doi.org/10.1002/jgrg.20042, 2013.
- Harris, I. C.: CRU JRA v2.1: A forcings dataset of gridded land surface blend of Climatic Research Unit
- 1062 (CRU) and Japanese reanalysis (JRA) data; Jan. 1901 Dec. 2019, Centre for Environmental Data Analysis,
- 1063 University of East Anglia Climatic Research Unit,
- 1064 https://catalogue.ceda.ac.uk/uuid/10d2c73e5a7d46f4ada08b0a26302ef7, 2020.
- 1065 Hegglin, M., Kinnison, D., and Lamarque, J.-F.: Wet and dry NHx and NOy deposition data,
- 1066 input4MIPs.CMIP6.CMIP.NCAR. Version 2016-11-15, Earth System Grid Federation,
- 1067 https://doi.org/10.22033/ESGF/input4MIPs.10448, 2016.

- 1068 Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A.,
- 1069 Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan,
- 1070 R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m:
- 1071 Global gridded soil information based on machine learning, PLOS ONE, 12, 1–40,
- 1072 https://doi.org/10.1371/journal.pone.0169748, 2017.
- Hobeichi, S., Abramowitz, G., and Evans, J.: Conserving Land-Atmosphere Synthesis Suite (CLASS), J Clim,2019.
- Hornberger, G. M. and Spear, R. C.: Approach to the preliminary analysis of environmental systems, JEnv. Manage U. S., 12:1, 1981.
- 1077 van den Hurk, B., Kim, H., Krinner, G., Seneviratne, S. I., Derksen, C., Oki, T., Douville, H., Colin, J.,
- 1078 Ducharne, A., Cheruy, F., Viovy, N., Puma, M. J., Wada, Y., Li, W., Jia, B., Alessandri, A., Lawrence, D. M.,
- 1079 Weedon, G. P., Ellis, R., Hagemann, S., Mao, J., Flanner, M. G., Zampieri, M., Materia, S., Law, R. M., and
- 1080 Sheffield, J.: LS3MIP (v1.0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model
- 1081 Intercomparison Project aims, setup and expected outcome, Geosci. Model Dev., 9, 2809–2832,
- 1082 https://doi.org/10.5194/gmd-9-2809-2016, 2016.
- Hurtt, G. C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B. L., Calvin, K., Doelman, J. C., Fisk, J., Fujimori,
 S., Klein Goldewijk, K., Hasegawa, T., Havlik, P., Heinimann, A., Humpenöder, F., Jungclaus, J., Kaplan, J.
 O., Kennedy, J., Krisztin, T., Lawrence, D., Lawrence, P., Ma, L., Mertz, O., Pongratz, J., Popp, A., Poulter,
 B., Riahi, K., Shevliakova, E., Stehfest, E., Thornton, P., Tubiello, F. N., van Vuuren, D. P., and Zhang, X.:
 Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6,
 Geosci Model Dev, 13, 5425–5464, 2020a.
- Hurtt, G. C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B. L., Calvin, K., Doelman, J. C., Fisk, J., Fujimori,
 S., Klein Goldewijk, K., Hasegawa, T., Havlik, P., Heinimann, A., Humpenöder, F., Jungclaus, J., Kaplan, J.
 O., Kennedy, J., Krisztin, T., Lawrence, D., Lawrence, P., Ma, L., Mertz, O., Pongratz, J., Popp, A., Poulter,
 B., Riahi, K., Shevliakova, E., Stehfest, E., Thornton, P., Tubiello, F. N., van Vuuren, D. P., and Zhang, X.:
 Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6,
 Geosci. Model Dev., 13, 5425–5464, https://doi.org/10.5194/gmd-13-5425-2020, 2020b.
- Jacobson, A. R., Schuldt, K. N., Miller, J. B., and Oda, T.: CarbonTracker Documentation CT2019 release,2020.
- Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana,
 G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, Sci Data, 6, 74,
 2019.
- Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S.,
 Bodesheim, P., Carvalhais, N., and others: Scaling carbon fluxes from eddy covariance sites to globe:
 synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 17, 1343–1365, 2020.
- 1103 Kato, S., Loeb, N. G., Rose, F. G., Doelling, D. R., Rutan, D. A., Caldwell, T. E., Yu, L., and Weller, R. A.:
 1104 Surface Irradiances Consistent with CERES-Derived Top-of-Atmosphere Shortwave and Longwave
 1105 Irradiances, J Clim, 26, 2719–2740, 2013.

- 1106 Kou-Giesbrecht, S. and Arora, V. K.: Representing the Dynamic Response of Vegetation to Nitrogen
- 1107 Limitation via Biological Nitrogen Fixation in the CLASSIC Land Model, Glob. Biogeochem. Cycles, 36, 1108 e2022GB007341, https://doi.org/10.1029/2022GB007341, 2022.
- 1109 Kyker-Snowman, E., Lombardozzi, D. L., Bonan, G. B., Cheng, S. J., Dukes, J. S., Frey, S. D., Jacobs, E. M.,
- 1110 McNellis, R., Rady, J. M., Smith, N. G., Thomas, R. Q., Wieder, W. R., and Grandy, A. S.: Increasing the
- 1111 spatial and temporal impact of ecological research: A roadmap for integrating a novel terrestrial process
- into an Earth system model, Glob. Change Biol., 28, 665–684, https://doi.org/10.1111/gcb.15894, 2022.
- 1113 Lawrence, P. J. and Chase, T. N.: Representing a new MODIS consistent land surface in the Community
- 1114 Land Model (CLM 3.0), J. Geophys. Res. Biogeosciences, 112, https://doi.org/10.1029/2006JG000168,
- 1115 2007.
- Li, J., Duan, Q., Wang, Y.-P., Gong, W., Gan, Y., and Wang, C.: Parameter optimization for carbon and
 water fluxes in two global land surface models based on surrogate modelling, Int. J. Climatol., 38,
 e1016–e1031, https://doi.org/10.1002/joc.5428, 2018a.
- Li, W., MacBean, N., Ciais, P., Defourny, P., Lamarche, C., Bontemps, S., Houghton, R. A., and Peng, S.:
- 1120 Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI

1121 land cover maps (1992–2015), Earth Syst. Sci. Data, 10, 219–234, https://doi.org/10.5194/essd-10-219-

- 1122 2018, 2018b.
- Li, X. and Xiao, J.: Mapping Photosynthesis Solely from Solar-Induced Chlorophyll Fluorescence: A Global,
 Fine-Resolution Dataset of Gross Primary Production Derived from OCO-2, Remote Sens., 11, 2563,
 2019.
- Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., Van Dijk, A., McCabe, M. F.,
 Evans, J., and Others: Developing an improved soil moisture dataset by blending passive and active
 microwave satellite-based retrievals, 2011.
- Lu, C. and Tian, H.: Global nitrogen and phosphorus fertilizer use for agriculture production in the past
 half century: shifted hot spots and nutrient imbalance, Earth Syst. Sci. Data, 9, 181–192,
 https://doi.org/10.5194/essd-9-181-2017, 2017.
- Mauder, M., Foken, T., and Cuxart, J.: Surface-Energy-Balance Closure over Land: A Review, Bound.Layer Meteorol., 177, 395–426, https://doi.org/10.1007/s10546-020-00529-6, 2020.
- Meiyappan, P. and Jain, A. K.: Three distinct global estimates of historical land-cover change and landuse conversions for over 200 years, Front. Earth Sci., 6, 122–139, https://doi.org/10.1007/s11707-0120314-2, 2012.
- Melton, J. R. and Arora, V. K.: Competition between plant functional types in the Canadian Terrestrial
 Ecosystem Model (CTEM) v. 2.0, Geosci. Model Dev., 9, 323–361, 2016a.

Melton, J. R. and Arora, V. K.: Competition between plant functional types in the Canadian Terrestrial
Ecosystem Model (CTEM) v. 2.0, Geosci Model Dev, 9, 323–361, https://doi.org/10.5194/gmd-9-3232016, 2016b.

- 1142 Melton, J. R., Arora, V. K., Wisernig-Cojoc, E., Seiler, C., Fortier, M., Chan, E., and Teckentrup, L.: CLASSIC
- 1143 v1.0: the open-source community successor to the Canadian Land Surface Scheme (CLASS) and the
- 1144 Canadian Terrestrial Ecosystem Model (CTEM) Part 1: Model framework and site-level performance,
- 1145 Geosci. Model Dev. Discuss., 2019, 1–40, https://doi.org/10.5194/gmd-2019-329, 2019.
- Mortimer, C., Mudryk, L., Derksen, C., Luojus, K., Brown, R., Kelly, R., and Tedesco, M.: Evaluation of
 long-term Northern Hemisphere snow water equivalent products, The Cryosphere, 14, 1579–1594,
 2020.
- 1149 Mudryk, L.: Historical gridded snow water equivalent and snow cover fraction over Canada from remote 1150 sensing and land surface models, available at http://climate-scenarios.canada.ca/?page=blended-snow-1151 data (last accessed Oct 2022), 2020.
- 1152 Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y.,
- 1153 Smith, G. R., Lotsch, A., Friedl, M., Morisette, J. T., Votava, P., Nemani, R. R., and Running, S. W.: Global
- 1154 products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sens
- 1155 Env., 83, 214–231, 2002.
- 1156 Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J.,
- 1157 Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C., Zhang, L., Amiro, B.,
- 1158 Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M.,
- 1159 Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C.,
- 1160 Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D.,
- 1161 Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns,
- 1162 S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti,
- A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha,
 H., Dai, X., Davis, K. J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R. C., Delpierre, N., Desai,
- 1164 A. R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E.,
- 1166 Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B.,
- 1167 Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno,
- M., Gharun, M., Gianelle, D., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for
 eddy covariance data, Sci Data, 7, 225, 2020.
- 1170 Peng, S., Ciais, P., Maignan, F., Li, W., Chang, J., Wang, T., and Yue, C.: Sensitivity of land use change
- emission estimates to historical land use and land cover mapping, Glob. Biogeochem. Cycles, 31, 626– 643, https://doi.org/10.1002/2015GB005360, 2017.
- Poulter, B., Hattermann, F., Hawkins, E., Zaehle, S., Sitch, S., Restrepo-Coupe, N., Heyder, U., and
 Cramer, W.: Robust dynamics of Amazon dieback to climate change with perturbed ecosystem model
 parameters, Glob. Change Biol., 16, 2476–2495, https://doi.org/10.1111/j.1365-2486.2009.02157.x,
 2010.
- 1177 Reusch, A. and Gibbs, H. K.: New IPCC Tier-1 Global Biomass Carbon Map For the Year 2000, Oak Ridge1178 National Laboratory, Oak Ridge, Tennessee, 2008.
- 1179 Rödenbeck, C., Zaehle, S., Keeling, R., and Heimann, M.: How does the terrestrial carbon exchange
- $1180 \qquad respond \ to \ inter-annual \ climatic \ variations? \ A \ quantification \ based \ on \ atmospheric \ CO_2 \ data,$
- 1181 Biogeosciences, 15, 2481–2498, 2018.

Santoro, M., Beaudoin, A., Beer, C., Cartus, O., Fransson, J. E. S., Hall, R. J., Pathe, C., Schmullius, C.,

- 1183 Schepaschenko, D., Shvidenko, A., Thurner, M., and Wegmüller, U.: Forest growing stock volume of the
- 1184 northern hemisphere: Spatially explicit estimates for 2010 derived from Envisat ASAR, Remote Sens
- 1185 Env., 168, 316–334, 2015.

1186 Schepaschenko, D., Chave, J., Phillips, O. L., Lewis, S. L., Davies, S. J., Réjou-Méchain, M., Sist, P., Scipal,

- 1187 K., Perger, C., Herault, B., Labrière, N., Hofhansl, F., Affum-Baffoe, K., Aleinikov, A., Alonso, A., Amani, C.,
- 1188 Araujo-Murakami, A., Armston, J., Arroyo, L., Ascarrunz, N., Azevedo, C., Baker, T., Ba\lazy, R., Bedeau,
- 1189 C., Berry, N., Bilous, A. M., Bilous, S. Y., Bissiengou, P., Blanc, L., Bobkova, K. S., Braslavskaya, T., Brienen,
- 1190 R., Burslem, D. F. R. P., Condit, R., Cuni-Sanchez, A., Danilina, D., Del Castillo Torres, D., Derroire, G.,
- 1191 Descroix, L., Sotta, E. D., d'Oliveira, M. V. N., Dresel, C., Erwin, T., Evdokimenko, M. D., Falck, J.,
- Feldpausch, T. R., Foli, E. G., Foster, R., Fritz, S., Garcia-Abril, A. D., Gornov, A., Gornova, M., GothardBassébé, E., Gourlet-Fleury, S., Guedes, M., Hamer, K. C., Susanty, F. H., Higuchi, N., Coronado, E. N. H.,
- Hubau, W., Hubbell, S., Ilstedt, U., Ivanov, V. V., Kanashiro, M., Karlsson, A., Karminov, V. N., Killeen, T.,
- 1195 Koffi, J.-C. K., Konovalova, M., Kraxner, F., Krejza, J., Krisnawati, H., Krivobokov, L. V., Kuznetsov, M. A.,
- 1196 Lakyda, I., Lakyda, P. I., Licona, J. C., Lucas, R. M., Lukina, N., Lussetti, D., Malhi, Y., Manzanera, J. A.,
- 1197 Marimon, B., Junior, B. H. M., Martinez, R. V., Martynenko, O. V., Matsala, M., Matyashuk, R. K., Mazzei,
- 1198 L., Memiaghe, H., Mendoza, C., Mendoza, A. M., Moroziuk, O. V., Mukhortova, L., Musa, S., Nazimova, D.
- 1199 I., Okuda, T., Oliveira, L. C., Ontikov, P. V., et al.: The Forest Observation System, building a global
- 1200 reference dataset for remote sensing of forest biomass, Sci Data, 6, 198, 2019.
- Seiler, C., Melton, J. R., Arora, V. K., and Wang, L.: CLASSIC v1.0: the open-source community successor
 to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM) –
- 1203 Part 2: Global benchmarking, Geosci. Model Dev., 14, 2371–2417, 2021a.
- Seiler, C., Melton, J. R., Arora, V. K., and Wang, L.: CLASSIC v1.0: the open-source community successor
 to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM) –
 Part 2: Global benchmarking, Geosci. Model Dev., 14, 2371–2417, https://doi.org/10.5194/gmd-142371-2021, 2021b.
- Seiler, C., Melton, J. R., Arora, V. K., Sitch, S., Friedlingstein, P., Anthoni, P., Goll, D., Jain, A. K., Joetzjer,
 E., Lienert, S., Lombardozzi, D., Luyssaert, S., Nabel, J. E. M. S., Tian, H., Vuichard, N., Walker, A. P., Yuan,
- 1210 W., and Zaehle, S.: Are Terrestrial Biosphere Models Fit for Simulating the Global Land Carbon Sink?, J.
- 1211 Adv. Model. Earth Syst., 14, e2021MS002946, https://doi.org/10.1029/2021MS002946, 2022.
- Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H.: A global soil data set for earth system modeling, J.
 Adv. Model. Earth Syst., 6, 249–263, https://doi.org/10.1002/2013MS000293, 2014.
- Slevin, D., Tett, S. F. B., Exbrayat, J.-F., Bloom, A. A., and Williams, M.: Global evaluation of gross primary
 productivity in the JULES land surface model v3.4.1, Geosci. Model Dev., 10, 2651–2670,
- 1216 https://doi.org/10.5194/gmd-10-2651-2017, 2017.
 - Stackhouse, P. W., Jr, Gupta, S. K., Cox, S. J., Zhang, T., Mikovitz, J. C., and Hinkelman, L. M.: The
 NASA/GEWEX surface radiation budget release 3.0: 24.5-year dataset, Gewex News, 21, 10–12, 2011.

¹²¹⁹ Strahler, A. H., Muller, J., Lucht, W., Schaaf, C., and others: MODIS BRDF/albedo product: algorithm 1220 theoretical basis document version 5.0, MODIS, 1999.

- 1221 Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V.,
- 1222 Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A.,
- 1223 Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model
- version 5 (CanESM5.0.3), Geosci. Model Dev., 12, 4823–4873, https://doi.org/10.5194/gmd-12-48232019, 2019.
- Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections,
 Philos. Trans. R. Soc. Math. Phys. Eng. Sci., 365, 2053–2075, https://doi.org/10.1098/rsta.2007.2076,
 2007.
- 1229 Tian, Y., Dickinson, R. E., Zhou, L., and Shaikh, M.: Impact of new land boundary conditions from
- 1230 Moderate Resolution Imaging Spectroradiometer (MODIS) data on the climatology of land surface
- 1231 variables, J. Geophys. Res. Atmospheres, 109, https://doi.org/10.1029/2003JD004499, 2004.
- Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., and
 Allison, S. D.: Causes of variation in soil carbon simulations from CMIP5 Earth system models and
 comparison with observations 10, 1717–1736, 2013
- 1234 comparison with observations, 10, 1717–1736, 2013.
- 1235 Tukey, J. W.: Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.
- 1236 Verger, A., Baret, F., and Weiss, M.: Near real-time vegetation monitoring at global scale, IEEE J. Sel. Top.1237 In, 2014.
- Verseghy, D. L.: Class—A Canadian land surface scheme for GCMS. I. Soil model, Int. J. Climatol., 11,
 111–133, https://doi.org/10.1002/joc.3370110202, 1991.
- 1240 Verseghy, D. L., McFarlane, N. A., and Lazare, M.: Class—A Canadian land surface scheme for GCMS, II.
- 1241 Vegetation model and coupled runs, Int. J. Climatol., 13, 347–370,
- 1242 https://doi.org/10.1002/joc.3370130402, 1993.
- Wang, A., Price, D. T., and Arora, V.: Estimating changes in global vegetation cover (1850–2100) for use
 in climate models, Glob. Biogeochem. Cycles, 20, https://doi.org/10.1029/2005GB002514, 2006.
- Wang, L., Bartlett, P., Chan, E., and Xiao, M.: Mapping of Plant Functional Type from Satellite-Derived
 Land Cover Datasets for Climate Models, in: IGARSS 2018 2018 IEEE International Geoscience and
- 1247 Remote Sensing Symposium, 3416–3419, https://doi.org/10.1109/IGARSS.2018.8518046, 2018.
- 1248 Wang, L., Bartlett, P., Pouliot, D., Chan, E., Lamarche, C., Wulder, M. A., Defourny, P., and Brady, M.:
- Comparison and Assessment of Regional and Global Land Cover Datasets for Use in CLASS over Canada,
 Remote Sens., 11, 2286, https://doi.org/10.3390/rs11192286, 2019.
- 1251 Wang, L., Arora, V. K., Bartlett, P., Chan, E., and Curasi, S. R.: Mapping of ESA-CCI land cover data to plant
- 1252 functional types for use in the CLASSIC land model, EGUsphere, 2022, 1–43,
- 1253 https://doi.org/10.5194/egusphere-2022-923, 2022.
- 1254 Wieder, W.: Regridded Harmonized World Soil Database v1.2, ,
- 1255 https://doi.org/10.3334/ORNLDAAC/1247, 2014.

- 1256 Wu, Z., Ahlström, A., Smith, B., Ardö, J., Eklundh, L., Fensholt, R., and Lehsten, V.: Climate data induced
- 1257 uncertainty in model-based estimations of terrestrial primary productivity, Environ. Res. Lett., 12,
- 1258 064013, https://doi.org/10.1088/1748-9326/aa6fd8, 2017.
- Xue, B.-L., Guo, Q., Hu, T., Wang, G., Wang, Y., Tao, S., Su, Y., Liu, J., and Zhao, X.: Evaluation of modeled
 global vegetation carbon dynamics: Analysis based on global carbon flux and above-ground biomass
 data Ecol Modell, 255, 84, 06, 2017
- 1261 data, Ecol Modell, 355, 84–96, 2017.
- 1262 Zhang, Y. and Liang, S.: Fusion of Multiple Gridded Biomass Datasets for Generating a Global Forest1263 Aboveground Biomass Map, Remote Sens., 12, 2559, 2020.
- Zhang, Y., Xiao, X., Wu, X., Zhou, S., Zhang, G., Qin, Y., and Dong, J.: A global moderate resolution dataset
 of gross primary production of vegetation for 2000–2016, Sci. Data, 4, 170165, 2017.
- 1266

1267 Code/data availability

- 1268
- 1269 More information about the CLASSIC land surface model and its Fortran code are available at 1270 https://cccma.gitlab.io/classic_pages/.
- 1271
- 1272 AMBER source code as well as the scripts required for reproducing the computational environment,
- 1273 including all dependencies on other R-packages, can be found at
- 1274 https://doi.org/10.5281/zenodo.5670387.
- 1275
- 1276 The full suite of results from AMBER for the eight simulations presented in this study can be
- 1277 found at <u>https://cseiler.shinyapps.io/ShinyCLASSIC/</u>.
- 1278

1279 Author contribution

- 1280
- 1281 VA and SKG performed the simulations, and VA wrote the majority of the manuscript. CS performed the
- 1282 AMBER related analysis. LW put together the ESA CCI land cover. CS, LW, and SKG provided comments
- 1283 on the entire manuscript and also wrote their respective sections.
- 1284
- 1285 Competing interests
- 1286
- 1287 There are no competing interests.
- 1288

1289 Acknowledgment

- 1290
- 1291 We thank Joe Melton for providing comments on an earlier version of this paper. We also thank
- 1292 Benjamin Bond-Lamberty for taking this paper on as an Associate Editor, and the two anonymous
- 1293 reviewers for providing helpful comments which greatly improved this paper.

Appendix 1294 1295 1296 A1: Automated Model Benchmarking R Package (AMBER) 1297 1298 The Automated Model Benchmarking R package quantifies model performance using five scores 1299 that assess a model's bias (S_{bias}), root-mean-square-error (S_{rmse}), seasonality (S_{phase}), inter-annual variability (S_{iav}) , and spatial distribution (S_{dist}) . All scores are dimensionless and range from zero 1300 to one, where increasing values imply better performance. The exact definition of each skill score 1301 1302 is provided below. A1.1 Bias Score (Sbias) 1303 1304 The bias is defined as the difference between the time-mean values of model and reference data: $bias(\lambda,\phi) = \overline{v_{mod}}(\lambda,\phi) - \overline{v_{ref}}(\lambda,\phi),$ 1305 (A1) where $\overline{v_{mod}}(\lambda, \phi)$ and $\overline{v_{ref}}(\lambda, \phi)$ are the mean values in time (t) of a variable v as a function of 1306 1307 longitude λ and latitude ϕ for model and reference data, respectively. Nondimensionalization is achieved by dividing the bias by the standard deviation of the reference data (σ_{ref}): 1308 $\varepsilon_{bias}(\lambda,\phi) = \frac{|bias(\lambda,\phi)|}{\sigma_{ref}(\lambda,\phi)}$ 1309 (A2) 1310 Note that ε_{bias} is always positive, as it uses the absolute value of the bias. For evaluations against stream flow measurements, the bias is divided by the annual mean rather than the standard 1311 deviation of the reference data. This is because we assess streamflow on an annual rather than 1312 monthly basis, implying that the corresponding standard deviation is small. The same approach 1313 is applied to soil C and vegetation C mass, whose reference data provide a static snapshot in time. 1314 For both of these cases, $\varepsilon_{bias}(\lambda, \phi)$ becomes: 1315

1316
$$\varepsilon_{bias}(\lambda,\phi) = \frac{|bias(\lambda,\phi)|}{\overline{v_{ref}}(\lambda,\phi)}$$
(A3)

1318 A bias score that ranges from zero to one is calculated next:

1319
$$s_{bias}(\lambda,\phi) = e^{-\varepsilon_{bias}(\lambda,\phi)}$$
(A4)

While small relative errors yield score values close to one, large relative errors cause score values
to approach zero. Taking the mean of *s_{bias}* across all latitudes and longitudes, denoted by a double
bar over a variable, leads to the scalar score:

1323
$$S_{bias} = \overline{s_{bias}(\lambda, \phi)}$$
(A5)

1324

1325 A1.2 Root-Mean-Square-Error Score (Srmse)

1326 While the bias assesses the difference between time-mean values, the root-mean-square-error 1327 (*rmse*) is concerned with the residuals of the modeled and observed time series:

1328
$$rmse(\lambda,\phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left(v_{mod}(t,\lambda,\phi) - v_{ref}(t,\lambda,\phi) \right)^2 dt}$$
(A6)

1329

where t_0 and t_f are the initial and final time steps, respectively. A similar metric is the centralized rmse (crmse), which is based on the residuals of the anomalies:

1333
$$crmse(\lambda,\phi) = \sqrt{\frac{1}{t_f - t_0}} \int_{t_0}^{t_f} \left[\left(v_{mod}(t,\lambda,\phi) - \overline{v_{mod}}(\lambda,\phi) \right) - \left(v_{ref}(t,\lambda,\phi) - \overline{v_{ref}}(\lambda,\phi) \right) \right]^2 dt$$
(A7)

The *crmse*, therefore, assesses residuals that have been bias-corrected. Since we already assessed the model's bias through S_{bias} , it is convenient to assess the residuals using *crmse* rather than *rmse*. In a similar fashion to the bias, we then compute a relative error:

1338
$$\varepsilon_{rmse}(\lambda,\phi) = \frac{crmse(\lambda,\phi)}{\sigma_{ref}(\lambda,\phi)}$$
(A8)

1339 scale this error onto a unit interval:

1340
$$s_{rmse}(\lambda,\phi) = e^{-\varepsilon_{rmse}(\lambda,\phi)}$$
 (A9)

1341 and compute the spatial mean:

1342
$$S_{rmse} = \overline{s_{rmse}(\lambda, \phi)}$$
(A10)

1343 A3 Phase Score (Sphase)

1344 The skill score S_{phase} assesses how well the model reproduces the seasonality of a variable by 1345 computing the time difference $\theta(\lambda, \phi)$ between modeled and observed month of maxima of the 1346 climatological mean cycle:

$$\theta(\lambda, \phi) = \max(c_{mod}(t, \lambda, \phi)) - \max(c_{ref}(t, \lambda, \phi))$$
(A11)

where c_{mod} and c_{ref} are the climatological mean cycle of the model and reference data, respectively. The operator *maxima* in equation A11 calculates the month in which the maximum of a given quantity occurs. The time difference $\theta(\lambda, \phi)$ in months is then scaled from zero to one based on the consideration that the maximum possible time difference is 6 months:

1352
$$s_{phase}(\lambda,\phi) = \frac{1}{2} \left[1 + \cos\left(\frac{2\pi \,\theta(\lambda,\phi)}{12}\right) \right]$$
(A12)

1353 The spatial mean of s_{phase} then leads to the scalar score:

$$S_{phase} = \overline{S_{phase}(\lambda, \phi)}$$
(A13)

1355

1356 A4 Inter-Annual Variability Score (Siav)

1357 The skill score *S*_{*iav*} quantifies how well the model reproduces patterns of inter-annual variability.

1358 This score is based on data where the seasonal cycle (c_{mod} and c_{ref}) has been removed:

1359
$$iav_{mod}(\lambda,\phi) = \sqrt{\frac{1}{t_f - t_0}} \int_{t_0}^{t_f} \left(v_{mod}(t,\lambda,\phi) - c_{mod}(t,\lambda,\phi) \right)^2 dt$$
(A14)

1360
$$iav_{ref}(\lambda,\phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left(v_{ref}(t,\lambda,\phi) - c_{ref}(t,\lambda,\phi) \right)^2 dt} .$$
(A15)

1361

1362 The relative error, nondimensionalization, and spatial mean are computed next:

1363
$$\varepsilon_{iav}(\lambda,\phi) = \left| iav_{mod}(\lambda,\phi) - iav_{ref}(\lambda,\phi) \right| / iav_{ref}(\lambda,\phi)$$
(A16)

1364
$$s_{iav}(\lambda,\phi) = e^{-\varepsilon_{iav}(\lambda,\phi)}$$
(A17)

1365
$$S_{iav} = \overline{S_{iav}(\lambda, \phi)}$$
(A13)

1366 A5 Spatial Distribution Score (Sdist)

The spatial distribution score S_{dist} assesses how well the model reproduces the spatial pattern of a variable. The score considers the correlation coefficient *R* and the relative standard deviation σ between $\overline{v_{mod}}(\lambda, \phi)$ and $\overline{v_{ref}}(\lambda, \phi)$. The score S_{dist} increases from zero to one, the closer *R* and σ approach a value of one. No spatial integration is required as this calculation yields a single value:

1372
$$S_{dist} = 2(1+R)\left(\sigma + \frac{1}{\sigma}\right)^{-2}$$
 (A19)

1373 where σ is the ratio between the standard deviation of the model and reference data:

1374
$$\sigma = \sigma_{\overline{v_{mod}}} / \sigma_{\overline{v_{ref}}}$$
(A20)

1375 and $\sigma_{\overline{v_{mod}}}$ and $\sigma_{\overline{v_{ref}}}$ are the standard deviations of the annual mean values from the model and

1376 reference/observation-based data, respectively, and therefore are scalars.

1377 A6 Overall Score (Soverall)

1378 As a final step, scores are averaged to obtain an overall score:

1379
$$S_{overall} = \frac{S_{bias} + 2 S_{rmse} + S_{phase} + S_{iav} + S_{dist}}{1 + 2 + 1 + 1 + 1}$$
(A21)

1380 Note that *S_{rmse}* is weighted by a factor of two and is an entirely subjective decision but follows

1381 Collier et al. (2018).



1384 Figure A1: Comparison of monthly precipitation (upper panel) and temperature (lower panel)

- 1385 for five global regions (global, north of 25 °N, northern and southern tropics, and south of 25 °S)
- 1386 from the CRU-JRA and GSWP3 meteorological forcing data sets that are used to drive the
- 1387 CLASSIC model. The global and regional averages exclude Greenland and Antarctica. The legend
- 1388 entries show the annual mean values averaged over the 1997-2016 period. The thin lines show
- 1389 individual years and the thick line is their average.
- 1390
- 1391





Figure A2: Comparison of the frequency distribution of daily precipitation between the CRU-JRA and GSWP3 meteorological data sets for three broad regions and the period 1997-2016: a) the Amazonian region, b) the Sahel region, and c) the Midwest United States. The frequency is represented as a percentage of time daily precipitation is between x and x+1 mm/day, where x is the value on the x-axis. Panel (d) shows the location of these broad regions. The underlying map in panel (d) is from Google Maps.

Bolivia



- 1407
- 1408
- 1409
- 1410
- _ . _ •
- 1411



1412

1413 Figure A3: Time series of simulated globally-averaged annual soil moisture (a) and soil

- 1414 temperature (b) in the top 1m, global annual evapotranspiration (c), and runoff (d) from the
- 1415 eight simulations summarized in Table 1. The thin lines show the individual years and the thick
- 1416 lines show their 11-year moving average. Model values averaged over the pre-industrial (1851-
- 1417 1860) and present-day (1997-2016) time periods, and their difference, are also shown.

1419





1421

Figure A4: Time series of simulated globally-averaged annual energy fluxes from the eight simulations summarized in Table 1. Panel (a) shows net downward shortwave radiation, panel (b) shows net downward longwave radiation, panel (c) shows latent heat flux, and panel (d) shows sensible heat flux. The thin lines show the individual years and the thick lines show their 11-year moving average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown for individual simulations.

1430

1431



Figure A5: Time series of globally-averaged annual net downward longwave and shortwave radiation (over all land area excluding Greenland and Antarctica) averaged over the four ensemble members each that are driven with and without N cycle (panels a, b), driven with GLC 2000 and ESA CCI based land cover (panels c, d), and driven with GSWP3 and CRU-JRA meteorological data (panels e, f). The thin lines show the individual years and the thick lines show their 11-year moving average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown.

- 1442
- 1443
- 1444



Figure A6: Time series of simulated global annual vegetation carbon mass (a) and soil carbon (b) from the eight simulations summarized in Table 1. The global totals exclude Greenland and Antarctica. Panels (c) and (d) show the zonally-averaged values of vegetation carbon mass and soil carbon mass over land from the eight simulations averaged over the 1997-2016 period. The thin lines show the individual years and the thick lines show their 11-year moving average in panels (a) and (b). Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown in panels (a) and (b).

- 1453
- 1454
- 1455
- 1456
1458



1459

Figure A7: Time series of simulated global annual gross primary productivity (GPP) (a), autotrophic respiration (b), and heterotrophic respiration (c) from the eight simulations summarized in Table 1. Panel (d) shows the zonally-averaged values of GPP from the eight simulations averaged over the 1997-2016 period for each simulation. The thin lines show the individual years and the thick lines show their 11-year moving average in panels (a) to (c). Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown in panels (a) to (c).

1467

1468





1476 moving average. Model values averaged over the pre-industrial (1851-1860) and present-day

1477 (1997-2016) time periods, and their difference, are also shown.

1479



1480

1481

Figure A9: Time series of simulated global annual area burned (a) and fire CO₂ emissions (b) from the eight simulations summarized in Table 1. Panels (c) and (d) show the zonally-averaged area burned and fire CO₂ emissions from the eight simulations averaged over the 1997-2016 period. The thin lines for the time series show the individual years and the thick lines show their 1486 11-year moving average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown for panels (a) and (b).

1489







1504	Figure A11: Time series of simulated global annual net atmosphere-land CO ₂ flux (a) and its
1505	zonally-averaged values from the eight simulations summarized in Table 1 averaged over the
1506	1997-2016 period. In panel (a) simulated annual net atmosphere-land CO ₂ flux values are
1507	compared to the estimates from the Global Carbon Project (Friedlingstein et al., 2022). The thin
1508	lines for the time series in panel (a) show the individual years and the thick lines show their 11-
1509	year moving average. In panel (b) the simulated zonally-averaged values are compared to the
1510	range from 11 models that contributed to the TRENDY 2020 intercomparison and averaged
1511	over the 1997-2016 period.



Figure A12: Zonally-averaged values of soil moisture (a), runoff (b), latent heat flux (c), and sensible heat flux (d) from the eight simulations summarized in Table 1. The model results are shown as their mean (black) and the spread across the eight simulations indicated by 50%, 80%, and 100% ranges in different shades of grey. The observation-based estimates used in AMBER to calculate scores are shown in coloured lines.







Figure A13: Zonally-averaged values of surface albedo (a), snow water equivalent (b), net surface radiation (c), net longwave radiation (d), and net shortwave radiation (e) from the eight simulations summarized in Table 1. The model results are shown as their mean (black) and the spread across the eight simulations indicated by 50%, 80%, and 100% ranges in different shades of grey. The observation-based estimates used in AMBER to calculate scores are shown in coloured lines.

1536