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

Vivek. K. Arora[1], Christian Seiler[2], Libo Wang[2], and Sian Kou-Giesbrecht[1]

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

1    [2]Climate Processes Section, Climate Research Division, Environment and Climate Change Canada, Toronto, ON,
2    Canada
3
4    *Correspondence to*: Vivek K. Arora (vivek.arora@ec.gc.ca)

5    **Abstract**

7    Quantification of uncertainty in fluxes of energy, water, and $CO_2$ 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

11   over the historical period is implemented in the same way), as well as two model structures (with

12   and without coupling of carbon and nitrogen cycles), the uncertainty in simulated results over

13   the historical period is quantified for the Canadian Land Surface Scheme Including

14   Biogeochemical Cycles (CLASSIC) model. The resulting eight (2 x 2 x 2) model simulations are

15   evaluated using an in-house model evaluation framework that uses multiple observations-based

16   data sets for a range of quantities. The simulated area burned, fire $CO_2$ emissions, soil carbon

17   mass, vegetation carbon mass, runoff, heterotrophic respiration, gross primary productivity, and

18   sensible heat flux show the largest spread across the eight simulations relative to their mean.

19   Simulated net atmosphere-land $CO_2$ flux, a critical determinant of the performance of LSMs, is

20   found to be largely independent of the simulated pre-industrial vegetation and soil carbon mass

21   although our framework represents the historical increase in crop area in the same way in both

22   land cover representations. This indicates that models can provide reliable estimates of the

23   strength of the land carbon sink despite some biases in carbon stocks. Results show that

24   evaluating an ensemble of model results against multiple observations disentangles model

25   deficiencies from uncertainties in model inputs, observation-based data, and model

26   configuration.

## 1. Introduction

The current generation land surface models (LSMs) explicitly simulate the fluxes of energy, water, momentum, and trace gases (including $CO_2$, $CH_4$, and $N_2O$) between the atmosphere and the land surface. These models have become an essential tool in understanding what role the land surface plays in the global climate system under current and projected future changes in environmental conditions, including atmospheric $CO_2$ concentration (Bonan and Doney, 2018). LSMs are also an essential component of climate and Earth system models (ESMs), together with their ocean and atmosphere components. Within the framework of ESMs, LSMs are coupled interactively to their atmospheric components through the fluxes of energy, momentum, and matter.

The complexity of LSMs has increased over time as more physical and biogeochemical 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 physical and biogeochemical processes implies that different models respond differently even when driven with the same external forcings. One estimate of the uncertainty in our understanding of land surface physical and biogeochemical processes is obtained by evaluating the inter-model spread in a given quantity when models are forced in the same manner. Other than the uncertainty among models due to differences in their model structures and parameterizations of various processes, uncertainty also exists due to at least three other

47    reasons. These include uncertainty 1) in parameter values[1] of represented processes, 2) in driving

48    meteorological data, and 3) in the specification of the geophysical fields. LSMs are typically driven

49    with meteorological data consisting of seven primary variables (incoming long and shortwave

50    radiation, temperature, precipitation, specific humidity, wind speed, and pressure).  In addition,

51    the geophysical fields of land cover, soil texture, and soil permeable depth are also required.

52    Driving data for LSMs also consist of atmospheric $CO_2$ concentration and other model-specific

53    external forcings such as nitrogen deposition and fertilizer application rates for models that

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

58    are driven with specified external forcings. The simulations are performed from the year 1700 to

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 $CO_2$ emissions and

61    the ocean carbon sink (Friedlingstein et al., 2019). The external forcings used to drive LSMs in the

62    TRENDY intercomparison include, 1) six hourly meteorological data from 1901 to the present day

63    (the most recent 2020 TRENDY intercomparison used the CRU-JRA forcing obtained by blending

64    the climate research unit (CRU) monthly data and the Japanese reanalysis (JRA)); 2) atmospheric

65    $CO_2$ concentration; and 3) information about changes in crop area and other land use changes

66    (LUC) from the land use harmonization (LUH) product (Hurtt et al., 2020a). The information about

---

[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.

71        Models participating in the TRENDY simulations are thus driven with common

72    meteorological and LUC forcings as part of its protocol. The resulting spread across models

73    participating in the TRENDY project thus provides a measure of inter-model uncertainty, as

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

78    various sub-projects (Eyring et al., 2016) is another prominent example. MIPs now routinely form

79    the basis of evaluating models against observations and multi-model means of various quantities.

80    Multi-model means are also considered the best estimate for a given quantity (Tebaldi and

81    Knutti, 2007).

82        The modelling community has been long aware of the uncertainty associated with

83    parameter values, since a large fraction of physical and biogeochemical model processes are

84    parameterized, and such uncertainty analysis dates back to the early hydrological models (e.g.

85    Hornberger and Spear, 1981; Beven and Binley, 1992). More recent examples of parameter value

86    uncertainty in the context of a given LSM include Poulter et al. (2010), Booth et al. (2012), and Li

87    et al. (2018a). The land surface modelling community, however, has only recently begun to

88    address and quantify uncertainty related with driving meteorological data. Wu et al. (2017), for

89  example, illustrate the uncertainty in gross primary productivity (GPP) simulated by the Lund-

90  Potsdam-Jena General Ecosystem Simulator (LPJ-GUESS) model when driven by six different

91  meteorological data sets. Bonan et al. (2019) analyze the uncertainty in simulated carbon cycle

92  related variables using three versions of the Community Land Model (CLM) when driven with two

93  meteorological data sets over the historical period. Slevin et al. (2017) assess the uncertainty in

94  simulated GPP by the JULES land model when driven by three different meteorological data sets.

95  Studies that evaluate the effect of different land cover representations on model performance

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

98  the MODIS satellite data as implemented in CLM2 in the Community Atmosphere Model (CAM2)

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

102  representations to quantify the uncertainty associated with both these forcings. Other than

103  these, we also use two versions of the CLASSIC model: one that represents the interactions

104  between the carbon (C) and nitrogen (N) cycles and the other in which these interactions are

105  turned off. CLASSIC has contributed to the simulations for the TRENDY intercomparison, and the

106  GCP, since 2016 (formerly under the CLASS-CTEM name). Seiler et al. (2021a) have evaluated how

107  well the CLASSIC model performs when forced with three different meteorological data sets using

108  the model version without the N cycle. Using the two meteorological forcing data sets, two

109  representations of land cover, and two versions of the model we perform eight simulations over

110  the historical period since 1700. All of these simulations may be considered equally likely

111 representations of the modelled state of the land surface over the historical period. Yet, they all

112 have their own distinct biases since simulated land surface states and fluxes are different. We

113 use these simulations to illustrate the uncertainty associated with meteorological forcing and the

114 two different representations of land cover that are used to drive the model. We also use an in-

115 house open-source benchmarking system (see code/data availability section) to evaluate these

116 different simulations against observations-based data sets: AMBER (Automated Benchmarking R

117 Package) (Seiler et al., 2021b) uses gridded and in-situ observation-based estimates of 19 energy,

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

122 performed for this study. Section 4 analyses the results from the simulations to illustrate their

123 different states and reports results from the AMBER benchmarking exercise. Finally, the

124 discussion and conclusions are presented in Section 5. The use of more than one meteorological

125 forcing data sets and land cover representation yields a conundrum since tuning model

126 parameters for a given forcing data set is not a useful exercise anymore. We also report a new

127 finding that despite different land C states (characterized in terms of vegetation and soil C mass)

128 in the eight simulations considered here, the net atmosphere-land $CO_2$ flux over the historical

129 period in these simulations is consistent with estimates from the GCP. This and the discussion

130 about the broader question of model tuning are also presented in Section 5.

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

**2.1 The physical and carbon biogeochemical processes**

132

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

136     as the land component in the family of Canadian Earth System Models (Arora et al., 2009, 2011;

137     Swart et al., 2019). Melton et al. (2019) provide an overview of the CLASSIC land model and

138     launched it as a community model. The basis of the modelling of physical and biogeochemical

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 $CO_2$, $CH_4$, $N_2O$, $NO_x$, and $NH_3$ based on its

142     biogeochemical process. The representation of the terrestrial N cycle is a new addition to CLASSIC

143     (Asaadi and Arora, 2021; Kou-Giesbrecht and Arora, 2022) and allows for the simulation of the

144     interactions between the C and N cycles explicitly.

145     The CLASSIC model simulations can be performed over a spatial domain, which may be

146     global or regional, using gridded data or at a point scale, e.g. using meteorological and

147     geophysical data from a FluxNet site. The primary physical and biogeochemical processes of

148     CLASSIC are briefly summarized in the next two sections.

149     **2.1.1 Physical processes**

150     The calculations for physical processes in CLASSIC are performed over vegetated, snow,

151     and bare fractions at a time step of 30 minutes. In the version used here, the fractional coverage

152     of the four plant functional types (PFTs) (needleleaf trees, broadleaf trees, crops, and grasses)

153    characterizes vegetation for each grid cell. The fractional coverage of these four PFTs is specified

154    over the historical period in this study. The structure of vegetation is characterized by leaf area

155    index (LAI), vegetation height, canopy mass, and rooting distribution through the soil layers all of

156    which are dynamically simulated by the biogeochemical module of CLASSIC. Twenty ground

157    layers represent the soil profile, starting with 10 layers of 0.1 m thickness. The thickness of layers

158    gradually increases to 30 m for a total ground depth of over 61 m. The depth of permeable soil

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

163    prognostically modelled. The result of physics calculations yields fluxes of energy (primarily net

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**

167    The biogeochemical processes in CLASSIC, based on CTEM, are described in detail in the

168    appendix of Melton and Arora (2016). The biogeochemical processes simulate the land-

169    atmosphere exchange of $CO_2$ and as a result simulate vegetation as a dynamic component

170    depending on the environmental conditions.

171    The biogeochemical module of CLASSIC prognostically calculates the amount of C in the

172    model's three live (leaves, stem, and root) and two dead (litter and soil) C pools for each PFT. The

173    live vegetation pools are separated into their structural and non-structural components. The C

174    amount in these pools is represented per unit land area (kg C/m$^2$). The amount of C in the live

175    and dead C pools and all terrestrial ecosystem processes in the biogeochemical module in this

176    study are modelled for nine PFTs that map directly onto the four base PFTs used in the physics

177    module of CLASSIC. Needleleaf trees are divided into their deciduous and evergreen phenotypes,

178    broadleaf trees are divided into cold deciduous, drought deciduous, and evergreen phenotypes,

179    and crops and grasses are divided based on their photosynthetic pathways into C$_3$ and C$_4$

180    versions. The sub-division of PFTs is essential for modelling biogeochemical processes. For

181    instance, simulating the onset and offset of leaves is different between evergreen and deciduous

182    phenotypes of needleleaf and broadleaf trees. However, once the leaf area index (LAI) is known,

183    a physical process does not need information about the underlying deciduous or evergreen

184    nature of leaf phenology. For example, the interception of rain and snow by canopy leaves (that

185    is typically modelled as a function of LAI and a PFT-dependent parameter that accounts for leaf

186    orientation and shape) does not depend on the underlying evergreen or deciduous nature of the

187    leaf phenology. In general, biogeochemical processes benefit more in terms of realism than

188    physical processes when the number of PFTs is increased. For example, in CLASSIC, large changes

189    in leaf area index (LAI) do not change total latent heat flux considerably since the partitioning of

190    evapotranspiration into its sub-components (transpiration, soil evaporation, and

191    evaporation/sublimation of intercepted rain/snow) adjusts. A decrease in transpiration and

192    evaporation of intercepted precipitation, due to a decrease in LAI, is compensated by an increase

193    in soil evaporation. This is expected since water and energy fluxes are determined largely by

194    available energy and precipitation.

195    The litter and soil C pools are tracked for each soil layer but the movement of C between

196    the soil layers is not yet modelled. Other than photosynthesis and leaf respiration which are

197    modelled at a time step of 30 minutes all other biogeochemical processes are modelled at a daily

198    time step. These include: 1) allocation of C from leaves to stem and roots, 2) autotrophic

199    respiration from the live C pools and heterotrophic respirations from the dead C pools, 3) leaf

200    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    study and fractional coverage of the nine PFTs is specified based on the representation of the

203    land cover as explained in the next section.

204    When the N cycle is turned on, land-atmosphere fluxes of $N_2O$, $NO_x$, and $NH_3$, and N

205    leaching are also modelled in response to biological N fixation, N fertilizer inputs, and N

206    deposition from the atmosphere. In particular, when the N cycle interacts with the C cycle, the

207    maximum photosynthetic capacities of model PFTs ($V_{c,max}$) are determined prognostically as a

208    function of their leaf N content (Asaadi and Arora, 2021; Kou-Giesbrecht and Arora, 2022). When

209    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 $CO_2$ increases (Arora et al., 2009). N in all model components (leaves,

212    stem, roots, litter, and soil organic matter) is prognostically tracked, and therefore C:N ratio of

213    all components is prognostically modelled except for soil organic matter for which a C:N ratio of

214    13 is specified. In addition, N in the soil mineral pools of nitrate ($NO_3^-$) and ammonium ($NH_4^+$) is

215    also prognostically modelled.

216    **3. Driving data for CLASSIC and model simulations**

**3.1 Land cover**

217

218   Land cover is one of the most important geophysical fields that is required by LSMs and

219   at its most basic level provides information about fractional vegetation cover in each grid cell for

220   a given regional or global domain. Vegetation in LSMs is typically represented in terms of PFTs.

221   Models may choose to represent a basic set of a few PFTs (trees, grasses, shrubs, and crops) or a

222   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   pathway ($C_3$ or $C_4$), and geographical location (tropical, temperate, or boreal). The version of

225   CLASSIC in this study uses a somewhat smaller set of nine PFTs for biogeochemical processes as

226   described in the previous section. The fractional coverage of PFTs in a model may be dynamically

227   simulated based on competition between PFTs or prescribed based on observation-based land

228   cover information. While CLASSIC does have a parameterization of competition between its PFTs

229   (Arora and Boer, 2006; Melton and Arora, 2016b), for the historical simulations considered here

230   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   The first is a remotely-sensed land cover product that represents the geographical distribution of

235   land cover for the present day for a short period. Examples of this include the GLC 2000 land

236   cover product which represents the November 1999 to December 2000 period

237   (https://forobs.jrc.ec.europa.eu/products/glc2000/glc2000.php) and the more recent European

238   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

240    spatially and temporally varying cropland (and pasture) area for a much longer period, which in

241    this case is based on the data set provided by the land use harmonization (LUH) product as part

242    of the TRENDY protocol for the period 850-2018. The LUH product is comprehensive (Hurtt et al.,

243    2020b). For example, not all models use the pasture area and other information provided in the

244    LUH product.

245         The process of generating land cover for a given model's PFTs is a three-step process.

246    First, the fractional coverage of model PFTs is obtained from a remotely sensed land cover

247    product that represents the snapshot of land cover for the present day. This requires typically

248    mapping 20 – 40 land cover classes that exist in a remotely-sensed land cover product to a given

249    model's PFTs. This step introduces the largest uncertainty in the entire process. The original land

250    cover in the CLASSIC model is based on the GLC 2000 land cover product. Table 2 of Wang et al.

251    (2006) summarizes the mapping/reclassification of the 22 GLC 2000 land cover categories to the

252    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    of 60% broadleaf deciduous trees, 20% grasses, and 20% bare ground. This first step yields a

256    snapshot of land cover expressed in terms of the fractional coverage of CLASSIC's nine PFTs. The

257    second step of generating fractional coverage of PFTs for a given snapshot in time requires

258    replacing the fractional area of crop categories with values from the LUH data set for the same

259    year. For example, when using the GLC 2000 land cover product, the area of $C_3$ and $C_4$ crops from

260    the LUH data set for the year 2000 are used, and the fractional coverage of the other seven non-

261    crop CLASSIC PFTs is adjusted such that the total vegetation fraction in each grid cell stays the

262    same. Finally, in the last step, the temporally varying crop area from the LUH product is used to

263    go backward in time to 1700 from the year 2000 with typically decreasing crop area while the

264    area of other non-crop PFTs is adjusted in proportion to their existing fractional coverage such

265    that the total vegetation fraction in each grid cell stays the same. Similarly, the area of $C_3$ and $C_4$

266    crops from the LUH product is used from the year 2000 onwards to the present day. All these

267    steps yield a reconstruction of historical land cover, expressed in terms of fractional coverage of

268    CLASSIC's nine PFTs (as interpreted from the GLC 2000 land cover product), from 1700 to 2018,

269    in which crop area changes spatially and temporally according to the LUH product.

270         GLC 2000 is an older land cover product and more recent land cover products are now

271    available. Here, in addition to the GLC 2000 based land cover, we also use the European Space

272    Agency (ESA) Climate Change Initiative (CCI) land cover product. The ESA CCI land cover product

273    is available at 300 m spatial resolution for the period 1992-2018 and contains 37 land cover

274    categories (ESA, 2017). We use the land cover from the year 1992 to create a snapshot of CLASSIC

275    PFTs for the present day. Although there is some interannual variability overall the total

276    vegetated area doesn't change substantially from 1992-2018 in the ESA-CCI land cover. A default

277    mapping/reclassification table for converting the ESA CCI classes into PFTs is provided in its user

278    guide (ESA, 2017). However, it overestimates tree cover along the taiga-tundra transition zone

279    and underestimates it elsewhere in Canada (Wang et al., 2018, 2019). Wang et al. (2022) have

280    developed a new reclassification table for converting the 37 ESA CCI land cover categories to

281    CLASSIC's nine PFTs which is used in this study. A high-resolution land cover map over Canada

282    and a tree cover fraction data at 30 m resolution are used to compute the sub-pixel fractional

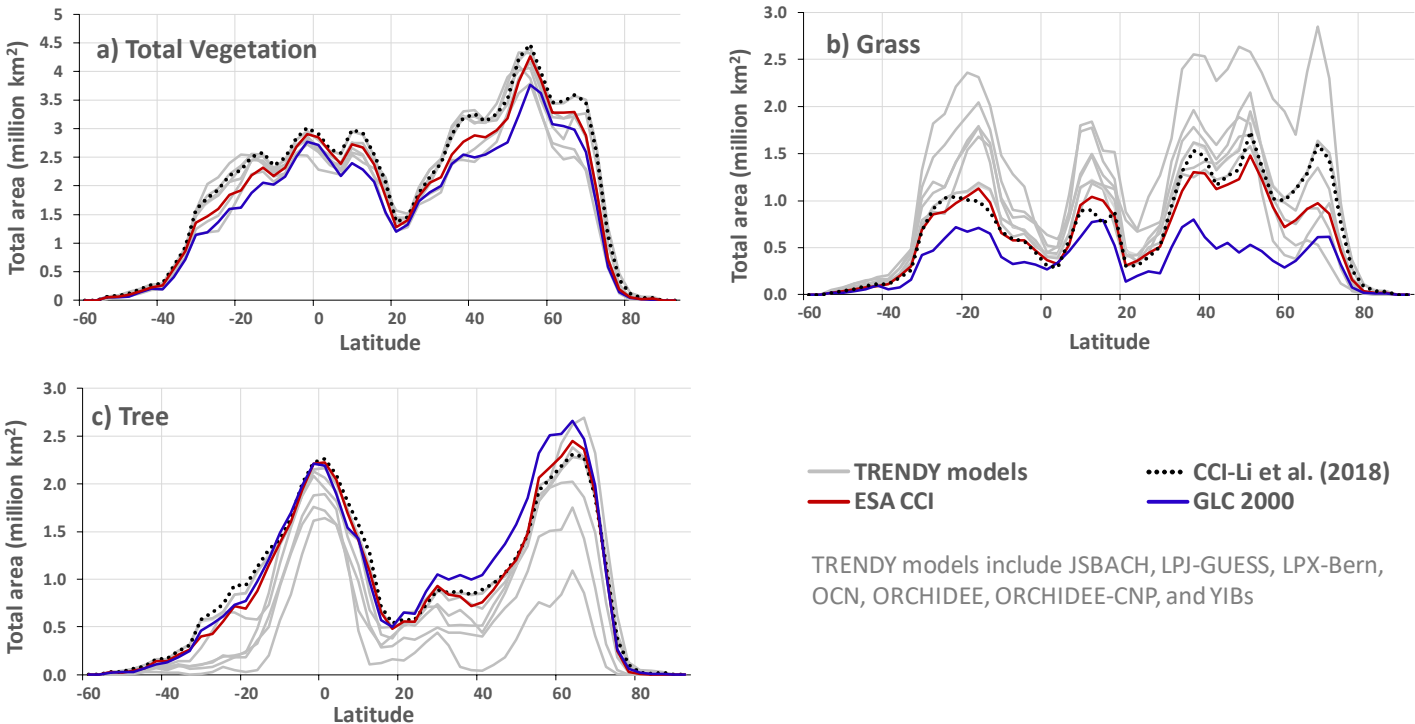
283 composition of each class in the ESA CCI dataset, which is then used to inform the cross-walking

284 reclassification procedure (Wang et al., 2022).

285

286

287

288

289 Figure 1: Comparison of zonally summed areas of total vegetation (a), grass (b), and tree (c) cover used
290 in the CLASSIC model based on GLC 2000 (blue line) and ESA CCI (dark red line) land cover products to
291 each other, to selected other models that participated in the 2020 TRENDY intercomparison (grey lines)
292 for which land cover information was available, and to Li et al. (2018) (dotted black line) who analyzed
293 the ESA CCI data. All data correspond to the 1992-2018 period. CLASSIC does not yet explicitly
294 represents shrub PFTs. Tall shrubs are merged into tree PFTs in CLASSIC. For the Li et al. (2018) data
295 plotted here, the shrub PFTs are combined with the tree PFTs for a consistent comparison to CLASSIC.

296

297

298     The above process yields two representations of land cover in which the geographical

299 distribution of CLASSIC PFTs is based on GLC 2000 and ESA CCI land cover products. Both these

300  representations include the same reconstruction of crop area over the historical period. Figure 1

301  illustrates the uncertainty in land cover by comparing zonally summed areas of total vegetation,

302  tree, and grass cover in CLASSIC, averaged over the period 1992-2018, when model land cover is

303  based on the GLC 2000 (blue line) and ESA CCI (dark red line) land cover products. These two

304  estimates are also compared to selected other models that participated in the 2020 TRENDY

305  intercomparison (grey lines), also for the period 1992-2018, for which land cover information was

306  available, and to Li et al. (2018b) (dotted black line) who analyzed the ESA CCI data based on the

307  default reclassification table from the ESA CCI user guide. Figure 1 shows while there is relatively

308  good agreement across TRENDY models in terms of total vegetation cover there's a much larger

309  uncertainty in its split between tree and grass PFTs. There are two reasons for the spread in total

310  vegetated, treed, and grassed areas across TRENDY models. First, modelling groups use different

311  remotely sensed land cover products for obtaining fractional cover of their model PFTs. Second,

312  the current process of mapping/reclassifying 20-40 land cover classes of a land cover product to

313  a model's PFTs is mainly based on the class description and expert judgment which introduces

314  subjectiveness in the process. Compared to the GLC 2000 based land cover in the CLASSIC model,

315  the newer ESA CCI based land cover yields a somewhat higher total vegetation cover, a higher

316  grass cover, and a somewhat lower tree cover area. Unlike the older GLC 2000 based land cover

317  used in CLASSIC, the newer ESA CCI based grass and tree cover area are within the range of the

318  TRENDY models reported here. Finally, Figure 1 also allows us to compare the results from the

319  analysis of Li et al. (2018b) for the ESA CCI land cover (dotted black line) to the ESA CCI

320  reclassification for CLASSIC (dark red line) by (Wang et al., 2022). Li et al. (2018b) used the default

321  mapping/reclassification table for converting the ESA CCI classes into PFTs. This comparison

322 illustrates that the remapping of the ESA CCI land cover classes to CLASSIC's PFTs yields total

323 vegetation, tree, and grass coverage that is broadly comparable to Li et al. (2018b) although some

324 differences remain for the grasses.

325 Our framework accounts for the uncertainty in land cover representation. However, since

326 both land cover representations in our study account for the increase in crop area over the

327 historical period in the same way by adjusting the area of non-crop PFTs in proportion to their

328 existing coverage using the LUH product, our framework is unable to account for the uncertainty

329 associated with the implementation of LUC. Di Vittorio et al. (2018) quantify this uncertainty by

330 implementing several approaches to account for the increase in crop area over the historical

331 period in the framework of an integrated assessment model: by preferentially converting grasses

332 and shrubs, by preferentially converting forests, and by proportionally adjusting areas of non-

333 crop PFTs in a way similar to ours. LUC emissions are higher if the increase in crop area is

334 preferentially obtained by converting forests. A similar uncertainty analysis for LUC emissions is

335 performed by Peng et al. (2017) using the ORCHIDEE land model who analyze the effect of using

336 different rules to incorporate the changes in crop and pasture area over the historical period. The

337 uncertainty related to incorporating LUC information to modify a model's land cover is further

338 illustrated in Di Vittorio et al. (2014) and Meiyappan and Jain (2012).

339

340 **3.2 Meteorological data**

341 As a land surface component of an ESM, CLASSIC requires meteorological forcing at a sub-

342 daily temporal resolution. In the offline simulations reported here, the model is run with half-

343    hourly values of meteorological data (incoming long and shortwave radiation, temperature,

344    precipitation, specific humidity, wind speed, and pressure). The first meteorological data set used

345    to drive CLASSIC is from the TRENDY protocol for the year 2020, CRU-JRA v2.1.5, which provides

346    6 hourly values of the seven variables from the Japanese reanalysis (JRA) with monthly values

347    adjusted to the climate research unit's data (CRU, https://crudata.uea.ac.uk/cru/data/hrg/). This

348    yields a blended product from year January 1901 to December 2019 with the 6-hourly temporal

349    resolution of a reanalysis but without the biases that may be present in reanalysis data (Harris,

350    2020). The second meteorological data set used here to drive CLASSIC is from the Global Soil

351    Wetness Project 3 (GSWP3). The GSWP3 forcing data are based on a dynamical downscaling of

352    the 20$^{th}$ century reanalysis (Compo et al., 2011) using a Global Spectral Model (GSM) run at about

353    50 km resolution. GSM is nudged towards the vertical structures of 20$^{th}$ century (20CR) zonal and

354    meridional air temperature and winds so that the synoptic features are retained at their higher

355    spatial resolution. Additional bias corrections are also performed as explained in van den Hurk et

356    al. (2016). The GSWP3 forcing is available for the 1901-2016 period. The 6-hourly values from

357    both the CRU-JRA and GSWP3 forcings are further disaggregated to half-hourly values for use by

358    CLASSIC.

359         Figure A1 compares the two meteorological forcings data sets, over the 1997-2016

360    period, to illustrate that although these two data sets are very similar there are differences

361    between the two. Global precipitation over land (excluding Greenland and Antarctica) in the

362    GSWP3 data set (857 mm/year) is somewhat higher than in the CRU-JRA data set (820 mm/year).

363    The global near-surface air temperature over land (excluding Greenland and Antarctica) is also

364    slightly higher in the GSWP3 data set (14.22 °C) compared to the CRU-JRA data set (14.08 °C). The

365 largest temperature difference occurs between the two data sets over the northern tropics

366 (panel h) where the GSWP3 data set is about 0.93 °C warmer than the CRU-JRA data set. The

367 geographical distribution of mean annual temperature is very similar between the two data sets

368 but there are some differences in the geographical distribution of precipitation (not shown).

369 Despite very similar total precipitation amounts and their seasonality over large global regions in

370 the two data sets, differences exist in the frequency distribution of precipitation. Figure A2

371 illustrates this over three broad regions, the Amazon, the Sahel, and the Midwest United States,

372 which shows the frequency distribution of daily precipitation amounts (mm/day) over the 1997-

373 2016 period from the two data sets. Figure A2 shows that the frequency of precipitation events

374 greater than about 5-10 mm/day is higher in the GSWP3 data set compared to the CRU-JRA data

375 set for the Amazonian, the Sahel, and the Midwest United States regions.

376 **3.3 Other forcings**

377 Other than the land cover and meteorological forcings CLASSIC requires globally averaged

378 atmospheric $CO_2$ concentration, geographically varying time-invariant soil texture and soil

379 permeable depth, population density, time-invariant monthly lightning, and geographically and

380 time-varying N fertilizer application rates and atmospheric N deposition rates. The atmospheric

381 $CO_2$ concentration values are provided by the TRENDY protocol. The soil texture information

382 consists of the percentage of sand, clay, and organic matter and is derived from Shangguan et

383 al. (2014). N fertilizer is specified according to the TRENDY protocol and based on Lu and Tian

384 (2017). N deposition is also specified according to the TRENDY protocol and based on model

385 forcings provided for the sixth phase of CMIP (CMIP6) through input4MIPs (Hegglin et al.,

386 2016). N deposition for the historical (1850-2014) period is used as is provided while that for

387    the period 2015-2018 is specified based on N deposition from the SSP5-85 scenario. For the

388    period 1700-1849, N deposition values from the year 1850 are used.

389
390    Table 1: Summary of simulations performed with two representations of the historical land
391    cover, two sets of meteorological data, and two versions of the CLASSIC land model.

392

| Simulation | Land cover reconstruction | Meteorological forcing | N cycle interactions with the C cycle |
|---|---|---|---|
| A | based on GLC 2000 | CRU-JRA v2.1.5 | On |
| B | based on GLC 2000 | GSWP3 | On |
| C | based on GLC 2000 | CRU-JRA v2.1.5 | Off |
| D | based on GLC 2000 | GSWP3 | Off |
| E | based on ESA CCI | CRU-JRA v2.1.5 | On |
| F | based on ESA CCI | GSWP3 | On |
| G | based on ESA CCI | CRU-JRA v2.1.5 | Off |
| H | based on ESA CCI | GSWP3 | Off |

393
394
395

396    **3.4 Model simulations**

397    Using the two representations of the historical land cover (based on the GLC 2000 and ESA CCI

398    land cover products), the two sets of meteorological data (CRU-JRA and GSWP3), and the two

399    versions of the CLASSIC model (with and without interactions between the C and N cycles) we

400    perform eight sets of pre-industrial and historical simulations as summarized in Table 1. Pre-

401    industrial simulations that correspond to the year 1700 are required before doing the historical

402    simulations (from which we analyze the model results) so that model pools can be spun up to

403    near equilibrium for each combination of land cover, meteorological forcing, and model

404    version. The pre-industrial simulations use 1901-1925 meteorological data repeatedly since this

405    period shows little trends in meteorological variables. Global thresholds of net atmosphere-

406    land C flux of 0.05 Pg/yr and net atmosphere-land N flux of 0.5 Tg N/yr, in simulations with the

407   N cycle turned on, are used to ensure the model pools have reached equilibrium. Each historical

408   simulation is then initialized from its corresponding pre-industrial simulation after it has

409   reached equilibrium. Simulations driven with the CRU-JRA meteorological data are performed

410   for the period 1701-2018, and the period 1701-2016 for simulations driven with the GSWP3

411   meteorological data. Similar to the pre-industrial simulations, meteorological data from 1901-

412   1925 is used repeatedly for the period 1701-1900. The global model simulations are performed

413   at a spatial resolution of about 2.81° (about 312 km at the equator) and the size of the spatial

414   longitude-latitude grid is 128 × 64. All model forcings are regridded to this common spatial

415   resolution. The model is run over about 1900 land grid cells at this resolution excluding glacial

416   cells in Greenland and Antarctica.

417

418   **3.5 Automated benchmarking**

419   The results from the eight CLASSIC simulations reported here are evaluated using an in-

420   house model benchmarking system called the Automated Model Benchmarking R package

421   (AMBER) (Seiler et al., 2021b). AMBER is based on a skill score system originally developed by

422   (Collier et al., 2018) which is used to quantify model performance and explained in detail in the

423   appendix. Five scores are used that assess a model's bias ($S_{bias}$), root-mean-square error ($S_{rmse}$),

424   seasonality ($S_{phase}$), interannual variability ($S_{iav}$), and spatial distribution ($S_{dist}$) against globally

425   gridded and in-situ data set(s) of observation-based estimates for a given quantity. A score is

426   computed by first calculating a dimensionless statistical metric, which is then scaled onto a unit

427   interval, and finally calculating its spatial mean. Scores range from 0 to 1 and are dimensionless.

428 Higher values indicate better performance. Finally, an overall score $S_{overall}$ is calculated as follows

429 by giving twice as much weight to $S_{rmse}$

430
$$S_{overall} = \frac{S_{bias} + 2S_{rmse} + S_{phase} + S_{iav} + S_{dist}}{1+2+1+1+1}.$$
(1)

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

432    The scores are calculated by comparing gridded and in-situ observation-based estimates,

433 referred to as reference data sets in Seiler et al. (2021b), for 19 energy (surface albedo, net

434 shortwave and longwave radiation, total net radiation, latent heat flux, sensible heat flux, ground

435 heat flux), water (soil moisture, snow, and runoff), and C cycle (GPP, net ecosystem exchange,

436 net biome productivity, aboveground biomass, soil C, LAI, area burnt, and fire $CO_2$ emissions)

437 related variables to model simulated quantities. Table 2 summarizes the source of these

438 observation-based data sets. The resulting model scores express to what extent simulated and

439 observation-based data agree. A low score does not necessarily indicate poor model

440 performance. Uncertainties in the meteorological forcing data and geophysical fields used to

441 drive the model, and/or in the observation-based data itself are possible reasons for the lack of

442 agreement. One way to assess uncertainties in observation-based data sets is to quantify the skill

443 score by comparing two independently-derived observation-based data sets (Seiler et al., 2022).

444 The resulting scores are referred to as benchmark scores and quantify the level of agreement

445 among the observation-based data sets themselves provided, of course, there are at least two

446 sets of observation-based data for a given quantity. The comparison of model scores against

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

448 data sets relative to the agreement between the observation-based data sets themselves.

449    Table 2: Observation-based data sets used for model evaluation in AMBER.

| 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 longwave radiation, net radiation | CERES | Radiative transfer model | Kato et al. (2013) |
| Net radiation, latent and sensible heat flux, ground heat flux, runoff | CLASSr | Blended product | Hobeichi et al. (2019) |
| Leaf area index | Copernicus | Artificial neural network | Verger et al. (2014) |
| Net biome productivity | 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 primary productivity | FLUXCOM | Machine learning | Jung et al. (2019, 2020) |
| Above ground biomass | GEOCARBON | Machine learning | Avitabile et al. (2016); Santoro et al. (2015) |
| Surface albedo, net shortwave and longwave radiation, net radiation | GEWEXSRB | Radiative transfer model | Stackhouse et al. (2011) |
| 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 Distribution Function | Strahler et al. (1999) |
| 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 sites)** | **Reference** |
| Leaf area index | CEOS | Transfer function (141) | Garrigues et al. (2008) |
| Latent, sensible, and ground heat flux, gross primary productivity, ecosystem respiration, net ecosystem exchange | FLUXNET 2015 | Eddy covariance (204) | Pastorello et al. (2020) |
| 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) |

450

451

452    **4. Results**

453       Figures 2 through 9 show the time series and/or zonally-averaged values of annual values

454    of a variable of interest when averaged across four ensemble members each according to

455    whether the N cycle is turned on or not, whether the GLC 2000 or ESA CCI based land cover is

456    used, and whether model simulations are driven by the CRU-JRA or GSWP3 meteorological data.

457    Figures A3, A4, A6, A7, A9, and A11 in the appendix, which are complementary to the above-

458    mentioned figures, show the physical and biogeochemical states of the land surface and primary

459    physical fluxes of water and energy, and primary biogeochemical fluxes of $CO_2$ simulated by

460    CLASSIC at the land-atmosphere boundary for all the eight simulations considered here. While

461    the figures in the appendix illustrate the range in simulated physical and biogeochemical states

462    and fluxes across the eight simulations, Figures 2 through 9 evaluate the effect of model

463    structure, meteorological forcing, and land cover on a given quantity. We also quantify the spread

464    across the eight simulations using the coefficient of variation (cv= standard deviation/mean)

465    calculated using annual global values for a given quantity averaged over the 1997-2016 20-year

466    period of each simulation. This time period is also used for other reported results.

467    **4.1 Physical land surface state and fluxes**

468       Figure A3, panels a and b, shows the globally-averaged simulated soil moisture and

469    temperature in the top 1 m soil layer. While simulated soil temperature in the top 1 m is fairly

470    similar across the eight simulations, the simulated soil moisture is distinctly separated into two

471    groups. The separation into these two groups is caused by the driving meteorological data as

472    shown in Figure 2. The coefficient of variation for soil moisture and temperature values averaged

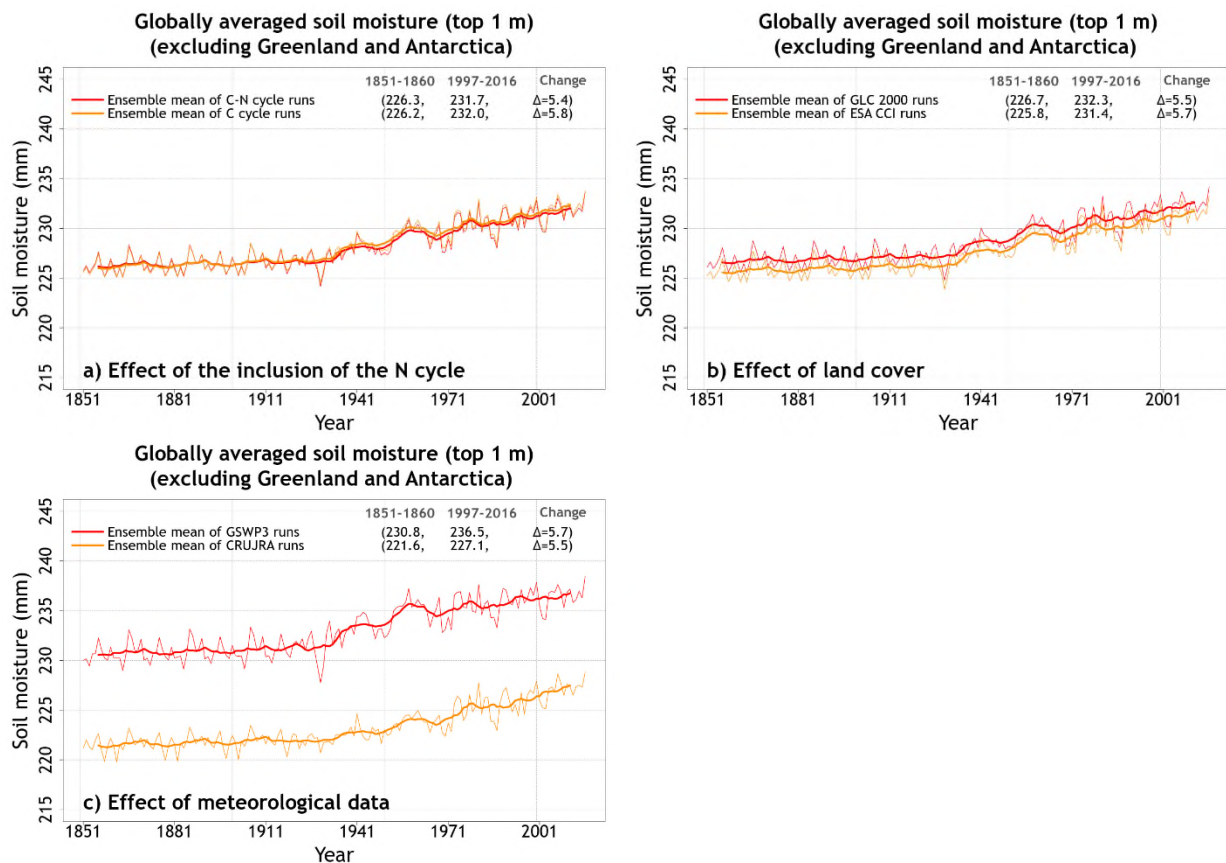
473    over the 1997-2016 period of each simulation are 0.02 and 0.004, respectively, indicating that

474    overall the variation in these quantities is relatively small compared to their means. The use of

475    the GSWP3 meteorological dataset yields slightly higher (~4%) globally-averaged soil moisture

476    compared to the CRU-JRA meteorological data set (236.5 mm vs. 227.1 mm, Figure 2c).

477

478

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

486

487

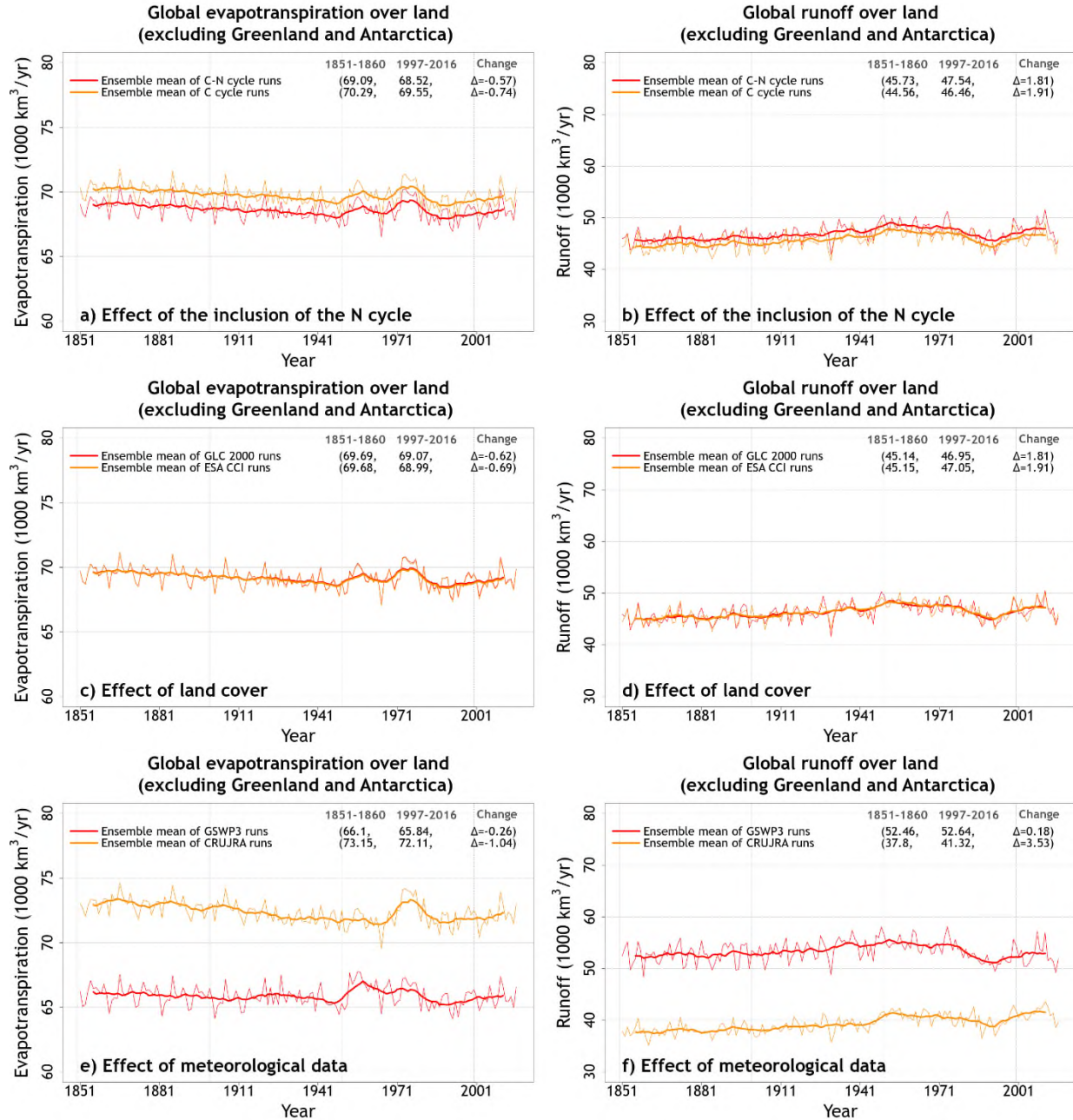



Figure 3: Time series of annual global evapotranspiration and runoff (over all land area excluding Greenland and Antarctica) averaged over the four ensemble members that are driven with and without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land cover (panels c, d), and driven with the 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, for each ensemble averaged over its set of four simulations are also shown.

496

497     Figure A3, panels c and d, shows the simulated fluxes of global evapotranspiration and

498     runoff across the eight simulations. Similar to soil moisture, evapotranspiration and runoff also

499     fall broadly into two groups and the reason for this again is the driving meteorological data.

500     Figure 3 shows that while the biggest factor that affects evapotranspiration and runoff is the

501     difference in driving meteorological data the interactive N cycle also affects these water fluxes.

502     Neither evapotranspiration nor runoff is significantly affected by the choice of land cover. The

503     reason an interactive N cycle affects evapotranspiration is that the N cycle in CLASSIC affects the

504     rate of photosynthesis through the prognostic determination of leaf N content. Photosynthesis

505     in turn affects canopy conductance, which affects transpiration through the canopy leaves.

506     Average evapotranspiration over the 1997-2016 period of the simulations driven with GSWP3

507     meteorological data is about 9% lower than in simulations driven with CRU-JRA meteorological

508     data (65.8 vs. 72.1 ×1000 km$^3$/year, Figure 3, panel e). An interactive N cycle reduces

509     evapotranspiration by about 2% due to lower photosynthesis rates as shown later (Figure 3, panel

510     a). Average runoff is about 27% higher in simulations driven with GSWP3 compared to

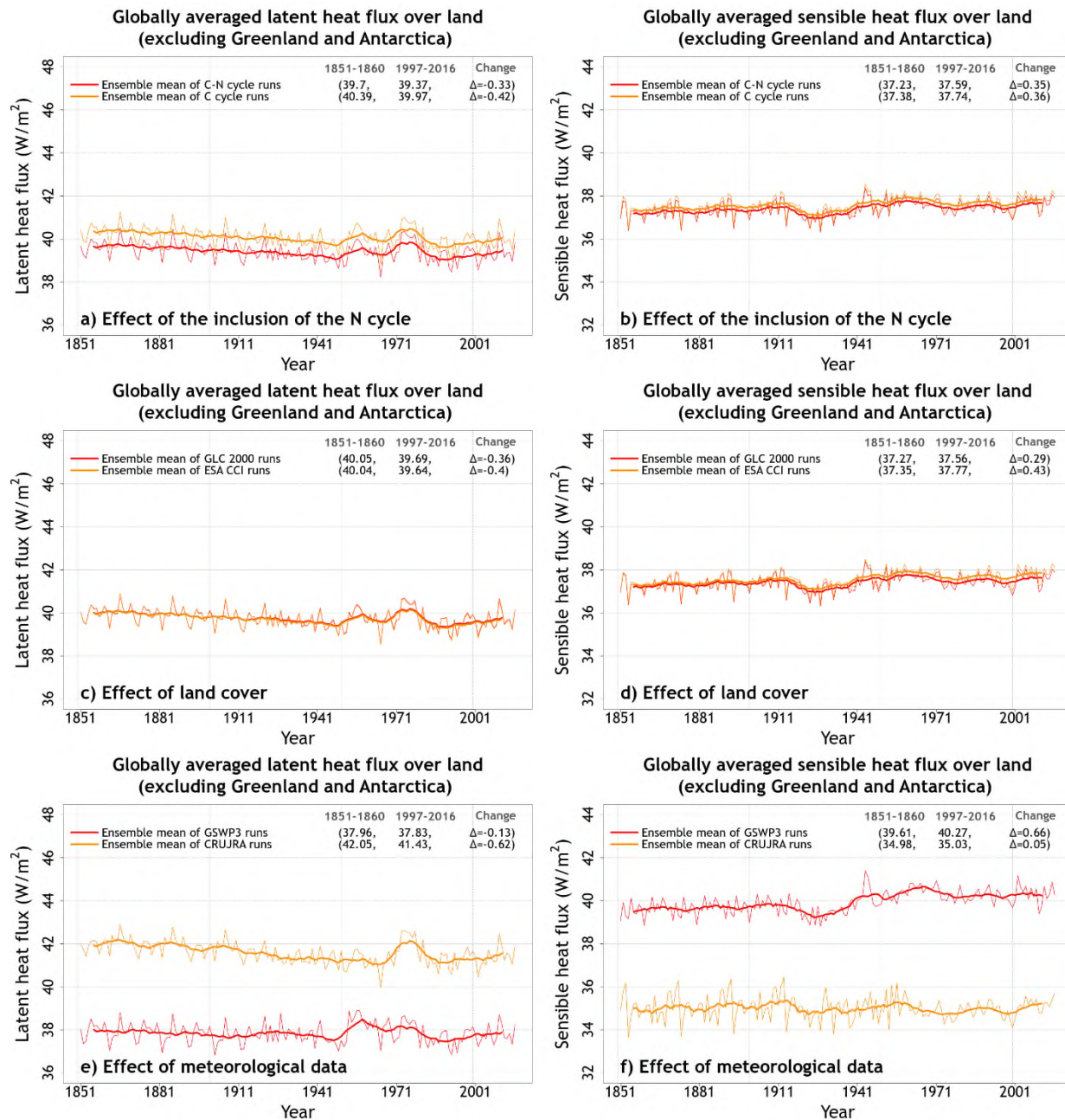
511     simulations driven with CRU-JRA meteorological data (52.6 vs 41.3 ×1000 km$^3$/year, Figure 3,

512     panel f). This is due to slightly high precipitation in the GSWP3 meteorological data set (Figure

513     A1) but is more so due to the simulated lower evapotranspiration when using the GSWP3 data

514     (Figure 3, panel e). The coefficient of variation for evapotranspiration and runoff values averaged

515     over the last 20 years of each simulation are 0.05 and 0.13, respectively.

516

517 Figure 4: Time series of annual global latent and sensible heat fluxes (over all land area excluding
518 Greenland and Antarctica) averaged over the four ensemble members that are driven with and
519 without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land
520 cover (panels c, d), and driven with GSWP3 and CRU-JRA meteorological data (panels e, f). The
521 thin lines show the individual years and the thick lines show their 11-year moving average. Model
522 values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods,
523 and their difference, for each ensemble averaged over its set of four simulations are also shown.

524

525       Figure A4 shows the primary energy fluxes from the eight simulations. These include net

526    downward shortwave and longwave radiation, and latent and sensible heat fluxes. Incoming

527    shortwave and longwave radiation are part of the driving meteorological data. Similar to water

528    fluxes, the differences in energy fluxes in CLASSIC are also primarily driven by differences in

529    meteorological data (Figure A4, A5, and Figure 4). Net shortwave radiation (Figure A4, panel a) is

530    equal to incoming shortwave radiation minus the fraction that is reflected back. Net longwave

531    radiation (Figure A4, panel b) is equal to incoming longwave radiation minus the longwave

532    radiation emitted by the land based on its surface temperature following the Stefan-Boltzmann

533    law. The difference in net shortwave radiation is also affected by simulated vegetation biomass

534    and leaf area index. The latter affects surface albedo which determines what fraction of incoming

535    shortwave radiation is reflected. This is the reason why an interactive N cycle affects net

536    shortwave radiation since the N cycle affects photosynthesis, and in turn, simulated vegetation

537    biomass and leaf area index (Figure A5, panel b). Latent heat flux is affected primarily by

538    meteorological data (Figure 4) but also if the N cycle is interactive or not since it is essentially

539    evapotranspiration but in energy units. Finally, differences in sensible heat fluxes are strongly

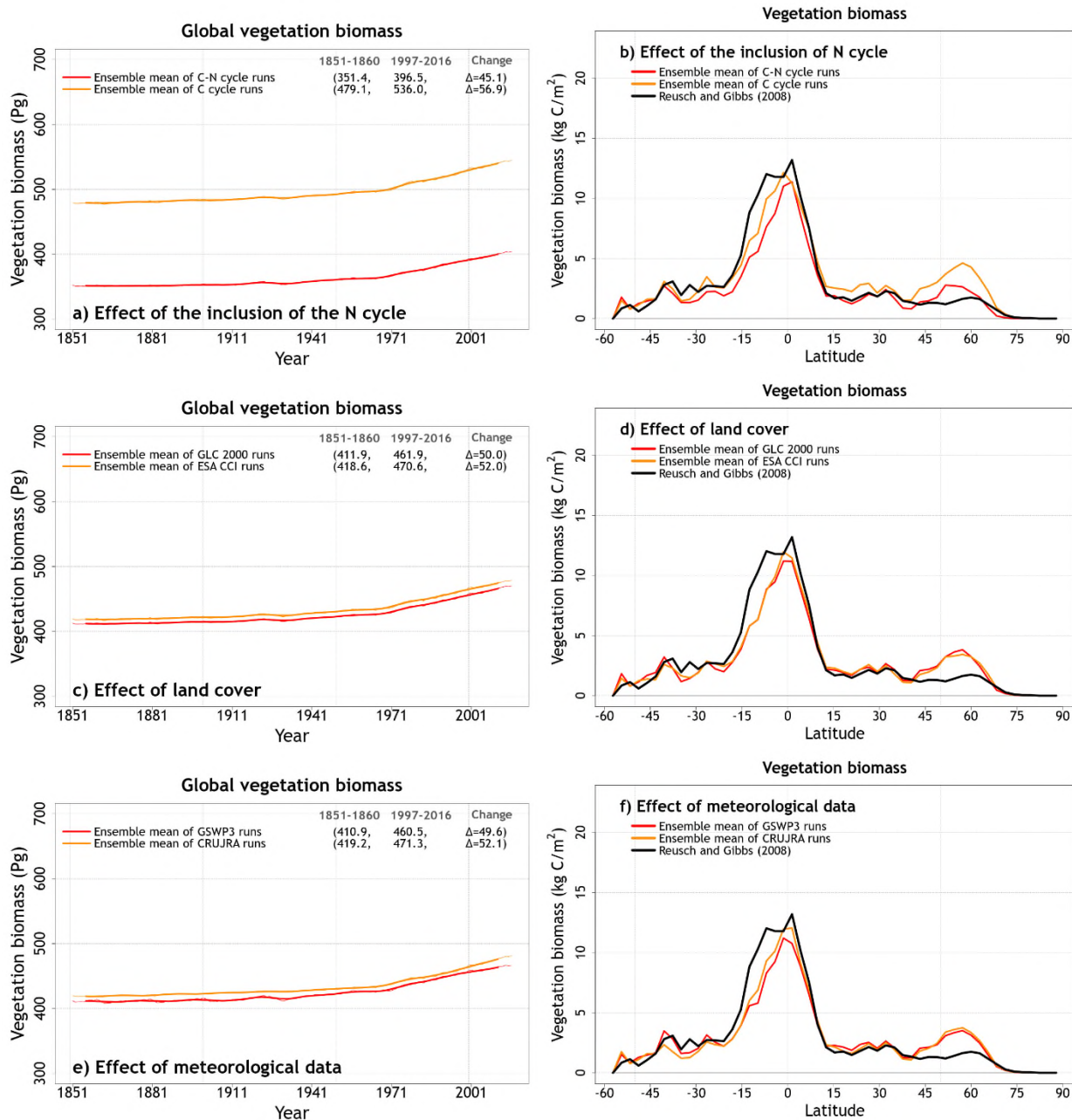
540    affected by differences in driving meteorological data (Figure 4). Globally-averaged sensible heat

541    flux in the simulations driven with GSWP3 data is ~14% higher compared to CRU-JRA driven

542    simulations (40.2 vs. 35.0 W/m$^2$). The coefficient of variation for latent and sensible heat flux

543    values averaged over the last 20 years of each simulation are 0.05 and  0.07, respectively. Net

544    shortwave (cv=0.006) and longwave (cv=0.03) radiative fluxes vary little across the eight

545    simulations.

546
547 Figure 5: Time series of annual global vegetation C mass (over all land area excluding Greenland
548 and Antarctica) (panels a, c, and e) and zonally-averaged values of vegetation C mass over land
549 (panels b, d, and f) averaged over the four ensemble members that are driven with and without
550 an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land cover
551 (panels c, d), and driven with GSWP3 and CRU-JRA meteorological data (panels e, f). The thin
552 lines for the time series show the individual years and the thick lines show their 11-year moving
553 average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-
554 2016) time periods, and their difference, are also shown in panels a, c, and e.

555

**4.2 Biogeochemical land surface state and fluxes**

**4.2.1 Primary CO$_2$ fluxes and C pools**

Figure A6 shows the simulated C state of the land surface expressed in terms of vegetation and soil C pools. Panels a and b show the annual time series of global vegetation and soil C mass from the eight simulations, and panels c and d show their zonally-averaged distributions averaged over the last 20 years of each simulation. The biggest difference in the time series of 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 eight simulations as indicated by their high cv values. The zonally-averaged values further provide 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 Figure A6 also show observation-based zonally-averaged values of vegetation and soil C mass 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 is provided in Section 4.3.

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

576

577

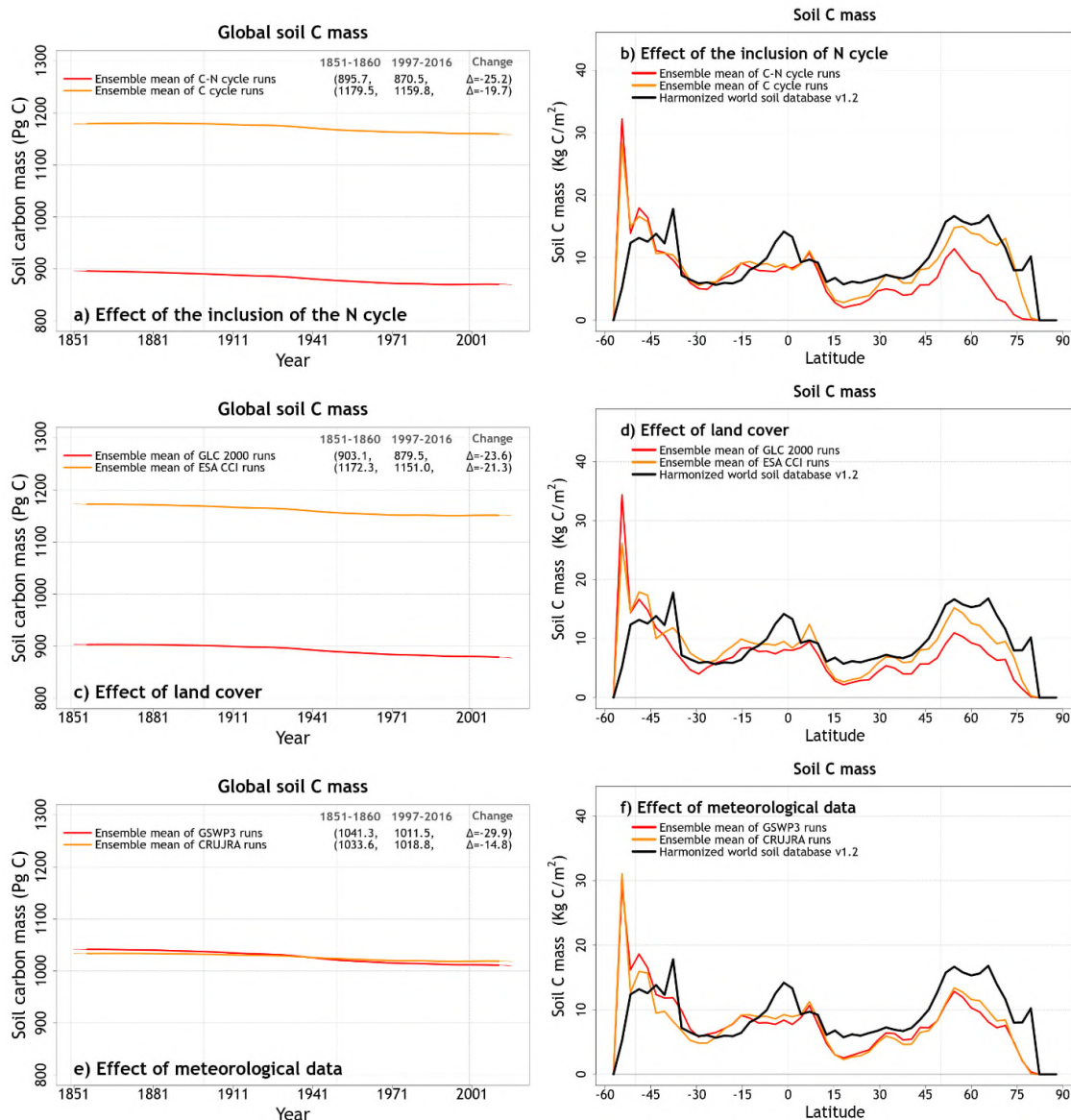



Figure 6: Time series of annual global soil carbon mass (over all land area excluding Greenland and Antarctica) (panels a, c, and e) and zonally-averaged values of soil carbon mass over land (panels b, d, and f) averaged over the four ensemble members that are driven with and without an interactive N 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 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) and present-day (1997-2016) time periods, and their difference, are also shown in panels a, c, and e.

586

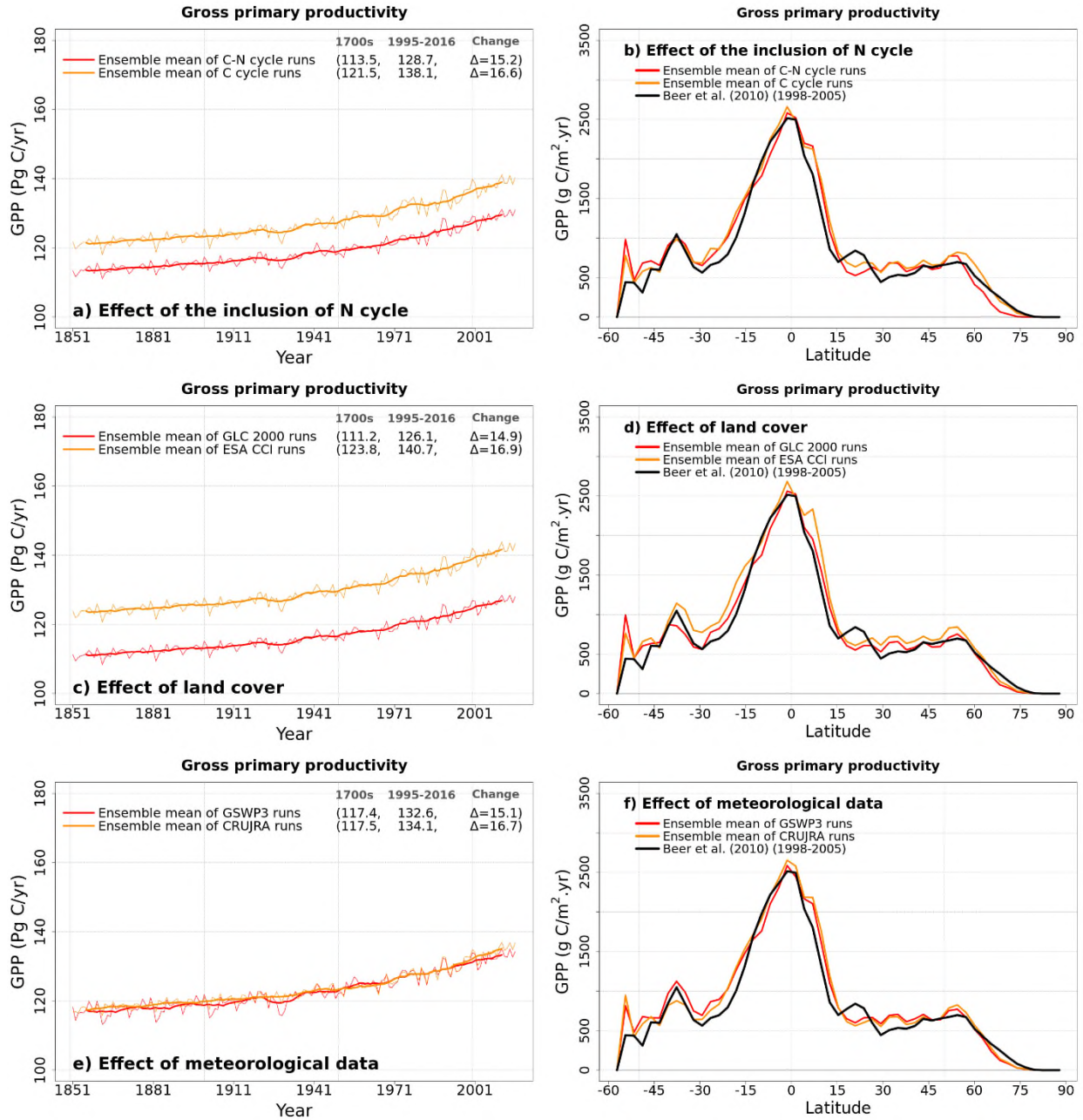



Figure 7: Time series of annual global gross primary productivity (over all land area excluding Greenland and Antarctica) (panels a, c, and e) and zonally-averaged values of gross primary productivity over land (panels b, d, and f) averaged over the four ensemble members that are driven with and without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land 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 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.

595  N cycle but also the choice of land cover (Figure 6). Since CLASSIC assumes that litter from grasses

596  is more recalcitrant than that from trees, the choice of ESA CCI based land cover leads to a higher

597  soil C mass because it has a higher grass area than the GLC 2000 based land cover (Figure 6,

598  panels c and d). The choice of meteorological data does not affect the magnitude of simulated

599  globally-summed soil C mass significantly but does affect its change over the historical period. In

600  Figure 6 (panel c) the decrease in soil C mass from the 1851-1860 period to the 1997-2016 period

601  is higher when using the GSWP3 (29.9 Pg C) compared to when using the CRU-JRA (14.8 Pg C)

602  meteorological data.

603      The reason why an interactive N cycle in CLASSIC affects vegetation C and soil C mass, and

604  why the ESA CCI based land cover yields high soil C, is seen in Figures A7 and 7. Figure A7 shows

605  the spread of primary C fluxes including gross primary productivity (GPP) (cv=0.07), and

606  autotrophic (cv=0.04) and heterotrophic (cv=0.10) respiratory fluxes, across the eight

607  simulations.  Since GPP is lower in the runs with the N cycle, both vegetation (Figure 5a) and soil

608  C mass (Figure 6a) are also lower. The lower GPP in the runs with the N cycle is due primarily to

609  lower GPP at high latitudes (Figure 7b) which yields low vegetation C mass at high latitudes

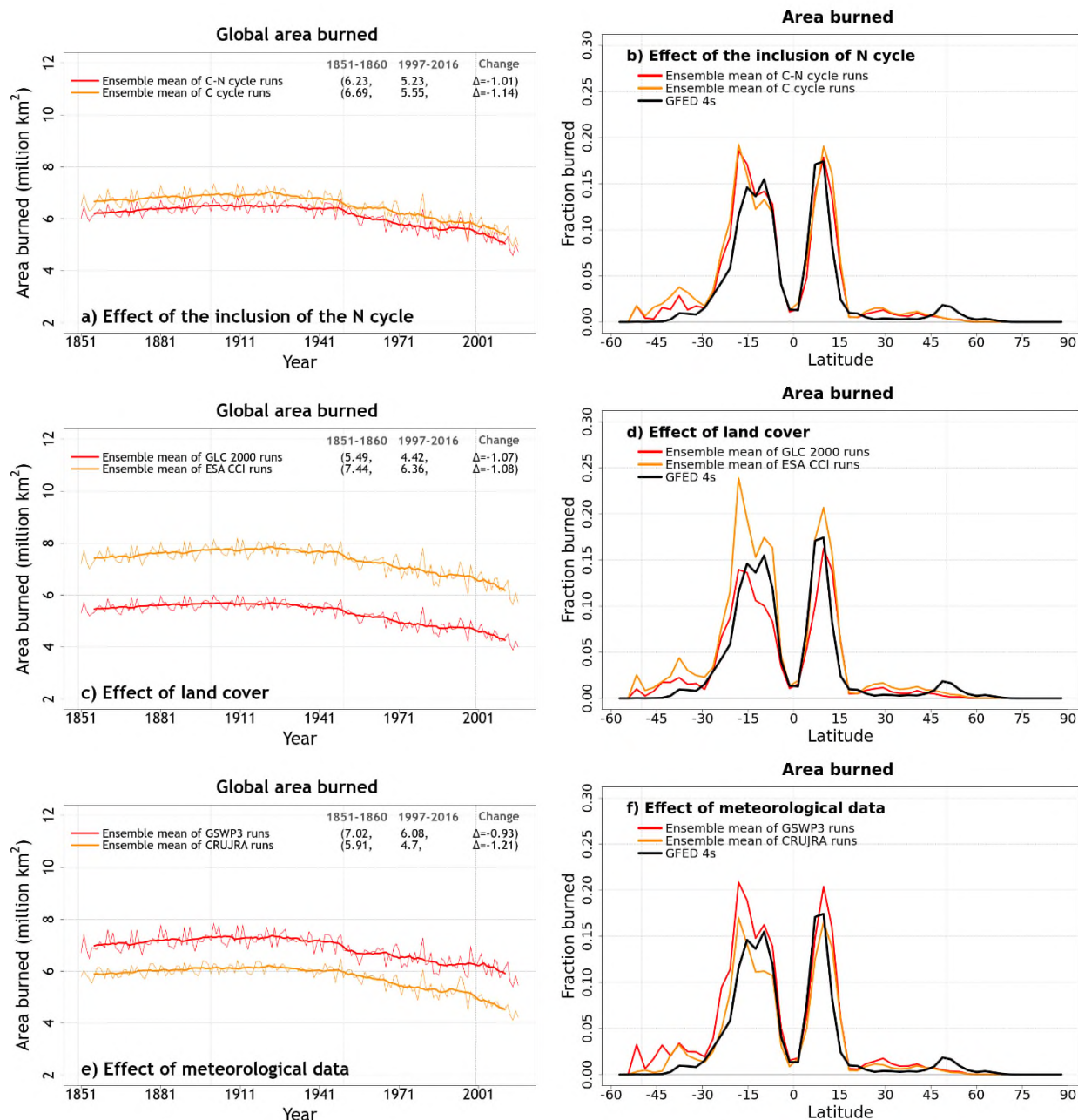
610  (Figure 5b). Low GPP at high latitudes translates to even larger relative differences in soil C given

611  the longer turnover time scales of soil C at high latitudes (Figure 6b). The use of the ESA CCI based

612  land cover which has a higher grass area than the GLC 2000 based land cover leads to higher GPP

613  (Figure 7d) and therefore higher soil C at all latitudes (Figure 6d). In Figure A8, global

614  heterotrophic and autotrophic respiratory fluxes are most affected by land cover and the

615  inclusion or absence of an interactive N cycle but not by the driving meteorological data.

616
617 Figure 8: Time series of annual area burned (over all land area excluding Greenland and
618 Antarctica) (panels a, c, and e) and zonally-averaged values of area burned (panels b, d, and f)
619 averaged over the four ensemble members that are driven with and without an interactive N
620 cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land cover (panels c, d), and
621 driven with GSWP3 and CRU-JRA meteorological data (panels e, f). The thin lines for the time
622 series show the individual years and the thick lines show their 11-year moving average in panels
623 (a), (c), and (e). Model values averaged over the pre-industrial (1851-1860) and present-day
624 (1997-2016) time periods, and their difference, are also shown for panels (a), (c), and (e).
625

626    Overall, while the primary biogeochemical fluxes (cv values vary from 0.04 to 0.10) vary

627    as much as the water and energy fluxes, the resulting spread in vegetation C mass (cv=0.16) and

628    soil C mass (cv=0.21) across the eight simulations is much larger and driven primarily by the

629    inclusion or absence of an interactive N cycle and the difference in land cover.

630

631    **4.2.2 Area burned and fire $CO_2$ emissions**

632    Figure A9 shows the time series of global area burned and global fire $CO_2$ emissions, and

633    their zonally-averaged values. We chose the area burned (cv=0.24) and fire $CO_2$ emissions

634    (cv=0.21) in addition to the primary biogeochemical fluxes since fire shows large variability both

635    in space and in time, and both these variables yield the largest spread across the eight

636    simulations, among all the fluxes and simulated quantities considered here. Figures A9 (panels c

637    and d) also show observation-based estimates for area burned and fire $CO_2$ emissions based on

638    GFED 4s (Giglio et al., 2013) to provide an observation-based context. Figures 8 and A10 help us

639    understand which factors contribute to this large variability. The variability in the area burned is

640    caused primarily by the choice of land cover and meteorological data and the variability is higher

641    in the southern hemisphere (Figure 8, panels d and f). An interactive N cycle does not affect the

642    zonal distribution of area burned and fire $CO_2$ emissions (Figures 8 and A10) as much. The reason

643    both area burned and fire $CO_2$ emissions are affected by the choice of land cover is because the

644    ESA CCI land cover has higher grass area and, as a result, it yields higher area burned and fire $CO_2$

645    emissions since a larger area is burned for grasses than for trees in the model. The choice of

646    driving meteorological data is a factor in the area burned and our simulations show that the use

647    of GSWP3 meteorological forcing yields a higher area burned than the CRU-JRA data. In particular

648 wind speed, which determines the rate of spread of fire in CLASSIC, is much higher in the GWSP3

649 than in the CRU-JRA meteorological data. Globally-averaged land wind speed (excluding

650 Greenland and Antarctica) in GSWP3 data is 6.1 m/s compared to 3.4 m/s in the CRU-JRA data

651 for the period 2000-2016.

652 Table 3: Simulated energy, water, and carbon cycle quantities considered in this study sorted
653 according to their coefficient of variation. The quantities are listed from the most variable at
654 the top to the least variable at the bottom. The coefficient of variation is based on annual
655 values averaged over the 1997-2016 period across the eight simulations. The last column shows
656 the dominant source of variability for each model simulated quantity.
657

| Energy, water, or carbon cycle quantities | Coefficient of variation | Dominant source of variability |
|---|---|---|
| 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 the N cycle |
| Vegetation carbon mass (Pg C) | 0.16 | The inclusion or the absence of the N cycle |
| Runoff (1000 km$^3$/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 (1000 km$^3$/year) | 0.05 | Meteorological forcing |
| 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 |

658

659 **4.2.3 Coefficient of variation summary**

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

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

662  the bottom according to their coefficient of variation. The area burned is found to be the most

663  variable quantity and soil temperature is the least variable quantity. Table 3 also shows the most

664  dominant source of variability for each simulated quantity: land cover, meteorological forcings,

665  or the inclusion or absence of an interactive N cycle.  Net atmosphere-land $CO_2$ flux (or net biome

666  productivity) and ground heat flux are not included in Table 3 because these fluxes are calculated

667  as the difference of larger fluxes and as a result, their values are closer to zero which yields a

668  large value of the coefficient of variation.

669  Overall, the results presented so far illustrate that different model simulated quantities

670  are sensitive to different forcings and model versions. As such it is not advisable to tune a model

671  to match observations when driven with a specific forcing data set.

**4.2.4 Net biome productivity**

673  Figure A11 shows the spread in the time series of annual global net atmosphere-land $CO_2$

674  flux and their zonally-averaged values across the eight simulations averaged over the 1995-2016

675  period from each simulation. The global net atmosphere-land $CO_2$ flux or net biome productivity

676  (NBP) is considered a critical determinant of the performance of LSMs, and is treated as such by

677  TRENDY, because this flux ultimately affects the changes in the atmospheric $CO_2$ burden. TRENDY

678  requires that LSMs simulate a terrestrial C sink for the decades of the 1990s to the present to be

679  considered for inclusion in the TRENDY ensemble.

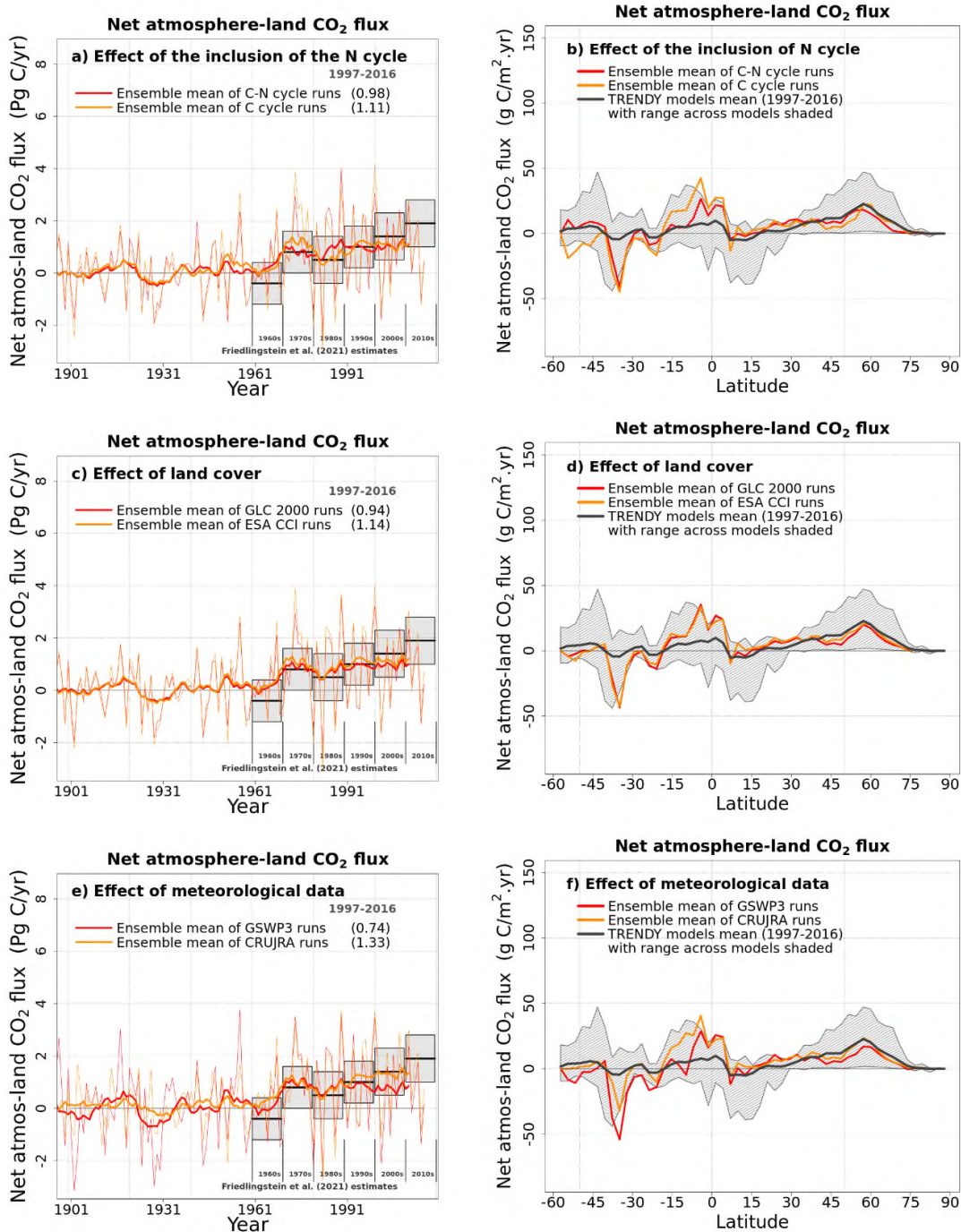



Figure 9: Time series of global net atmosphere-land $CO_2$ flux (over all land area excluding Greenland and Antarctica) (panels a, c, and e) and its zonally-averaged values (panels b, d, and f) averaged over the four ensemble members that are driven with and without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land 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 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).

688    Figure A11 also shows the estimates of global net atmosphere-land $CO_2$ flux from the

689    participating TRENDY models in grey boxes with mean and shaded ranges for the decades from

690    the 1960s to 2010s from the Global Carbon Project (Friedlingstein et al., 2022). Positive values in

691    Figure A11 indicate a C sink over land and negative values a C source to the atmosphere. In Figure

692    A11a, all eight simulations reported here would qualify for inclusion in the TRENDY ensemble

693    since they all simulate a terrestrial C sink from the 1990s to the present day. Before 1960, since

694    the atmospheric $CO_2$ concentration is not high enough,  the model yields both a land C sink and

695    source in response to interannual variability in meteorological data. In addition, the time series

696    of global NBP from all eight simulations lie within the uncertainty range of reported estimates

697    from the Global Carbon Project. Figure A11a suggests that based on global NBP, at least, it is not

698    possible to exclude any of the eight simulations. In Figure A11b, zonally-averaged NBP averaged

699    over the 1997-2016 period from each of the eight simulations mostly lie within the range of NBP

700    simulated by models that participated in TRENDY 2020. CLASSIC simulates a C sink at northern

701    high latitudes consistent with TRENDY models but it simulates a C sink on the stronger side of

702    TRENDY models in the southern tropics (0° - 20°S). This is likely because CLASSIC is known to

703    simulate low C emissions associated with LUC most of which are generated in tropical regions

704    (Asaadi and Arora, 2021).

705    Figure 9 provides additional insights into the effect of different forcings on the simulated

706    NBP. In Figure 9, averaged over the 1997-2016 period, an interactive N cycle leads to a somewhat

707    weaker C sink (panel a, 0.98 vs. 1.11 Pg C/yr), the choice of the ESA CCI based land cover leads to

708    a somewhat stronger C sink (panel c, 1.14 vs 0.94 Pg C/yr), and the choice of the GSWP3

709    meteorological data leads to a much weaker C sink (panel e, 0.74 vs 1.33 Pg C/yr) than the CRU-

710     JRA meteorological data. In Figure 9, panels a and b, the largest difference between the model

711     versions with and without the N cycle occurs in the tropics (~ 5°N - 20°S) where an interactive N

712     cycle leads to a weaker C sink. There are differences in zonally-averaged NBP with and without

713     the N cycle south of 45°S but the land area below this latitude is small so the averages are

714     calculated over only a few grid cells. The choice of the land cover (Figure 9, panels c and d) does

715     not substantially change the distribution of the zonally-averaged values of NBP although, as

716     noted above, the choice of ESA CCI based land cover leads to a somewhat stronger C sink. Finally,

717     the choice of the GSWP3 meteorological forcing leads to a weaker C sink at most latitudes (Figure

718     9, panels e and f).

719     **4.3 Automated benchmarking**

720     Figure 10 plots the overall score, $S_{overall}$, against benchmark scores for several of the energy,

721     water, and C cycle related variables. AMBER does not yet evaluate N cycle related variables for

722     which observations are more scarce than for C cycle related variables. The whiskers show the

723     range in the overall score both for the benchmark and model scores. The vertical whiskers show

724     the range of eight model scores when a given variable from all eight model simulations is

725     compared to an observation-based data set. The horizontal whiskers show the range when three

726     or more observation-based datasets are compared to each other. When only two observation-

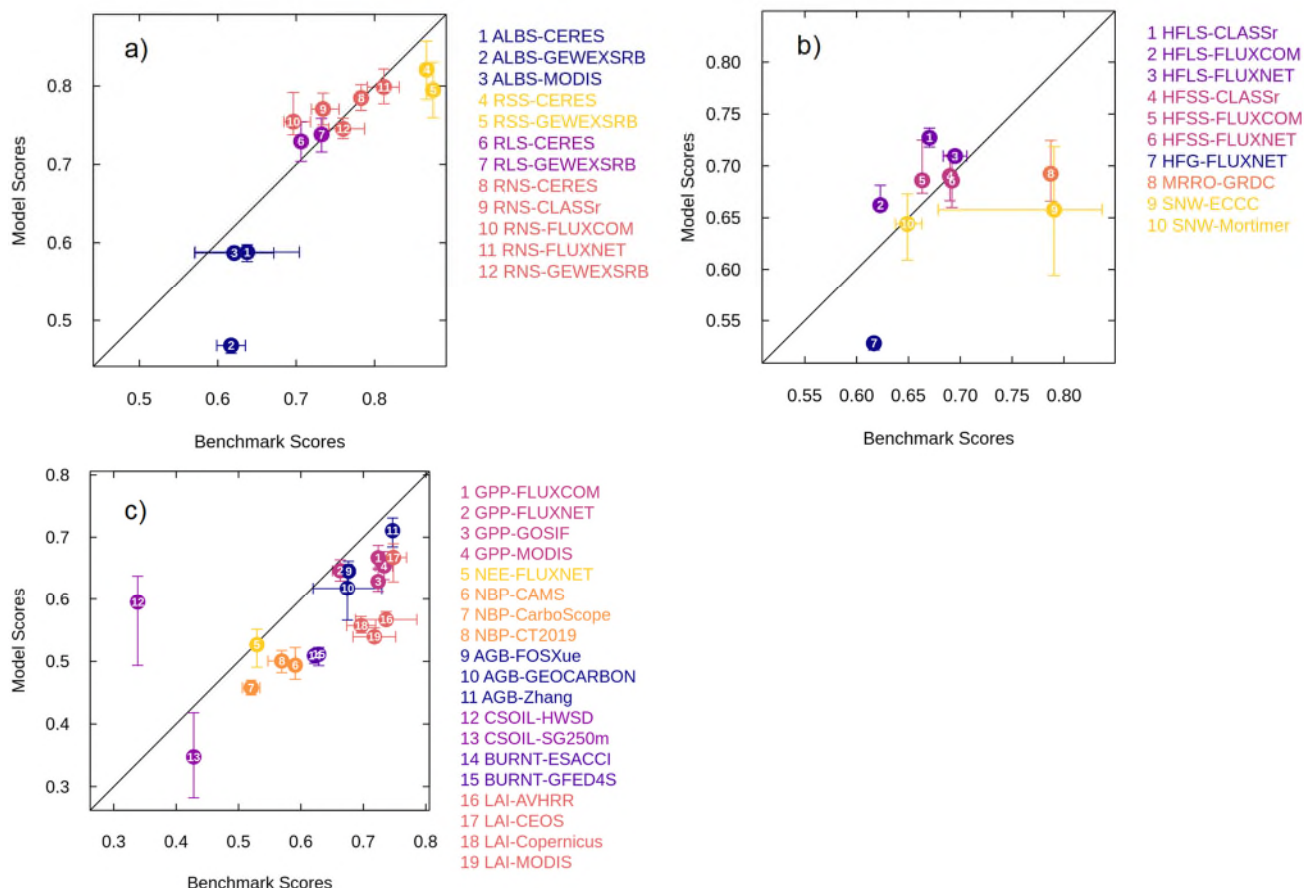
727     based data sets are compared to each other there is only one benchmark score, and therefore

728     there is no range. The range in model scores comes from the eight simulations, and the range in

729     benchmark scores comes from the different observation-based data sets. Figure 10 shows that

730     typically as the benchmark scores increase so do the overall model scores for a given quantity.

731 This indicates that uncertainty in observation-based estimates themselves leads to a poor

732 agreement between observations and model-simulated quantities.

733

734 Figure 10: Comparison of benchmark scores with model overall scores for a range of energy-,
735 water-, and carbon-related quantities. The whiskers indicate the range for benchmark scores
736 across different observation-based data sets and the range across the eight model simulations
737 for the overall model scores. The quantities in panel (a) are ALBS (surface albedo), RSS (net
738 shortwave radiation), RLS (net longwave radiation), and RNS (net radiation). Quantities in panel
739 (b) are HFLS (latent heat flux), HFSS (sensible heat flux), HFG (ground heat flux), MRRO (runoff),
740 and SNW (snow water equivalent). Quantities in panel (c) are GPP (gross primary productivity),
741 NEE (net ecosystem exchange), NBP (net biome productivity), AGB (aboveground biomass), CSOIL
742 (soil carbon mass), BURNT (area burned), and LAI (leaf area index).

743

744 For energy and water fluxes scores (panels a and b) the model overall scores lie around the 1:1

745 line indicating that model scores are generally as good as the benchmark scores, except for

746 surface albedo (ALBS), runoff (MRRO), ground heat flux (HFG), and comparison against one

747 observation-based estimate of snow water equivalent which lie below the 1:1 line. For C cycle

748 related variables most scores lie somewhat below the 1:1 line indicating that simulated quantities

749 do not agree as well with observations as observations agree among themselves. The lower

750 benchmark score for soil C (panel c) is because the SoilGrids250m (SG250m) data and the

751 Harmonized World Soil Database (HWSD) do not agree well amongst themselves because the

752 SG250m soil C data includes peatlands and permafrost C at high latitudes while the HWSD data

753 does not (see Figure 11b). Since the version of CLASSIC used here does not represent peatlands

754 and permafrost C it compares better with the HWSD data than with the SG250m data. In the case

755 of soil C, the choice of HSWD data for comparison against model values is obvious. However, for

756 other variables, it may not always be obvious which observation-based estimate is more

757 appropriate or better for comparison against model results. The uncertainty in forcing data sets

758 and in observation-based estimates, against which model results are evaluated, implies that even

759 a perfect model cannot be evaluated to its fullest extent.

760 Figure 11 shows the zonal distribution of vegetation C mass, LAI, area burnt, GPP, and fire

761 $CO_2$ emissions (which constitute standard output from AMBER) and illustrates how AMBER

762 compares the spread across the simulations indicated by 50%, 80%, and 100% shading against

763 observation-based estimates. The black and shades of grey indicate the model mean and the

764 spread across the eight model simulations, respectively, and the thick lines in other colours show

765 the mean values of observation-based estimates. The time period over which observations and

766 model quantities are averaged is chosen to be the same. In Figure 11a, for aboveground biomass,

767 the GEOCARBON data set uses one product for the extratropics and another for the tropics to

768      create a global aboveground biomass product. The Zhang product (Zhang and Liang, 2020) is

769      based on the fusion of multiple gridded biomass datasets for generating a global product. Both

770      products are described in detail in Seiler et al. (2022). The model results generally compare better

771      with the Zhang product outside the $10°N$ to $10°S$ region but with the GEOCARBON product within

772      this region. The values to the south of $40°S$ are generally less reliable because of the little

773      vegetated land area below this latitude. In Figure 11b, the model simulated values for soil organic

774      C compare better with the HWSD dataset compared to the SG250m data for reasons mentioned

775      in the previous paragraph. Simulated leaf area index (Figure 11c) and gross primary productivity

776      (Figure 11e) generally compare well their observation-based estimates. The simulated area

777      burned (Figure 11d) and fire emissions (Figure 11f) also compare well with observation-based

778      estimates except that the model is not able to capture the small area burned and emissions at

779      northern high latitudes between around $50°N$ to $70°N$. Figures A12 and A13 compare zonally

780      averaged values of other simulated quantities with observation-based estimates used in the

781      AMBER framework. Together Figures 11, A12, and A13 illustrate that the model is overall able

782      to capture the latitudinal distribution of most land surface quantities.

783

784

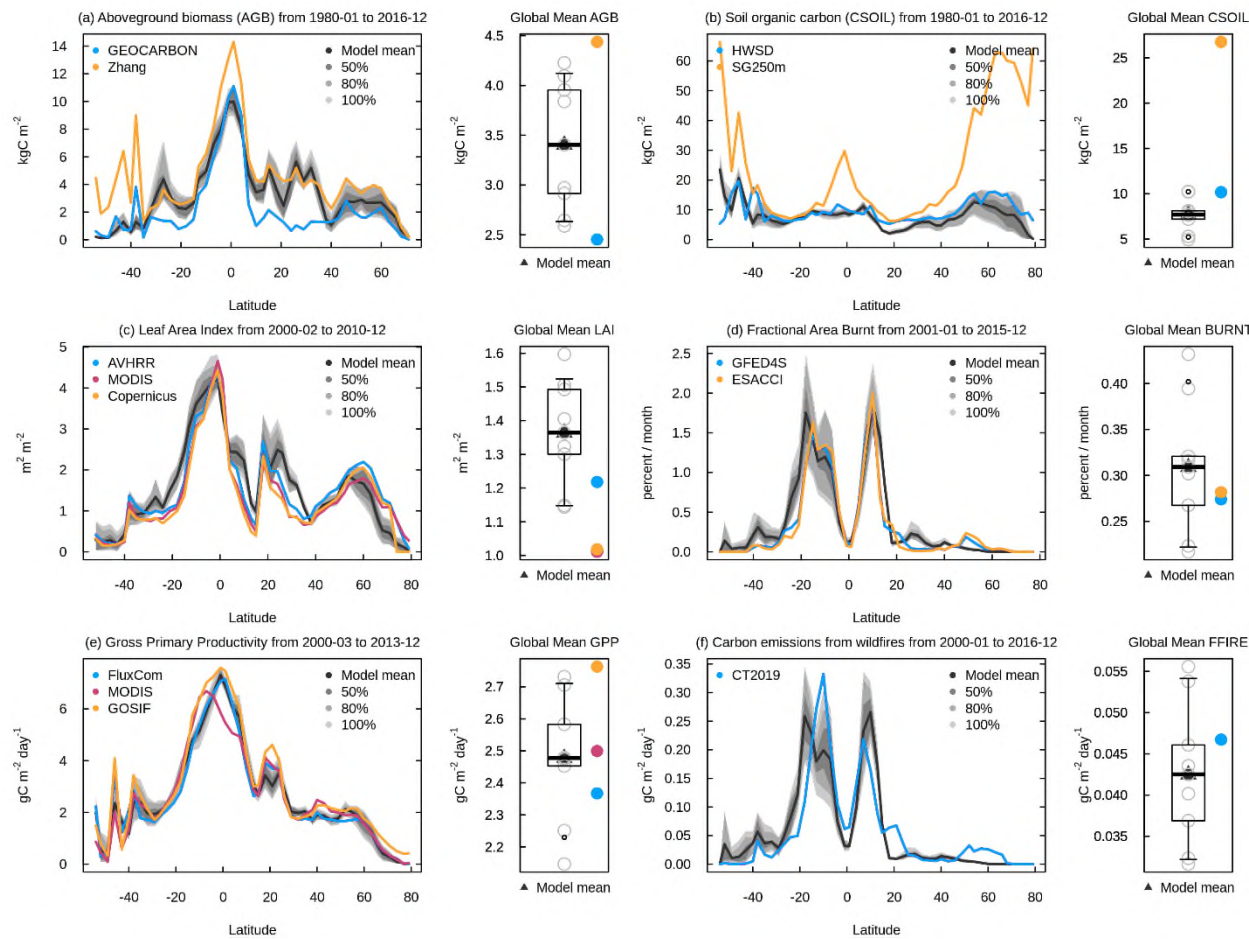
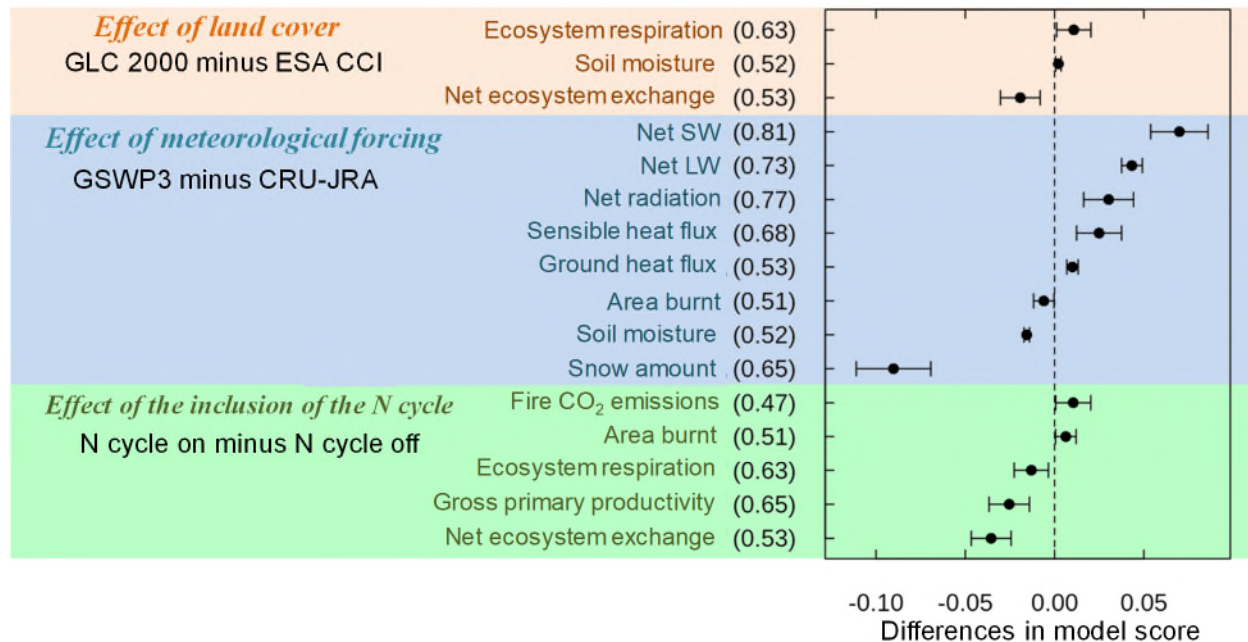



Figure 11: Zonally-averaged values of aboveground biomass (a), soil carbon mass (b), leaf area index (c), fractional area burnt (d), gross primary productivity (e), and fire $CO_2$ emissions (f) from the eight simulations summarized in Table 1. The model results are shown as their mean (black) and the 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.

791
792 Figure 12: Summary of difference in overall scores for model simulated quantities and
793 combinations for which the differences are statistically significant. The scores in parentheses for
794 each quantity are the average scores across the eight simulations and provide context. The error
795 bars denote the 95% confidence interval as explained in the text.

796

797 Since overall scores are available for all eight simulations for model quantities that are

798 compared to observations it is possible to evaluate how an interactive N cycle, and the choice of

799 meteorological data and land cover data affect model performance. Figure 12 summarizes the

800 difference in overall scores for model quantities and combinations for which the differences are

801 statistically significant at the 5% level based on Tukey's test (Tukey, 1977). The score indicated in

802 parentheses for each quantity is the average score across the eight simulations and provides

803 context. For example, when evaluating the effect of change in land cover for NEE the use of the

804 GLC 2000 based land cover, compared to the use of the ESA CCI based land cover, degrades the

805 average score for net ecosystem exchange by about 0.02 given that the average score for net

806    ecosystem exchange in 0.53. The error bars on the value 0.02 denote the 95% confidence interval

807    and in this case are calculated by differencing four simulations that use the GLC 2000 based land

808    cover versus four simulations that use the ESA CCI based land cover. The use of the GLC 2000

809    based land cover on the other hand slightly improves scores for ecosystem respiration and liquid

810    soil moisture. The use of GSWP3 data improves model scores for net shortwave, longwave, and

811    total radiation, for sensible and ground heat flux but degrades the overall score for area burned,

812    soil moisture, and more so for snow water equivalent. Finally, an interactive N cycle slightly

813    improves model performance for area burned and fire $CO_2$ emissions (due to improved

814    aboveground biomass in the tropics) but degrades it for ecosystem respiration, GPP, and net

815    ecosystem exchange. The inclusion of an interactive N cycle changes $V_{c,max}$ to a prognostic

816    variable for each PFT as opposed to being specified based on observations. This is analogous to

817    running an atmospheric model with a fully dynamic 3-dimensional ocean as opposed to using

818    specified sea surface temperatures (SST) and sea ice concentrations (SIC). Using a dynamic ocean

819    allows future projections (since future SSTs and SICs are not known) but invariably degrades a

820    model's performance for the present day since simulated SSTs and SICs will have their biases.

821    Similarly, using an interactive N cycle allows to project future changes in $V_{c,max}$ (based on changes

822    in N availability) but also degrades CLASSIC's performance for the present day since simulated

823    $V_{c,max}$ has its own biases. Overall, the model performance is most affected by the choice of the

824    driving meteorological data for water and energy fluxes, and by the inclusion or absence of an N

825    cycle and by the choice of land  cover for carbon-cycle related state variables and fluxes.

826

827

**5. Conclusions**

The response of the terrestrial biosphere over the historical period has been driven primarily by four global change drivers – increasing atmospheric $CO_2$, changing climate, LUC, and N deposition and fertilizer application. Our framework allows us to evaluate how a land surface model responds to increasing atmospheric $CO_2$, changing climate, and anthropogenic N additions to the coupled soil-vegetation system is dependent on two driving meteorological data sets, two land cover representations, and the two model variations (with and without an interactive N cycle). However, the framework used here does not quantify the uncertainty associated with LUC over the historical period since we use only one reconstruction of increasing crop area over the historical period. These results help draw three primary conclusions. First, even if the observations and models were perfect (including their structure and their parameterizations) the uncertainty associated with driving meteorological data and geophysical fields makes it difficult to evaluate LSMs. The uncertainty in global scale driving data 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 perfect results.

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 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

849     perfect model driven by perfect forcing data cannot be truly evaluated to its fullest extent since

850     observations themselves have uncertainties.

851        Third, with the caveat that our framework uses only one reconstruction of increase in

852     crop area over the historical period, the response of a model expressed in terms of net

853     atmosphere-land $CO_2$ flux to perturbation in meteorological, $CO_2$, and LUC forcing over the

854     historical period appears to be largely independent of its pre-industrial state as simulated here.

855     The pre-industrial soil and vegetation C mass for the eight simulations considered here vary

856     between 1035 ± 195 Pg C and 405 ± 58 Pg C (mean ± standard deviation), respectively. Both pre-

857     industrial and present-day vegetation and soil C pools explain only about 2% to 7% of the

858     variability in simulated net atmosphere-land $CO_2$ flux (Figure A11) over the 1997-2016 period of

859     each of the eight simulations. The net atmosphere-$CO_2$ flux from all eight simulations for the

860     period the 1960s to 2000s is found to lie within the uncertainty range provided by the GCP

861     (Friedlingstein et al., 2022). Given the current uncertainty in net atmosphere-land $CO_2$ flux, it is

862     therefore not possible to exclude any of the eight simulations at least on this basis. The finding

863     that a transient response of a model is independent of its pre-industrial state is also consistent

864     with land components of CMIP6 models. Arora et al. (2020) analyzed results from CMIP6

865     simulations in which atmospheric $CO_2$ increases at a rate of 1% per year from the year 1850 until

866     $CO_2$ quadruples from ~285 to ~1140 ppm. They found that the C-concentration and C-climate

867     feedback parameters for the land component of CMIP6 models do not depend on the absolute

868     values of their vegetation and soil C pools but rather how a given model responds to changes in

869     atmospheric $CO_2$ and the associated change in temperature. This conclusion is perhaps

870     somewhat comforting in that while pre-industrial states of LSMs may be different from their true

871 observed states they still have the ability to reproduce net atmosphere-land $CO_2$ flux over the

872 historical period that is consistent with current observation-based estimates. Clearly, this

873 reasoning does not apply if pre-industrial vegetation or soil C mass are zero. One reason why

874 present day net atmosphere-land $CO_2$ flux is independent of a LSM's pre-industrial state is

875 because the model is first spun up to equilibrium conditions and then forced with time-variant

876 forcings. However, successful reproduction of atmosphere-land $CO_2$ fluxes over the historical

877 period is no guarantee that future projections from LSMs are reliable.

878    The ensemble-based approach used here also allows for the evaluation of the effect of a

879 given meteorological forcing and land cover, and the effect of an interactive N cycle on model

880 simulated quantities in a robust manner. Ensemble averages of simulations that use the CRU-JRA

881 and GSWP3 meteorological forcing show that the use of the GSWP3 meteorological forcing yields

882 lower evapotranspiration (latent heat flux), higher runoff, higher sensible heat flux, higher

883 burned area, and a weaker land C sink for the present day compared to when the CRU-JRA

884 meteorological forcing is used. The use of the ESA CCI land cover leads to higher soil C, higher

885 GPP, and higher area burned primarily because of the larger grass area when land cover is based

886 on the ESA CCI product compared to the GLC 2000 product. The use of the ESA CCI based land

887 cover also leads to a slightly weaker land C sink for the present day. Finally, the comparison of

888 simulations with and without the N cycle averaged over all meteorological data and land cover

889 combinations allows us to identify the effect of the N cycle. Simulated vegetation C mass and GPP

890 are lower in the model version with the interactive N cycle. In particular, we found that the

891 somewhat low productivity at high latitudes, when the N cycle is turned on, leads to relatively

892 large differences in soil C at high latitudes regardless of the meteorological data or land cover

893    being used to drive the model. Although, this is not the reason for differences in net atmosphere-

894    land $CO_2$ flux between models with and without N cycling: as mentioned above present-day net

895    atmosphere-land $CO_2$ flux is independent of both the pre-industrial and present-day vegetation

896    and soil C pools. Given the knowledge about the effect of N cycling on model behaviour, the

897    reasons can now be investigated to further improve the N cycle component of CLASSIC.

898        It is logical to assume that the results presented here are sensitive to the horizontal

899    resolution of the model. Both forcing data that are used to drive the model, and observations

900    against which model results are compared, are regridded to be consistent with the model's

901    spatial resolution. For example, at the scale of a few meters, meteorological variables measured

902    at a given site will indeed be less uncertain than their spatially-averaged values say for a 2.81°

903    grid cell. Similarly, observations at a scale of a few meters for soil C and/or vegetation C mass will

904    also likely be more certain than their values at large spatial scales. This is one reason why AMBER

905    uses both gridded and in-situ observation-based estimates to calculate its scores. Fluxes of latent

906    and sensible heat, on the other hand, may not be any more certain at a given site than over large

907    spatial scales. This is because of the problems associated with energy budget closure (Mauder et

908    al., 2020) which, at the point scale, prevent the sum of annual latent and sensible heat flux to be

909    equal to net radiation (average of ground heat fluxes is close to zero at an annual time scale).

910        LSMs have become increasingly complex over the years and so has the requirement for

911    forcing data to drive these models. The evaluation of LSMs has also become complex as the

912    models now generate a multitude of variables that must be evaluated against their observation-

913    based estimates. Estimates of observation-based data to evaluate models, and the availability of

914    forcing data, have also increased. Given the uncertainties associated with model inputs, model

915     structure, and observation-based data, it is unrealistic to expect LSMs to perfectly reproduce

916     observations for large-scale global simulations. It is not known *a priori* which model structure,

917     forcing data sets, and observation data sets are better. Driving data including meteorological data

918     sets and land cover representations may be more realistic in some parts of world and less in

919     others. Observation-based data sets also have their limitations and attributes which may make

920     them better or ill-suited for comparison with a given model.  A more robust model evaluation

921     must therefore take into account the uncertainties both in the forcing and observation-based

922     data. A comprehensive and robust model evaluation can be performed by comparing multiple

923     model realizations against multiple observation-based data sets.

924

## References

925 **References**

926 Agustí-Panareda, A., Diamantakis, M., Massart, S., Chevallier, F., Muñoz-Sabater, J., Barré, J., Curcoll, R.,
927 Engelen, R., Langerock, B., Law, R. M., Loh, Z., Morguí, J. A., Parrington, M., Peuch, V.-H., Ramonet, M.,
928 Roehl, C., Vermeulen, A. T., Warneke, T., and Wunch, D.: Modelling $CO_2$ weather – why horizontal
929 resolution matters, Atmos Chem Phys, 19, 7347–7376, 2019.

930 Arora, V. K. and Boer, G. J.: A parameterization of leaf phenology for the terrestrial ecosystem
931 component of climate models, Glob. Change Biol., 11, 39–59, https://doi.org/10.1111/j.1365-
932 2486.2004.00890.x, 2005.

933 Arora, V. K. and Boer, G. J.: Simulating Competition and Coexistence between Plant Functional Types in a
934 Dynamic Vegetation Model, Earth Interact., 10, 1–30, 2006.

935 Arora, V. K. and Melton, J. R.: Reduction in global area burned and wildfire emissions since 1930s
936 enhances carbon uptake by land, Nat. Commun., 9, 1326, https://doi.org/10.1038/s41467-018-03838-0,
937 2018.

938 Arora, V. K., Boer, G. J., Christian, J. R., Curry, C. L., Denman, K. L., Zahariev, K., Flato, G. M., Scinocca, J.
939 F., Merryfield, W. J., and Lee, W. G.: The Effect of Terrestrial Photosynthesis Down Regulation on the
940 Twentieth-Century Carbon Budget Simulated with the CCCma Earth System Model, J. Clim., 22, 6066–
941 6088, https://doi.org/10.1175/2009JCLI3037.1, 2009.

942 Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W.
943 G., and Merryfield, W. J.: Carbon emission limits required to satisfy future representative concentration
944 pathways of greenhouse gases, Geophys. Res. Lett., 38, https://doi.org/10.1029/2010GL046270, 2011.

945 Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp,
946 L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina,
947 T., Joetzjer, E., Kawamiya, M., Koven, C. D., Krasting, J. P., Law, R. M., Lawrence, D. M., Lenton, A.,
948 Lindsay, K., Pongratz, J., Raddatz, T., Séférian, R., Tachiiri, K., Tjiputra, J. F., Wiltshire, A., Wu, T., and
949 Ziehn, T.: Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison
950 to CMIP5 models, Biogeosciences, 17, 4173–4222, https://doi.org/10.5194/bg-17-4173-2020, 2020.

951 Asaadi, A. and Arora, V. K.: Implementation of nitrogen cycle in the CLASSIC land model, Biogeosciences,
952 18, 669–706, https://doi.org/10.5194/bg-18-669-2021, 2021.

953 Avitabile, V., Herold, M., Heuvelink, G. B. M., and others: An integrated pan tropical biomass map using
954 multiple reference datasets, Glob Chang Biol, 2016.

955 Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction,
956 Hydrol. Process., 6, 279–298, https://doi.org/10.1002/hyp.3360060305, 1992.

957 Bonan, G. B. and Doney, S. C.: Climate, ecosystems, and planetary futures: The challenge to predict life
958 in Earth system models, Science, 359, eaam8328, https://doi.org/10.1126/science.aam8328, 2018.

959 Bonan, G. B., Lombardozzi, D. L., Wieder, W. R., Oleson, K. W., Lawrence, D. M., Hoffman, F. M., and
960 Collier, N.: Model Structure and Climate Data Uncertainty in Historical Simulations of the Terrestrial

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

963    Booth, B. B. B., Jones, C. D., Collins, M., Totterdell, I. J., Cox, P. M., Sitch, S., Huntingford, C., Betts, R. A.,
964    Harris, G. R., and Lloyd, J.: High sensitivity of future global warming to land carbon cycle processes,
965    Environ. Res. Lett., 7, 024002, https://doi.org/10.1088/1748-9326/7/2/024002, 2012.

966    Chuvieco, E., Lizundia-Loiola, J., Pettinari, M. L., Ramo, R., Padilla, M., Tansey, K., Mouillot, F., Laurent,
967    P., Storm, T., Heil, A., and Others: Generation and analysis of a new global burned area product based on
968    MODIS 250 m reflectance bands and thermal anomalies, Earth Syst. Sci. Data, 10, 2015–2031, 2018.

969    Claverie, M., Matthews, J. L., Vermote, E. F., and Justice, C. O.: A 30+ Year AVHRR LAI and FAPAR Climate
970    Data Record: Algorithm Description and Validation, Remote Sens., 8, 263, 2016.

971    Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and
972    Randerson, J. T.: The International Land Model Benchmarking (ILAMB) System: Design, Theory, and
973    Implementation, J. Adv. Model. Earth Syst., 10, 2731–2754, https://doi.org/10.1029/2018MS001354,
974    2018.

975    Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R.
976    S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman,
977    P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T.
978    F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The Twentieth Century Reanalysis Project,
979    Q. J. R. Meteorol. Soc., 137, 1–28, https://doi.org/10.1002/qj.776, 2011.

980    Dai, A. and Trenberth, K. E.: Estimates of Freshwater Discharge from Continents: Latitudinal and
981    Seasonal Variations, J Hydrometeorol, 3, 660–687, 2002.

982    Di Vittorio, A. V., Chini, L. P., Bond-Lamberty, B., Mao, J., Shi, X., Truesdale, J., Craig, A., Calvin, K., Jones,
983    A., Collins, W. D., Edmonds, J., Hurtt, G. C., Thornton, P., and Thomson, A.: From land use to land cover:
984    restoring the afforestation signal in a coupled integrated assessment–earth system model and the
985    implications for CMIP5 RCP simulations, Biogeosciences, 11, 6435–6450, https://doi.org/10.5194/bg-11-
986    6435-2014, 2014.

987    Di Vittorio, A. V., Mao, J., Shi, X., Chini, L., Hurtt, G., and Collins, W. D.: Quantifying the Effects of
988    Historical Land Cover Conversion Uncertainty on Global Carbon and Climate Estimates, Geophys. Res.
989    Lett., 45, 974–982, https://doi.org/10.1002/2017GL075124, 2018.

990    ESA: Land Cover CCI Product User Guide Version 2 Technical Report, European Space Agency. Available
991    at http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf, 2017.

992    Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of
993    the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization,
994    Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

995    Fischer, G., Nachtergaele, F., Prieler, S., van Velthuizen, H. T., Verelst, L., and Wiberg, D.: Global Agro-
996    ecological Zones Assessment for Agriculture (GAEZ 2008), IIASA and FAO, Laxenburg, Austria and Rome,
997    Italy, 2008.

998      Fisher, R. A. and Koven, C. D.: Perspectives on the Future of Land Surface Models and the Challenges of
999      Representing Complex Terrestrial Systems, J Adv Model Earth Syst, 12, 2020.

1000    Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Hauck, J., Peters, G. P., Peters, W.,
1001    Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D. C. E., Canadell, J. G., Ciais, P., Jackson, R. B., Anthoni, P.,
1002    Barbero, L., Bastos, A., Bastrikov, V., Becker, M., Bopp, L., Buitenhuis, E., Chandra, N., Chevallier, F.,
1003    Chini, L. P., Currie, K. I., Feely, R. A., Gehlen, M., Gilfillan, D., Gkritzalis, T., Goll, D. S., Gruber, N.,
1004    Gutekunst, S., Harris, I., Haverd, V., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Joetzjer, E., Kaplan, J.
1005    O., Kato, E., Klein Goldewijk, K., Korsbakken, J. I., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lenton, A.,
1006    Lienert, S., Lombardozzi, D., Marland, G., McGuire, P. C., Melton, J. R., Metzl, N., Munro, D. R., Nabel, J.
1007    E. M. S., Nakaoka, S.-I., Neill, C., Omar, A. M., Ono, T., Peregon, A., Pierrot, D., Poulter, B., Rehder, G.,
1008    Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P. P., Tian, H.,
1009    Tilbrook, B., Tubiello, F. N., van der Werf, G. R., Wiltshire, A. J., and Zaehle, S.: Global Carbon Budget
1010    2019, Earth Syst. Sci. Data, 11, 1783–1838, https://doi.org/10.5194/essd-11-1783-2019, 2019.

1011    Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., Le Quéré, C.,
1012    Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni,
1013    P., Bates, N. R., Becker, M., Bellouin, N., Bopp, L., Chau, T. T. T., Chevallier, F., Chini, L. P., Cronin, M.,
1014    Currie, K. I., Decharme, B., Djeutchouang, L. M., Dou, X., Evans, W., Feely, R. A., Feng, L., Gasser, T.,
1015    Gilfillan, D., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Houghton, R. A., Hurtt,
1016    G. C., Iida, Y., Ilyina, T., Luijkx, I. T., Jain, A., Jones, S. D., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer,
1017    J., Korsbakken, J. I., Körtzinger, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lienert, S., Liu, J.,
1018    Marland, G., McGuire, P. C., Melton, J. R., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Niwa, Y., Ono,
1019    T., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M.,
1020    Schwinger, J., Schwingshackl, C., Séférian, R., Sutton, A. J., Sweeney, C., Tanhua, T., Tans, P. P., Tian, H.,
1021    Tilbrook, B., Tubiello, F., van der Werf, G. R., Vuichard, N., Wada, C., Wanninkhof, R., Watson, A. J.,
1022    Willis, D., Wiltshire, A. J., Yuan, W., Yue, C., Yue, X., Zaehle, S., and Zeng, J.: Global Carbon Budget 2021,
1023    Earth Syst. Sci. Data, 14, 1917–2005, https://doi.org/10.5194/essd-14-1917-2022, 2022.

1024    Garrigues, S., Lacaze, R., Baret, F., Morisette, J. T., Weiss, M., Nickeson, J. E., Fernandes, R., Plummer, S.,
1025    Shabanov, N. V., Myneni, R. B., and Others: Validation and intercomparison of global Leaf Area Index
1026    products derived from remote sensing data, J. Geophys. Res. Biogeosciences, 113, 2008.

1027    Giglio, L., Randerson, J. T., van der Werf, G. R., Kasibhatla, P. S., Collatz, G. J., Morton, D. C., and DeFries,
1028    R. S.: Assessing variability and long-term trends in burned area by merging multiple satellite fire
1029    products, 7, 1171–1186, 2010.

1030    Giglio, L., Randerson, J. T., and van der Werf, G. R.: Analysis of daily, monthly, and annual burned area
1031    using the fourth-generation global fire emissions database (GFED4), J. Geophys. Res. Biogeosciences,
1032    118, 317–328, https://doi.org/10.1002/jgrg.20042, 2013.

1033    Harris, I. C.: CRU JRA v2.1: A forcings dataset of gridded land surface blend of Climatic Research Unit
1034    (CRU) and Japanese reanalysis (JRA) data; Jan. 1901 - Dec. 2019, Centre for Environmental Data Analysis,
1035    University of East Anglia Climatic Research Unit,
1036    https://catalogue.ceda.ac.uk/uuid/10d2c73e5a7d46f4ada08b0a26302ef7, 2020.

1037    Hegglin, M., Kinnison, D., and Lamarque, J.-F.: Wet and dry NHx and NOy deposition data,
1038    input4MIPs.CMIP6.CMIP.NCAR. Version 2016-11-15, Earth System Grid Federation,
1039    https://doi.org/10.22033/ESGF/input4MIPs.10448, 2016.

1040 Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A.,
1041 Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan,
1042 R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m:
1043 Global gridded soil information based on machine learning, PLOS ONE, 12, 1–40,
1044 https://doi.org/10.1371/journal.pone.0169748, 2017.

1045 Hobeichi, S., Abramowitz, G., and Evans, J.: Conserving Land-Atmosphere Synthesis Suite (CLASS), J Clim,
1046 2019.

1047 Hornberger, G. M. and Spear, R. C.: Approach to the preliminary analysis of environmental systems, J
1048 Env. Manage U. S., 12:1, 1981.

1049 van den Hurk, B., Kim, H., Krinner, G., Seneviratne, S. I., Derksen, C., Oki, T., Douville, H., Colin, J.,
1050 Ducharne, A., Cheruy, F., Viovy, N., Puma, M. J., Wada, Y., Li, W., Jia, B., Alessandri, A., Lawrence, D. M.,
1051 Weedon, G. P., Ellis, R., Hagemann, S., Mao, J., Flanner, M. G., Zampieri, M., Materia, S., Law, R. M., and
1052 Sheffield, J.: LS3MIP (v1.0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model
1053 Intercomparison Project – aims, setup and expected outcome, Geosci. Model Dev., 9, 2809–2832,
1054 https://doi.org/10.5194/gmd-9-2809-2016, 2016.

1055 Hurtt, G. C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B. L., Calvin, K., Doelman, J. C., Fisk, J., Fujimori,
1056 S., Klein Goldewijk, K., Hasegawa, T., Havlik, P., Heinimann, A., Humpenöder, F., Jungclaus, J., Kaplan, J.
1057 O., Kennedy, J., Krisztin, T., Lawrence, D., Lawrence, P., Ma, L., Mertz, O., Pongratz, J., Popp, A., Poulter,
1058 B., Riahi, K., Shevliakova, E., Stehfest, E., Thornton, P., Tubiello, F. N., van Vuuren, D. P., and Zhang, X.:
1059 Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6,
1060 Geosci Model Dev, 13, 5425–5464, 2020a.

1061 Hurtt, G. C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B. L., Calvin, K., Doelman, J. C., Fisk, J., Fujimori
1062 S., Klein Goldewijk, K., Hasegawa, T., Havlik, P., Heinimann, A., Humpenöder, F., Jungclaus, J., Kaplan, J.
1063 O., Kennedy, J., Krisztin, T., Lawrence, D., Lawrence, P., Ma, L., Mertz, O., Pongratz, J., Popp, A., Poulter,
1064 B., Riahi, K., Shevliakova, E., Stehfest, E., Thornton, P., Tubiello, F. N., van Vuuren, D. P., and Zhang, X.:
1065 Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6,
1066 Geosci. Model Dev., 13, 5425–5464, https://doi.org/10.5194/gmd-13-5425-2020, 2020b.

1067 Jacobson, A. R., Schuldt, K. N., Miller, J. B., and Oda, T.: CarbonTracker Documentation CT2019 release,
1068 2020.

1069 Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana,
1070 G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, Sci Data, 6, 74,
1071 2019.

1072 Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S.,
1073 Bodesheim, P., Carvalhais, N., and others: Scaling carbon fluxes from eddy covariance sites to globe:
1074 synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 17, 1343–1365, 2020.

1075 Kato, S., Loeb, N. G., Rose, F. G., Doelling, D. R., Rutan, D. A., Caldwell, T. E., Yu, L., and Weller, R. A.:
1076 Surface Irradiances Consistent with CERES-Derived Top-of-Atmosphere Shortwave and Longwave
1077 Irradiances, J Clim, 26, 2719–2740, 2013.

1078  Kou-Giesbrecht, S. and Arora, V. K.: Representing the Dynamic Response of Vegetation to Nitrogen
1079  Limitation via Biological Nitrogen Fixation in the CLASSIC Land Model, Glob. Biogeochem. Cycles, 36,
1080  e2022GB007341, https://doi.org/10.1029/2022GB007341, 2022.

1081  Kyker-Snowman, E., Lombardozzi, D. L., Bonan, G. B., Cheng, S. J., Dukes, J. S., Frey, S. D., Jacobs, E. M.,
1082  McNellis, R., Rady, J. M., Smith, N. G., Thomas, R. Q., Wieder, W. R., and Grandy, A. S.: Increasing the
1083  spatial and temporal impact of ecological research: A roadmap for integrating a novel terrestrial process
1084  into an Earth system model, Glob. Change Biol., 28, 665–684, https://doi.org/10.1111/gcb.15894, 2022.

1085  Lawrence, P. J. and Chase, T. N.: Representing a new MODIS consistent land surface in the Community
1086  Land Model (CLM 3.0), J. Geophys. Res. Biogeosciences, 112, https://doi.org/10.1029/2006JG000168,
1087  2007.

1088  Li, J., Duan, Q., Wang, Y.-P., Gong, W., Gan, Y., and Wang, C.: Parameter optimization for carbon and
1089  water fluxes in two global land surface models based on surrogate modelling, Int. J. Climatol., 38,
1090  e1016–e1031, https://doi.org/10.1002/joc.5428, 2018a.

1091  Li, W., MacBean, N., Ciais, P., Defourny, P., Lamarche, C., Bontemps, S., Houghton, R. A., and Peng, S.:
1092  Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI
1093  land cover maps (1992–2015), Earth Syst. Sci. Data, 10, 219–234, https://doi.org/10.5194/essd-10-219-
1094  2018, 2018b.

1095  Li, X. and Xiao, J.: Mapping Photosynthesis Solely from Solar-Induced Chlorophyll Fluorescence: A Global,
1096  Fine-Resolution Dataset of Gross Primary Production Derived from OCO-2, Remote Sens., 11, 2563,
1097  2019.

1098  Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., Van Dijk, A., McCabe, M. F.,
1099  Evans, J., and Others: Developing an improved soil moisture dataset by blending passive and active
1100  microwave satellite-based retrievals, 2011.

1101  Lu, C. and Tian, H.: Global nitrogen and phosphorus fertilizer use for agriculture production in the past
1102  half century: shifted hot spots and nutrient imbalance, Earth Syst. Sci. Data, 9, 181–192,
1103  https://doi.org/10.5194/essd-9-181-2017, 2017.

1104  Mauder, M., Foken, T., and Cuxart, J.: Surface-Energy-Balance Closure over Land: A Review, Bound.-
1105  Layer Meteorol., 177, 395–426, https://doi.org/10.1007/s10546-020-00529-6, 2020.

1106  Meiyappan, P. and Jain, A. K.: Three distinct global estimates of historical land-cover change and land-
1107  use conversions for over 200 years, Front. Earth Sci., 6, 122–139, https://doi.org/10.1007/s11707-012-
1108  0314-2, 2012.

1109  Melton, J. R. and Arora, V. K.: Competition between plant functional types in the Canadian Terrestrial
1110  Ecosystem Model (CTEM) v. 2.0, Geosci. Model Dev., 9, 323–361, 2016a.

1111  Melton, J. R. and Arora, V. K.: Competition between plant functional types in the Canadian Terrestrial
1112  Ecosystem Model (CTEM) v. 2.0, Geosci Model Dev, 9, 323–361, https://doi.org/10.5194/gmd-9-323-
1113  2016, 2016b.

1114    Melton, J. R., Arora, V. K., Wisernig-Cojoc, E., Seiler, C., Fortier, M., Chan, E., and Teckentrup, L.: CLASSIC
1115    v1.0: the open-source community successor to the Canadian Land Surface Scheme (CLASS) and the
1116    Canadian Terrestrial Ecosystem Model (CTEM) – Part 1: Model framework and site-level performance,
1117    Geosci. Model Dev. Discuss., 2019, 1–40, https://doi.org/10.5194/gmd-2019-329, 2019.

1118    Mortimer, C., Mudryk, L., Derksen, C., Luojus, K., Brown, R., Kelly, R., and Tedesco, M.: Evaluation of
1119    long-term Northern Hemisphere snow water equivalent products, The Cryosphere, 14, 1579–1594,
1120    2020.

1121    Mudryk, L.: Historical gridded snow water equivalent and snow cover fraction over Canada from remote
1122    sensing and land surface models, available at http://climate-scenarios.canada.ca/?page=blended-snow-
1123    data (last accessed Oct 2022), 2020.

1124    Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y.,
1125    Smith, G. R., Lotsch, A., Friedl, M., Morisette, J. T., Votava, P., Nemani, R. R., and Running, S. W.: Global
1126    products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sens
1127    Env., 83, 214–231, 2002.

1128    Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J.,
1129    Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C., Zhang, L., Amiro, B.,
1130    Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M.,
1131    Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C.,
1132    Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D.,
1133    Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns,
1134    S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti,
1135    A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha,
1136    H., Dai, X., Davis, K. J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R. C., Delpierre, N., Desai,
1137    A. R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E.,
1138    Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B.,
1139    Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno,
1140    M., Gharun, M., Gianelle, D., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for
1141    eddy covariance data, Sci Data, 7, 225, 2020.

1142    Peng, S., Ciais, P., Maignan, F., Li, W., Chang, J., Wang, T., and Yue, C.: Sensitivity of land use change
1143    emission estimates to historical land use and land cover mapping, Glob. Biogeochem. Cycles, 31, 626–
1144    643, https://doi.org/10.1002/2015GB005360, 2017.

1145    Poulter, B., Hattermann, F., Hawkins, E., Zaehle, S., Sitch, S., Restrepo-Coupe, N., Heyder, U., and
1146    Cramer, W.: Robust dynamics of Amazon dieback to climate change with perturbed ecosystem model
1147    parameters, Glob. Change Biol., 16, 2476–2495, https://doi.org/10.1111/j.1365-2486.2009.02157.x,
1148    2010.

1149    Reusch, A. and Gibbs, H. K.: New IPCC Tier-1 Global Biomass Carbon Map For the Year 2000, Oak Ridge
1150    National Laboratory, Oak Ridge, Tennessee, 2008.

1151    Rödenbeck, C., Zaehle, S., Keeling, R., and Heimann, M.: How does the terrestrial carbon exchange
1152    respond to inter-annual climatic variations? A quantification based on atmospheric $CO_2$ data,
1153    Biogeosciences, 15, 2481–2498, 2018.

Santoro, M., Beaudoin, A., Beer, C., Cartus, O., Fransson, J. E. S., Hall, R. J., Pathe, C., Schmullius, C., Schepaschenko, D., Shvidenko, A., Thurner, M., and Wegmüller, U.: Forest growing stock volume of the northern hemisphere: Spatially explicit estimates for 2010 derived from Envisat ASAR, Remote Sens Env., 168, 316–334, 2015.

Schepaschenko, D., Chave, J., Phillips, O. L., Lewis, S. L., Davies, S. J., Réjou-Méchain, M., Sist, P., Scipal, K., Perger, C., Herault, B., Labrière, N., Hofhansl, F., Affum-Baffoe, K., Aleinikov, A., Alonso, A., Amani, C., Araujo-Murakami, A., Armston, J., Arroyo, L., Ascarrunz, N., Azevedo, C., Baker, T., Ba\lazy, R., Bedeau, C., Berry, N., Bilous, A. M., Bilous, S. Y., Bissiengou, P., Blanc, L., Bobkova, K. S., Braslavskaya, T., Brienen, R., Burslem, D. F. R. P., Condit, R., Cuni-Sanchez, A., Danilina, D., Del Castillo Torres, D., Derroire, G., 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., Gothard-Bassébé, 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., Koffi, J.-C. K., Konovalova, M., Kraxner, F., Krejza, J., Krisnawati, H., Krivobokov, L. V., Kuznetsov, M. A., Lakyda, I., Lakyda, P. I., Licona, J. C., Lucas, R. M., Lukina, N., Lussetti, D., Malhi, Y., Manzanera, J. A., Marimon, B., Junior, B. H. M., Martinez, R. V., Martynenko, O. V., Matsala, M., Matyashuk, R. K., Mazzei, L., Memiaghe, H., Mendoza, C., Mendoza, A. M., Moroziuk, O. V., Mukhortova, L., Musa, S., Nazimova, D. I., Okuda, T., Oliveira, L. C., Ontikov, P. V., et al.: The Forest Observation System, building a global 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) – 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-14-2371-2021, 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, W., and Zaehle, S.: Are Terrestrial Biosphere Models Fit for Simulating the Global Land Carbon Sink?, J. 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, 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.

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

1193    Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V.,
1194    Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A.,
1195    Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model
1196    version 5 (CanESM5.0.3), Geosci. Model Dev., 12, 4823–4873, https://doi.org/10.5194/gmd-12-4823-
1197    2019, 2019.

1198    Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections,
1199    Philos. Trans. R. Soc. Math. Phys. Eng. Sci., 365, 2053–2075, https://doi.org/10.1098/rsta.2007.2076,
1200    2007.

1201    Tian, Y., Dickinson, R. E., Zhou, L., and Shaikh, M.: Impact of new land boundary conditions from
1202    Moderate Resolution Imaging Spectroradiometer (MODIS) data on the climatology of land surface
1203    variables, J. Geophys. Res. Atmospheres, 109, https://doi.org/10.1029/2003JD004499, 2004.

1204    Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., and
1205    Allison, S. D.: Causes of variation in soil carbon simulations from CMIP5 Earth system models and
1206    comparison with observations, 10, 1717–1736, 2013.

1207    Tukey, J. W.: Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.

1208    Verger, A., Baret, F., and Weiss, M.: Near real-time vegetation monitoring at global scale, IEEE J. Sel. Top.
1209    In, 2014.

1210    Verseghy, D. L.: Class—A Canadian land surface scheme for GCMS. I. Soil model, Int. J. Climatol., 11,
1211    111–133, https://doi.org/10.1002/joc.3370110202, 1991.

1212    Verseghy, D. L., McFarlane, N. A., and Lazare, M.: Class—A Canadian land surface scheme for GCMS, II.
1213    Vegetation model and coupled runs, Int. J. Climatol., 13, 347–370,
1214    https://doi.org/10.1002/joc.3370130402, 1993.

1215    Wang, A., Price, D. T., and Arora, V.: Estimating changes in global vegetation cover (1850–2100) for use
1216    in climate models, Glob. Biogeochem. Cycles, 20, https://doi.org/10.1029/2005GB002514, 2006.

1217    Wang, L., Bartlett, P., Chan, E., and Xiao, M.: Mapping of Plant Functional Type from Satellite-Derived
1218    Land Cover Datasets for Climate Models, in: IGARSS 2018 - 2018 IEEE International Geoscience and
1219    Remote Sensing Symposium, 3416–3419, https://doi.org/10.1109/IGARSS.2018.8518046, 2018.

1220    Wang, L., Bartlett, P., Pouliot, D., Chan, E., Lamarche, C., Wulder, M. A., Defourny, P., and Brady, M.:
1221    Comparison and Assessment of Regional and Global Land Cover Datasets for Use in CLASS over Canada,
1222    Remote Sens., 11, 2286, https://doi.org/10.3390/rs11192286, 2019.

1223    Wang, L., Arora, V. K., Bartlett, P., Chan, E., and Curasi, S. R.: Mapping of ESA-CCI land cover data to plant
1224    functional types for use in the CLASSIC land model, EGUsphere, 2022, 1–43,
1225    https://doi.org/10.5194/egusphere-2022-923, 2022.

1226    Wieder, W.: Regridded Harmonized World Soil Database v1.2, ,
1227    https://doi.org/10.3334/ORNLDAAC/1247, 2014.

1228 Wu, Z., Ahlström, A., Smith, B., Ardö, J., Eklundh, L., Fensholt, R., and Lehsten, V.: Climate data induced
1229 uncertainty in model-based estimations of terrestrial primary productivity, Environ. Res. Lett., 12,
1230 064013, https://doi.org/10.1088/1748-9326/aa6fd8, 2017.

1231 Xue, B.-L., Guo, Q., Hu, T., Wang, G., Wang, Y., Tao, S., Su, Y., Liu, J., and Zhao, X.: Evaluation of modeled
1232 global vegetation carbon dynamics: Analysis based on global carbon flux and above-ground biomass
1233 data, Ecol Modell, 355, 84–96, 2017.

1234 Zhang, Y. and Liang, S.: Fusion of Multiple Gridded Biomass Datasets for Generating a Global Forest
1235 Aboveground Biomass Map, Remote Sens., 12, 2559, 2020.

1236 Zhang, Y., Xiao, X., Wu, X., Zhou, S., Zhang, G., Qin, Y., and Dong, J.: A global moderate resolution dataset
1237 of gross primary production of vegetation for 2000–2016, Sci. Data, 4, 170165, 2017.

1238

**Code/data availability**

1241 More information about the CLASSIC land surface model and its Fortran code are available at
1242 https://cccma.gitlab.io/classic_pages/.

1244 AMBER source code as well as the scripts required for reproducing the computational environment,
1245 including all dependencies on other R-packages, can be found at
1246 https://doi.org/10.5281/zenodo.5670387.

1248 The full suite of results from AMBER for the eight simulations presented in this study can be
1249 found at https://cseiler.shinyapps.io/ShinyCLASSIC/.

**Author contribution**

1253 VA and SKG performed the simulations, and VA wrote the majority of the manuscript. CS performed the
1254 AMBER related analysis. LW put together the ESA CCI land cover. CS, LW, and SKG provided comments
1255 on the entire manuscript and also wrote their respective sections.

**Competing interests**

1259 There are no competing interests.

**Acknowledgment**

1263 We thank Joe Melton for providing comments on an earlier version of this paper. We also thank
1264 Benjamin Bond-Lamberty for taking this paper on as an Associate Editor, and the two anonymous
1265 reviewers for providing helpful comments which greatly improved this paper.

# Appendix

1266

1267
1268  **A1: Automated Model Benchmarking R Package (AMBER)**

1269

1270  The Automated Model Benchmarking R package quantifies model performance using five scores

1271  that assess a model's bias ($S_{bias}$), root-mean-square-error ($S_{rmse}$), seasonality ($S_{phase}$), inter-annual

1272  variability ($S_{iav}$), and spatial distribution ($S_{dist}$). All scores are dimensionless and range from zero

1273  to one, where increasing values imply better performance. The exact definition of each skill score

1274  is provided below.

1275  **A1.1 Bias Score ($S_{bias}$)**

1276  The bias is defined as the difference between the time-mean values of model and reference data:

1277  $$bias(\lambda, \phi) = \overline{v_{mod}}(\lambda, \phi) - \overline{v_{ref}}(\lambda, \phi), \tag{A1}$$

1278  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

1279  longitude $\lambda$ and latitude $\phi$ for model and reference data, respectively. Nondimensionalization is

1280  achieved by dividing the bias by the standard deviation of the reference data ($\sigma_{ref}$):

1281  $$\varepsilon_{bias}(\lambda, \phi) = \frac{|bias(\lambda,\phi)|}{\sigma_{ref}(\lambda,\phi)} \tag{A2}$$

1282  Note that $\varepsilon_{bias}$ is always positive, as it uses the absolute value of the bias. For evaluations against

1283  stream flow measurements, the bias is divided by the annual mean rather than the standard

1284  deviation of the reference data. This is because we assess streamflow on an annual rather than

1285  monthly basis, implying that the corresponding standard deviation is small. The same approach

1286  is applied to soil C and vegetation C mass, whose reference data provide a static snapshot in time.

1287  For both of these cases, $\varepsilon_{bias}(\lambda, \phi)$ becomes:

1288
$$\varepsilon_{bias}(\lambda, \phi) = \frac{|bias(\lambda, \phi)|}{\overline{v_{ref}}(\lambda, \phi)} \qquad \text{(A3)}$$

1289

1290  A bias score that ranges from zero to one is calculated next:

1291
$$s_{bias}(\lambda, \phi) = e^{-\varepsilon_{bias}(\lambda, \phi)} \qquad \text{(A4)}$$

1292  While small relative errors yield score values close to one, large relative errors cause score values

1293  to approach zero. Taking the mean of $s_{bias}$ across all latitudes and longitudes, denoted by a double

1294  bar over a variable, leads to the scalar score:

1295
$$S_{bias} = \overline{\overline{s_{bias}(\lambda, \phi)}} \qquad \text{(A5)}$$

1296

1297  **A1.2 Root-Mean-Square-Error Score ($S_{rmse}$)**

1298  While the bias assesses the difference between time-mean values, the root-mean-square-error

1299  (*rmse*) is concerned with the residuals of the modeled and observed time series:

1300
$$rmse(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left( v_{mod}(t, \lambda, \phi) - v_{ref}(t, \lambda, \phi) \right)^2 dt} \qquad \text{(A6)}$$

1301

1302  where $t_0$ and $t_f$ are the initial and final time steps, respectively. A similar metric is the centralized

1303  *rmse* (*crmse*), which is based on the residuals of the anomalies:

1304

1305  $crmse(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left[ \left( v_{mod}(t, \lambda, \phi) - \overline{v_{mod}}(\lambda, \phi) \right) - \left( v_{ref}(t, \lambda, \phi) - \overline{v_{ref}}(\lambda, \phi) \right) \right]^2 dt}$ (A7)

1306

1307  The *crmse*, therefore, assesses residuals that have been bias-corrected. Since we already

1308  assessed the model's bias through $S_{bias}$, it is convenient to assess the residuals using *crmse* rather

1309  than *rmse*. In a similar fashion to the bias, we then compute a relative error:

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

1310

1311  scale this error onto a unit interval:

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

1312

1313  and compute the spatial mean:

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

1314

1315  **A3 Phase Score ($S_{phase}$)**

1316  The skill score $S_{phase}$ assesses how well the model reproduces the seasonality of a variable by

1317  computing the time difference $\theta(\lambda, \phi)$ between modeled and observed maxima of the

1318  climatological mean cycle:

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

1319

1320  where $c_{mod}$ and $c_{ref}$ are the climatological mean cycle of the model and reference data,

1321  respectively. This time difference is then scaled from zero to one based on the consideration that

1322  the maximum possible time difference is 6 months:

$$s_{phase}(\lambda, \phi) = \frac{1}{2}\left[1 + \cos\left(\frac{2\pi\,\theta(\lambda,\phi)}{365}\right)\right] \tag{A12}$$

1323

1324  The spatial mean of $s_{phase}$ then leads to the scalar score:

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

1325

1326

1327  **A4 Inter-Annual Variability Score ($S_{iav}$)**

1328  The skill score $S_{iav}$ quantifies how well the model reproduces patterns of inter-annual variability.

1329  This score is based on data where the seasonal cycle ($c_{mod}$ and $c_{ref}$) has been removed:

1330  $$iav_{mod}(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left( v_{mod}(t, \lambda, \phi) - c_{mod}(t, \lambda, \phi) \right)^2 dt} \qquad \text{(A14)}$$

1331  $$iav_{ref}(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left( v_{ref}(t, \lambda, \phi) - c_{ref}(t, \lambda, \phi) \right)^2 dt} \, . \qquad \text{(A15)}$$

1332

1333  The relative error, nondimensionalization, and spatial mean are computed next:

1334  $$\varepsilon_{iav}(\lambda, \phi) = \left| iav_{mod}(\lambda, \phi) - iav_{ref}(\lambda, \phi) \right| / iav_{ref}(\lambda, \phi) \qquad \text{(A16)}$$

1335  $$s_{iav}(\lambda, \phi) = e^{-\varepsilon_{iav}(\lambda, \phi)} \qquad \text{(A17)}$$

1336  $$S_{iav} = \overline{\overline{s_{iav}(\lambda, \phi)}} \qquad \text{(A13)}$$

1337  **A5 Spatial Distribution Score ($S_{dist}$)**

1338  The spatial distribution score $S_{dist}$ assesses how well the model reproduces the spatial pattern of

1339  a variable. The score considers the correlation coefficient $R$ and the relative standard deviation $\sigma$

1340  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

1341  $\sigma$ approach a value of one. No spatial integration is required as this calculation yields a single

1342  value:

1343  $$S_{dist} = 2(1 + R) \left( \sigma + \frac{1}{\sigma} \right)^{-2} \qquad \text{(A19)}$$

1344    where $\sigma$ is the ratio between the standard deviation of the model and reference data:

1345
$$\sigma = \sigma_{\overline{v_{mod}}} / \sigma_{\overline{v_{ref}}} \qquad (A20)$$

1346    and $\sigma_{\overline{v_{mod}}}$ and $\sigma_{\overline{v_{ref}}}$ are the standard deviations of the annual mean values from the model and

1347    reference/observation-based data, respectively, and therefore are scalars.

1348    **A6 Overall Score ($S_{overall}$)**

1349    As a final step, scores are averaged to obtain an overall score:

1350
$$S_{overall} = \frac{S_{bias} + 2\,S_{rmse} + S_{phase} + S_{iav} + S_{dist}}{1+2+1+1+1} \qquad (A21)$$

1351    Note that $S_{rmse}$ is weighted by a factor of two and is an entirely subjective decision but follows

1352    Collier et al. (2018).

1353
1354

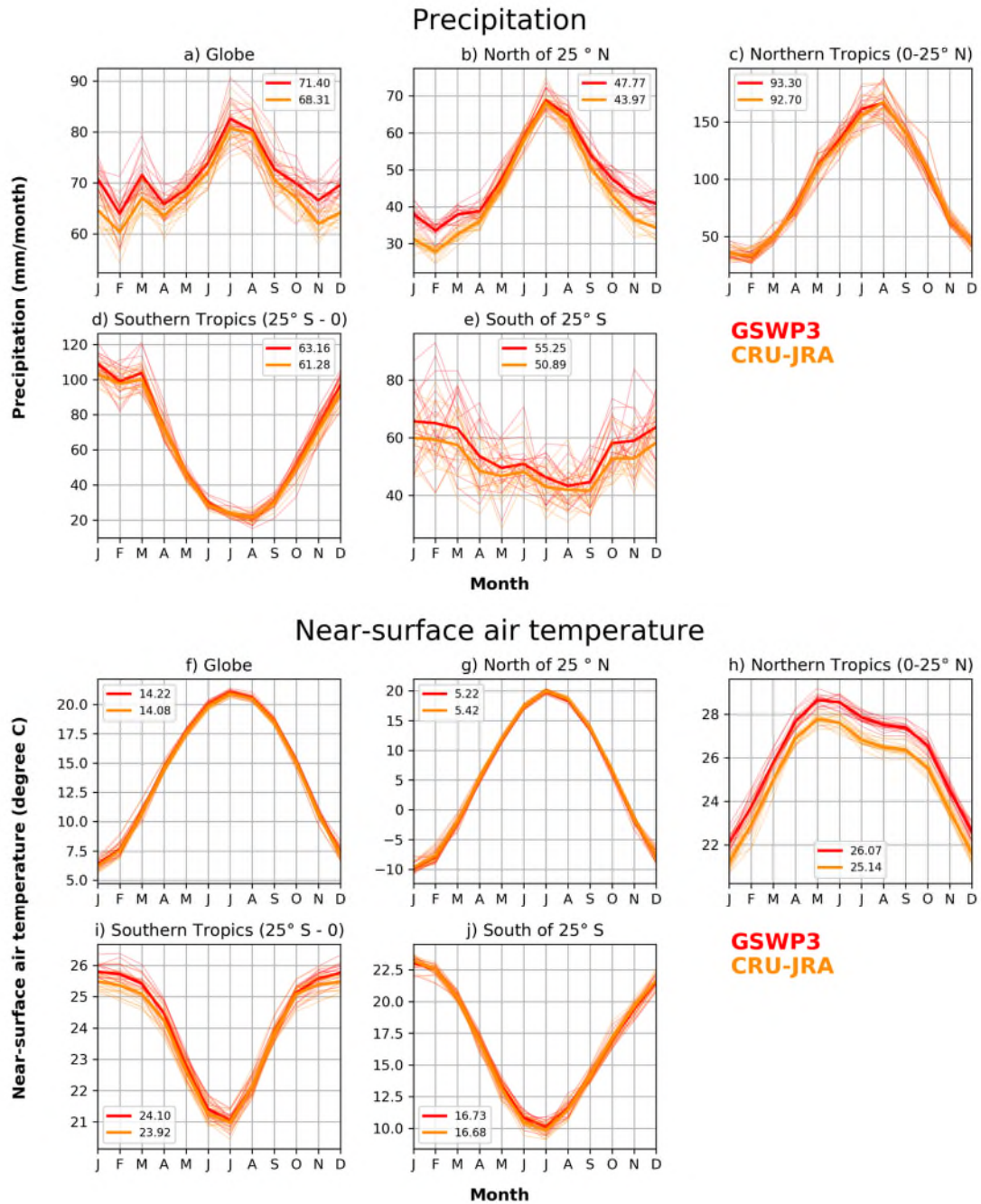
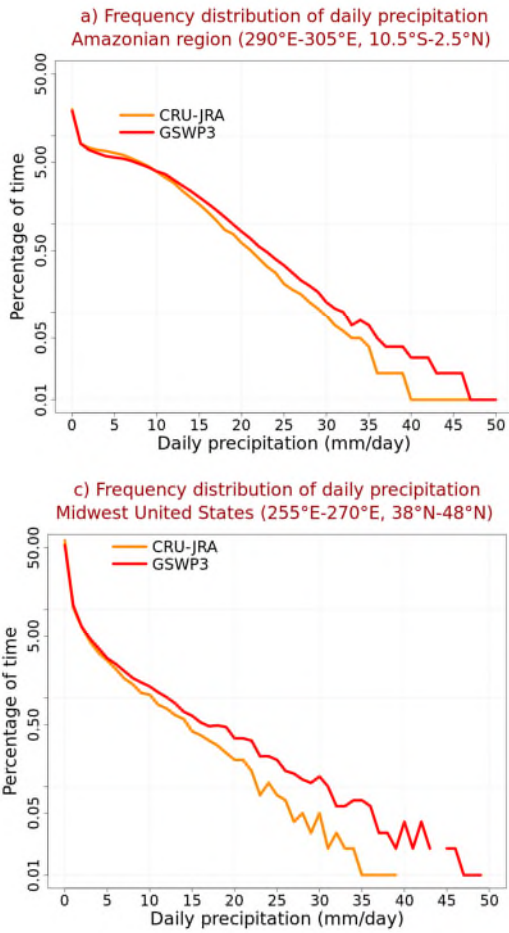



Figure A1: Comparison of monthly precipitation (upper panel) and temperature (lower panel) for five global regions (global, north of 25 °N, northern and southern tropics, and south of 25 °S) from the CRU-JRA and GSWP3 meteorological forcing data sets that are used to drive the CLASSIC model. The global and regional averages exclude Greenland and Antarctica. The legend entries show the annual mean values averaged over the 1997-2016 period. The thin lines show individual years and the thick line is their average.

1363

1364

1365
1366
1367
1368 Figure A2: Comparison of the frequency distribution of daily precipitation between the CRU-JRA
1369 and GSWP3 meteorological data sets for three broad regions and the period 1997-2016: a) the
1370 Amazonian region, b) the Sahel region, and c) the Midwest United States. The frequency is
1371 represented as a percentage of time daily precipitation is between *x* and *x*+1 mm/day, where x
1372 is the value on the x-axis. Panel (d) shows the location of these broad regions. The underlying
1373 map in panel (d) is from Google Maps.

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

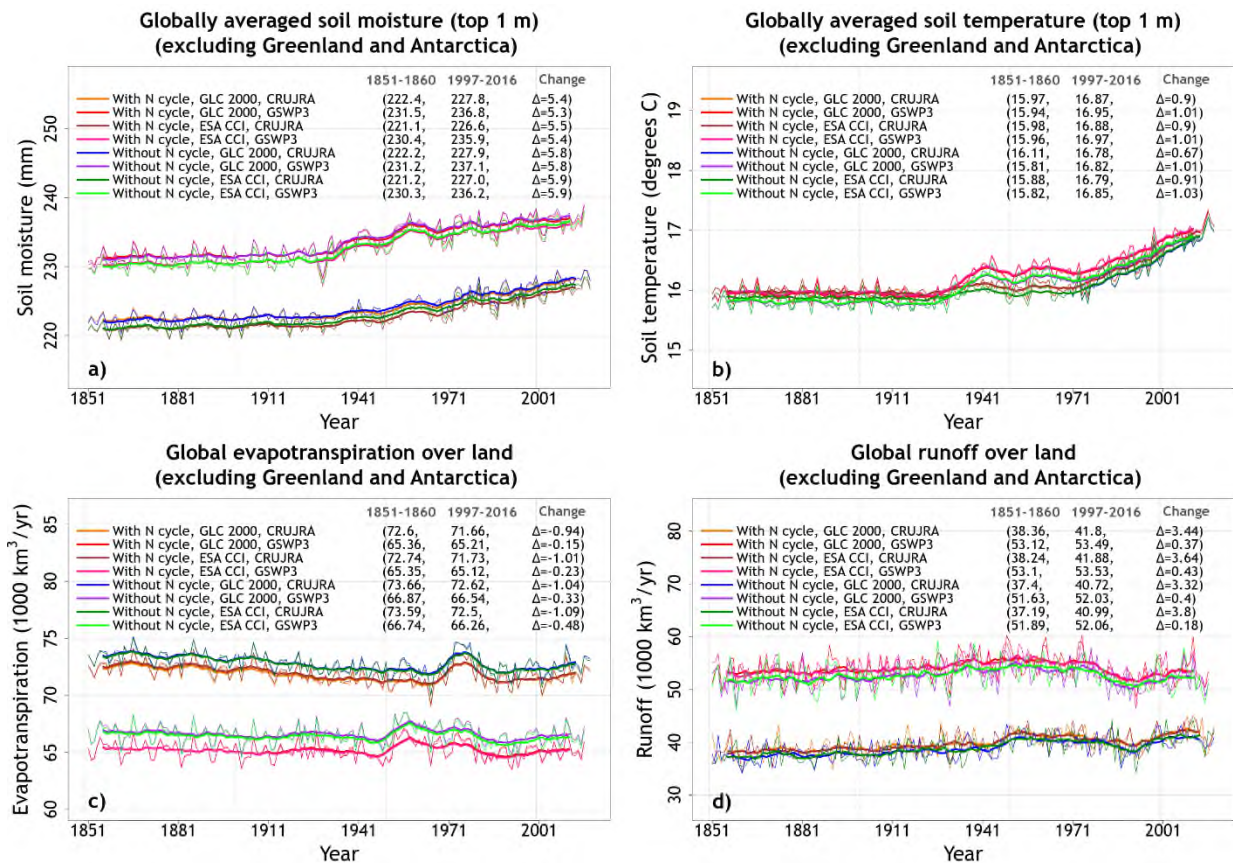
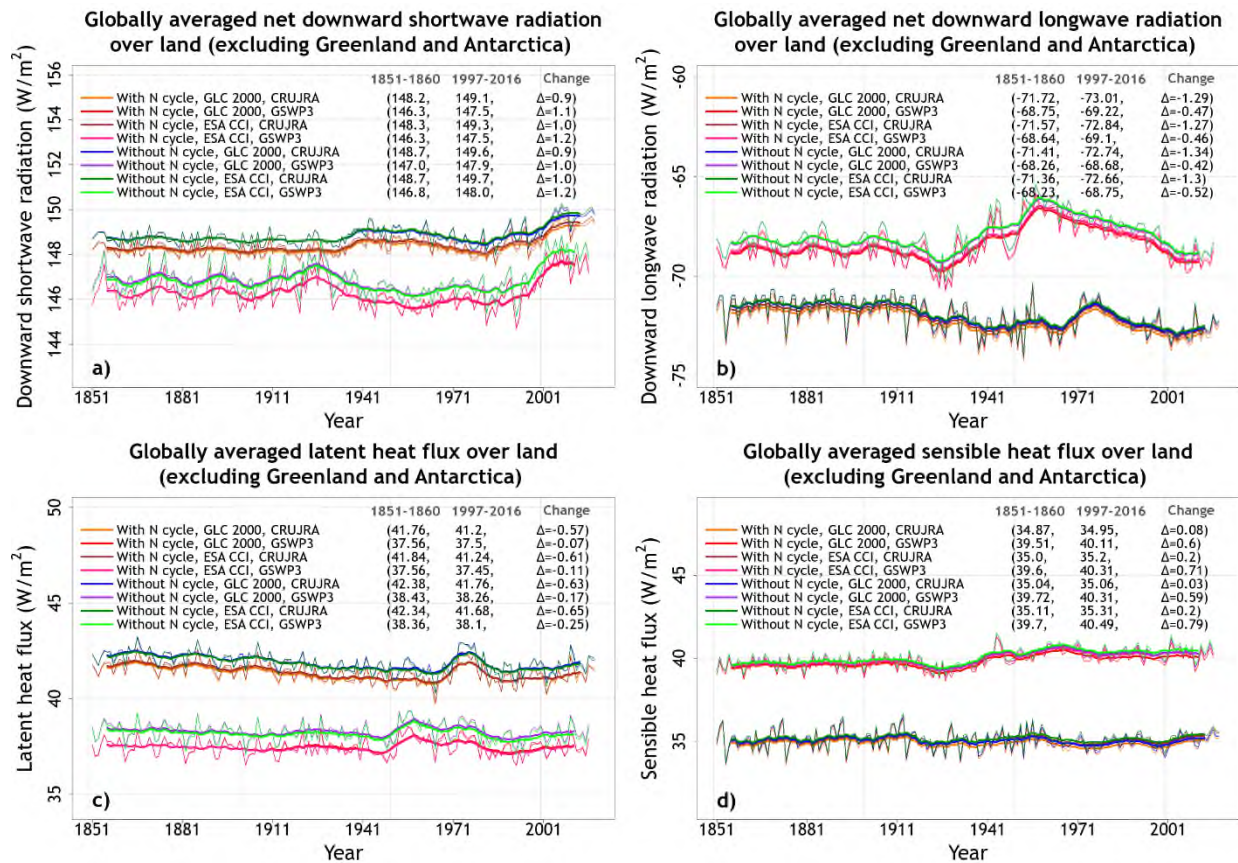



Figure A3: Time series of simulated globally-averaged annual soil moisture (a) and soil temperature (b) in the top 1m, global annual evapotranspiration (c), and runoff (d) from the eight simulations summarized in Table 1. The thin lines show the individual years and the thick lines show their 11-year moving average. Model values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods, and their difference, are also shown.

1389

1390

1391

1392

1393    Figure A4: Time series of simulated globally-averaged annual energy fluxes from the eight
1394    simulations summarized in Table 1. Panel (a) shows net downward shortwave radiation, panel
1395    (b) shows net downward longwave radiation, panel (c) shows latent heat flux, and panel (d)
1396    shows sensible heat flux. The thin lines show the individual years and the thick lines show their
1397    11-year moving average. Model values averaged over the pre-industrial (1851-1860) and
1398    present-day (1997-2016) time periods, and their difference, are also shown for individual
1399    simulations.

1400

1401

1402

1403

1404

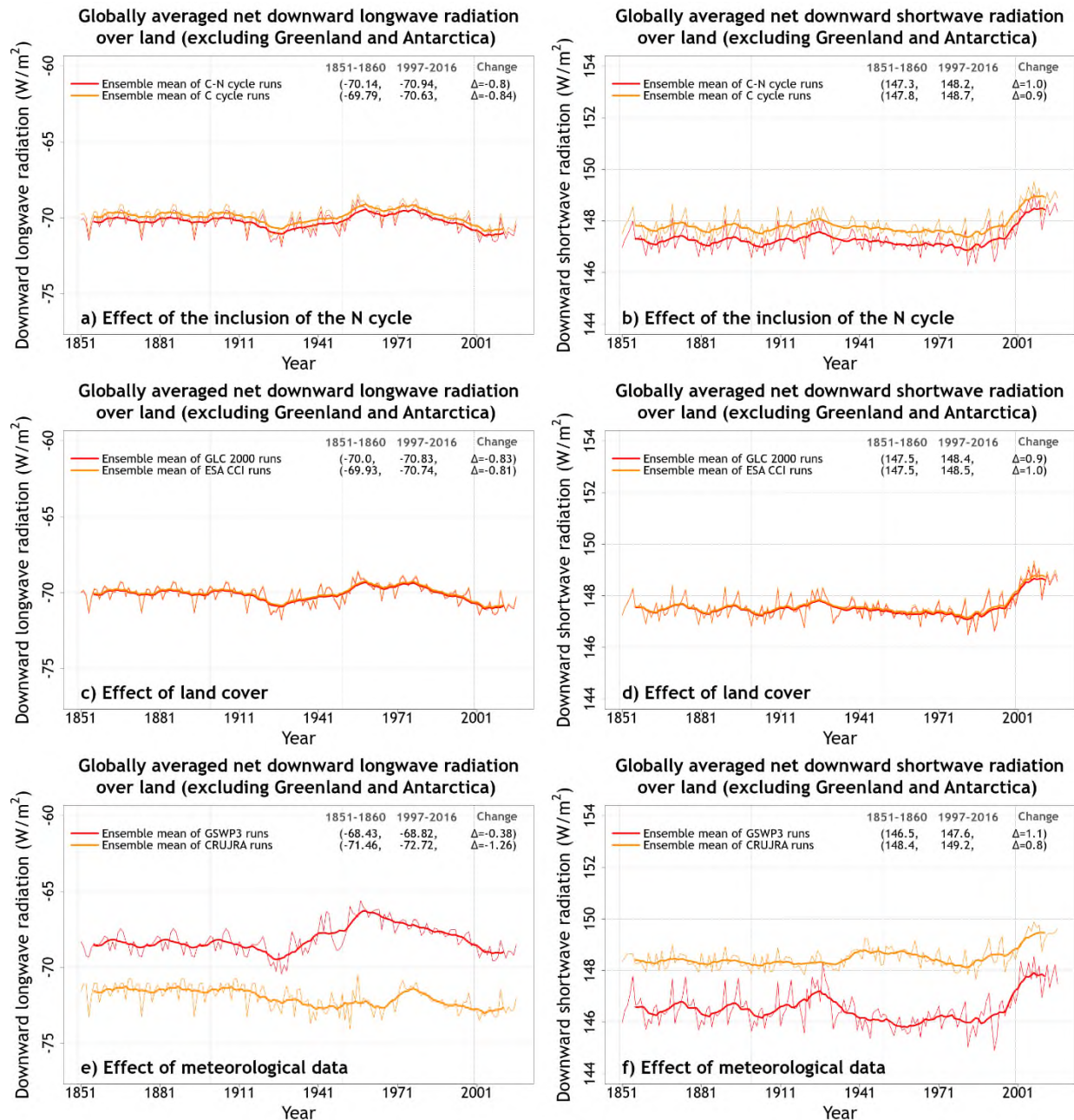



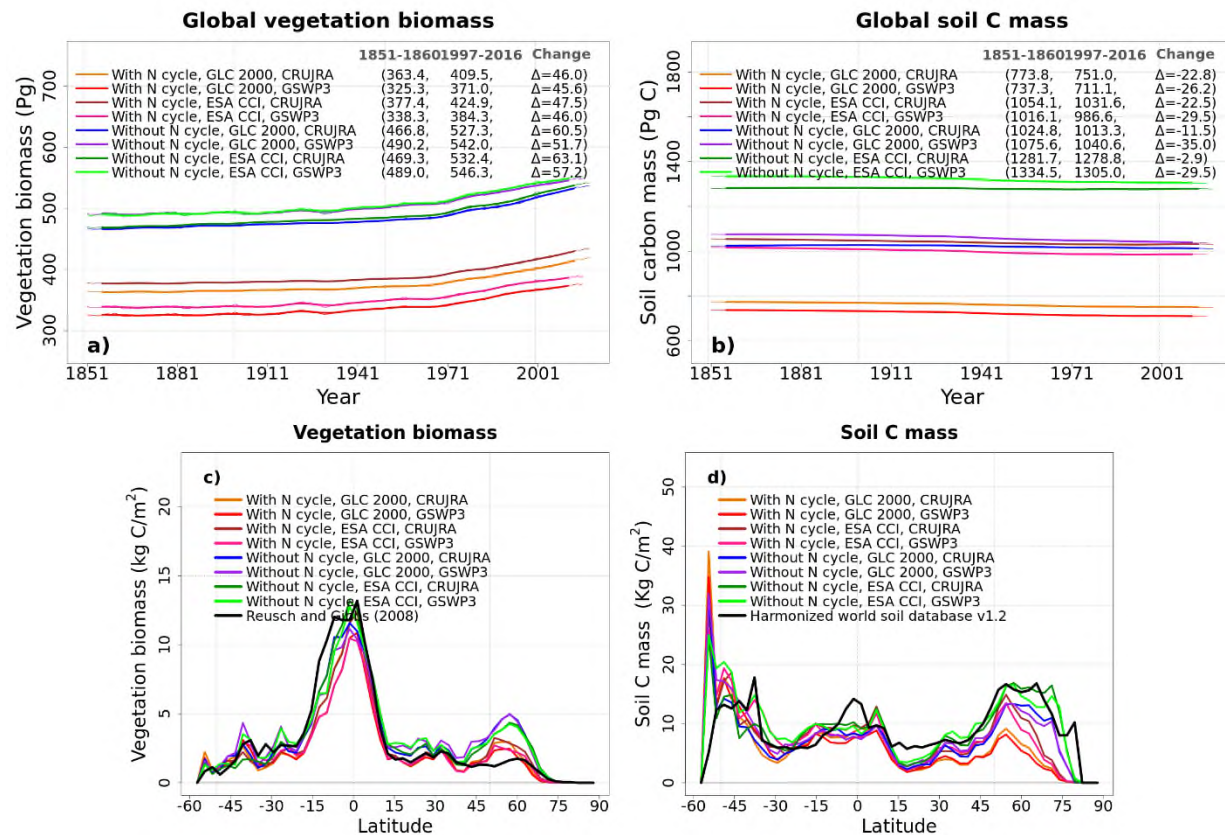
Figure A5: Time series of globally-averaged annual net downward longwave and shortwave radiation (over all land area excluding Greenland and Antarctica) averaged over the four ensemble members each that are driven with and without N cycle (panels a, b), driven with GLC 2000 and ESA CCI based land cover (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.

1412

1413

1414

1415

1416

1417 Figure A6: Time series of simulated global annual vegetation carbon mass (a) and soil carbon (b)
1418 from the eight simulations summarized in Table 1. The global totals exclude Greenland and
1419 Antarctica. Panels (c) and (d) show the zonally-averaged values of vegetation carbon mass and
1420 soil carbon mass over land from the eight simulations averaged over the 1997-2016 period. The
1421 thin lines show the individual years and the thick lines show their 11-year moving average in
1422 panels (a) and (b). Model values averaged over the pre-industrial (1851-1860) and present-day
1423 (1997-2016) time periods, and their difference, are also shown in panels (a) and (b).

1424

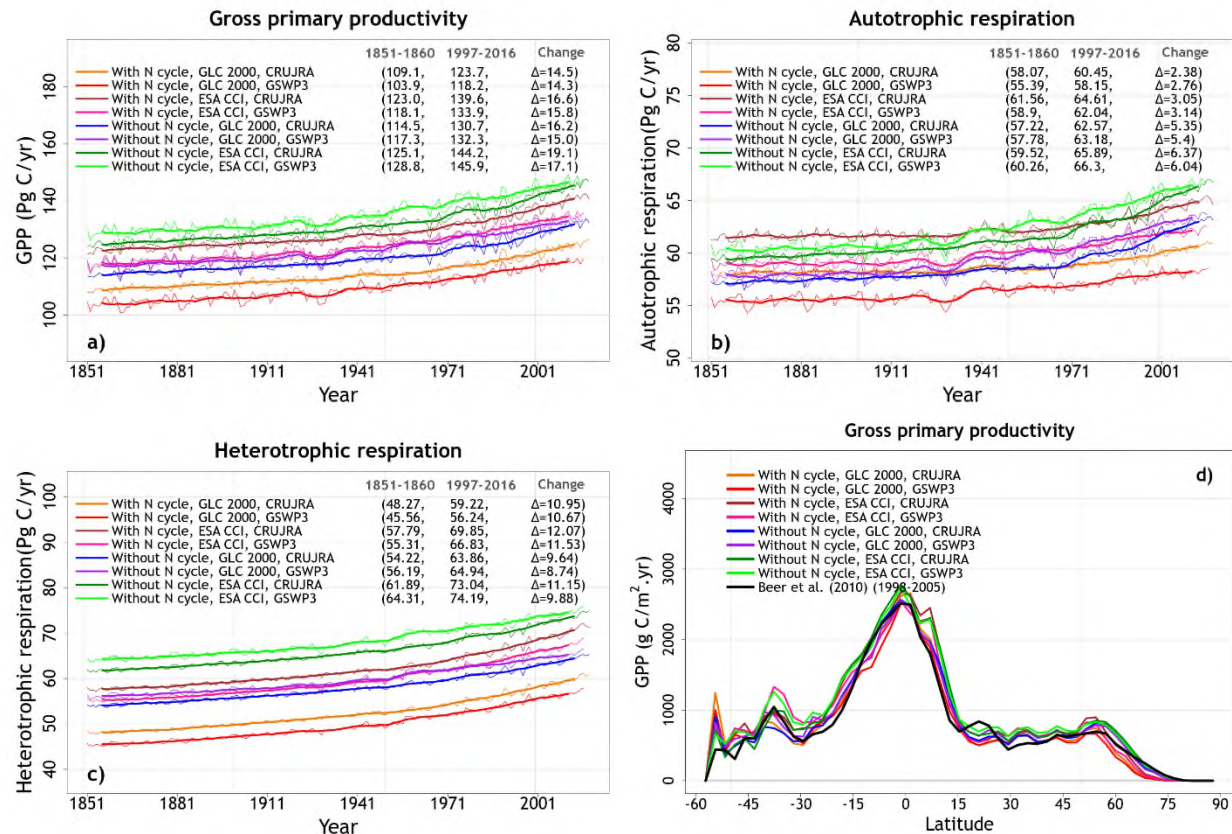
1425

1426

1427

1428

1429

1430

1431 Figure A7: Time series of simulated global annual gross primary productivity (GPP) (a),
1432 autotrophic respiration (b), and heterotrophic respiration (c) from the eight simulations
1433 summarized in Table 1. Panel (d) shows the zonally-averaged values of GPP from the eight
1434 simulations averaged over the 1997-2016 period for each simulation. The thin lines show the
1435 individual years and the thick lines show their 11-year moving average in panels (a) to (c). Model
1436 values averaged over the pre-industrial (1851-1860) and present-day (1997-2016) time periods,
1437 and their difference, are also shown in panels (a) to (c).

1438

1439

1440

1441

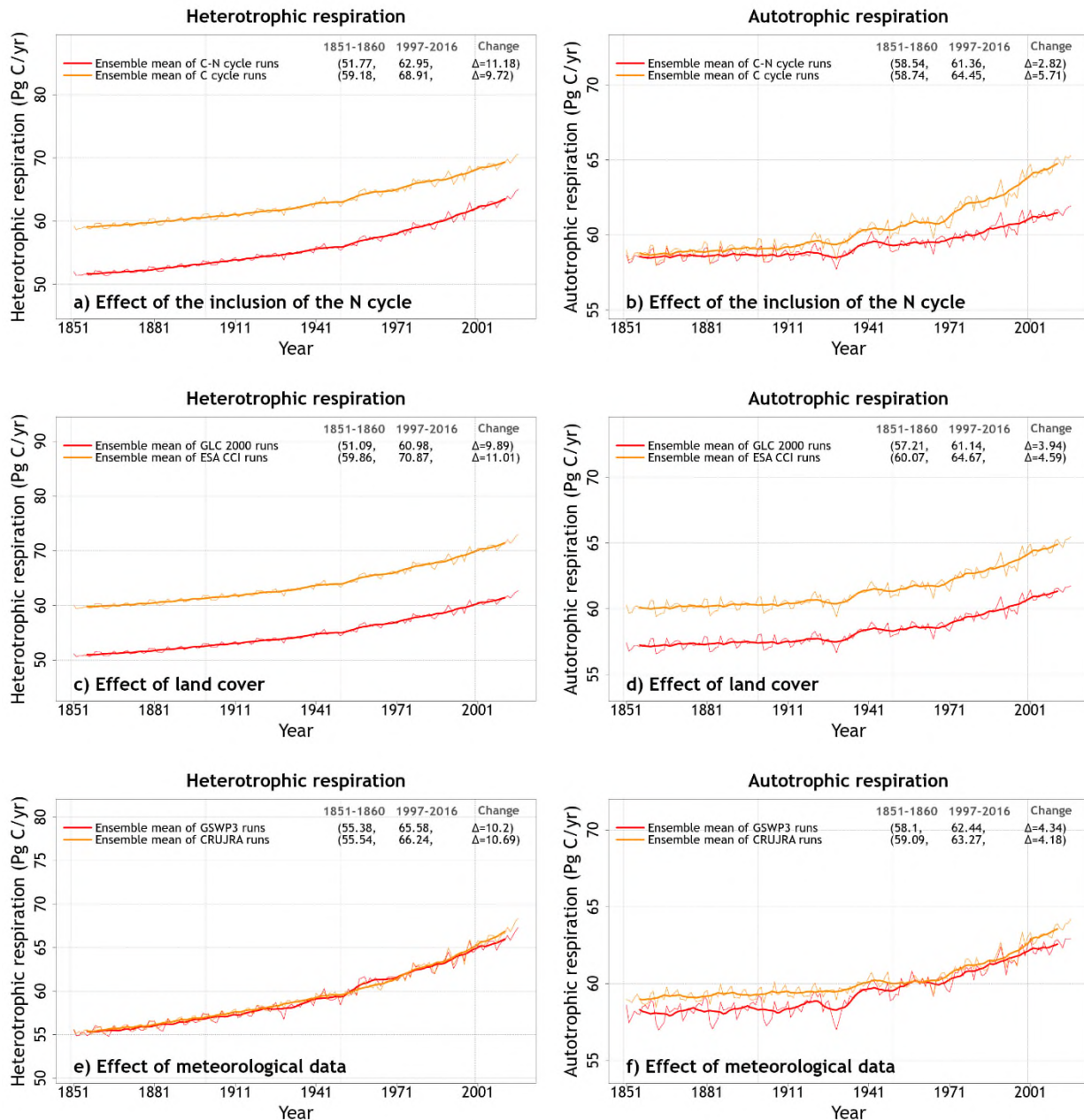



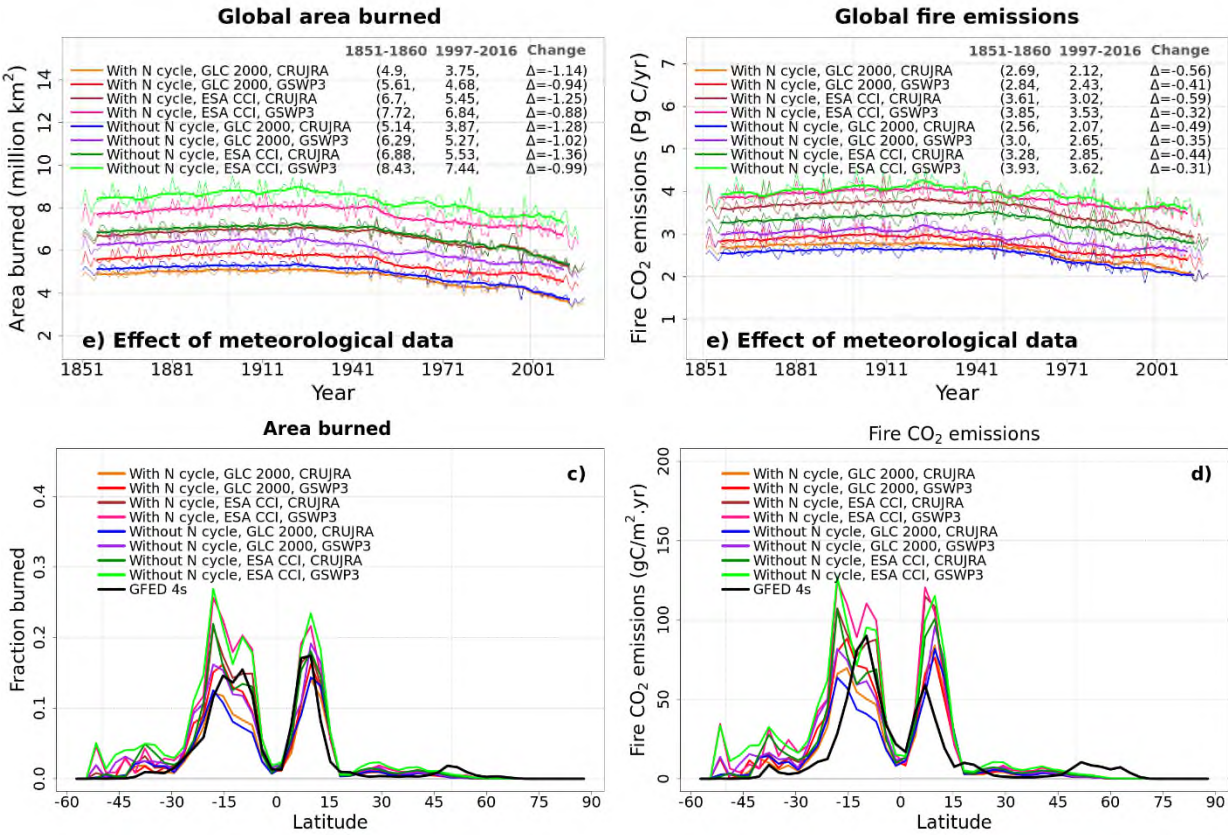
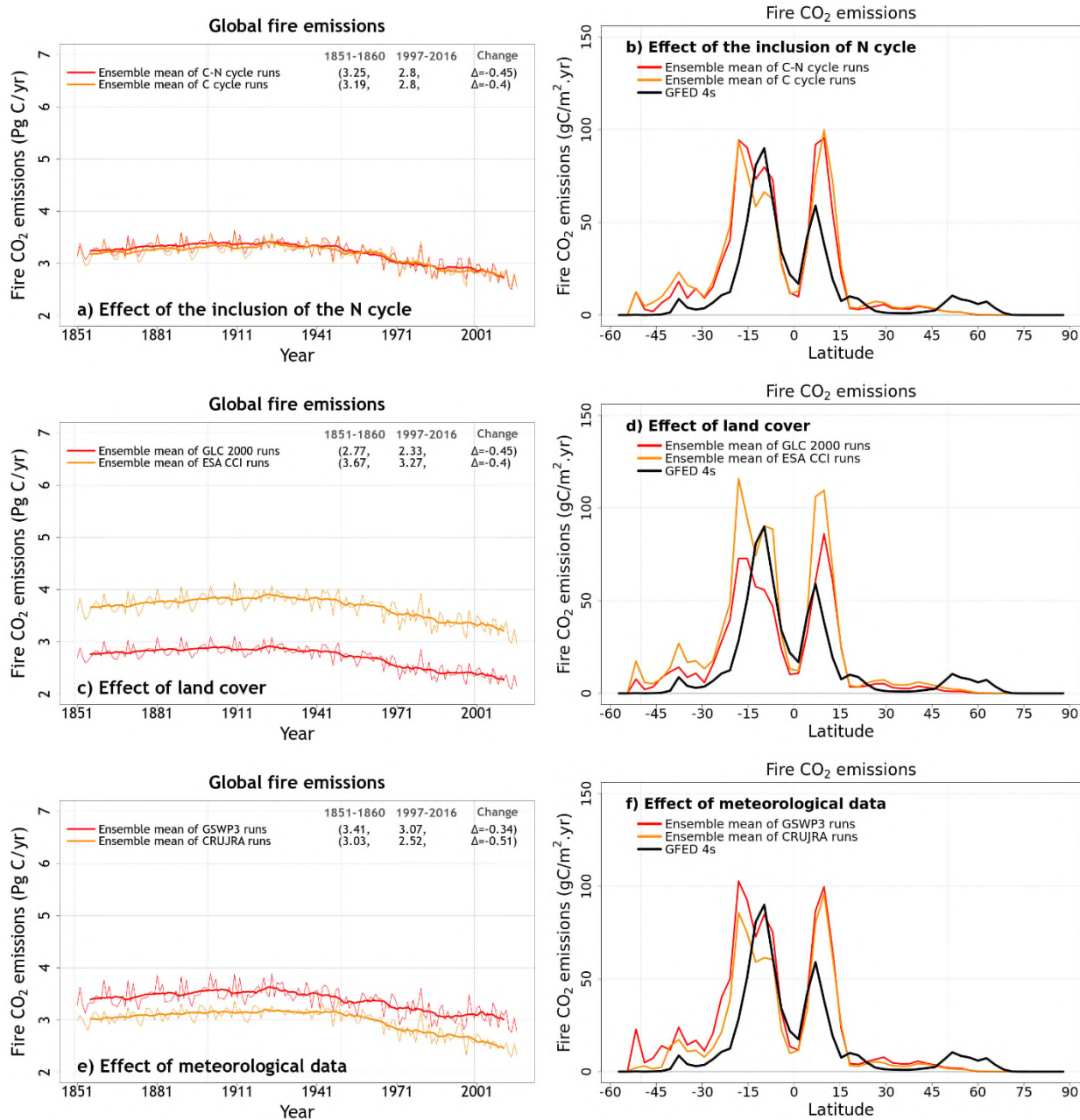
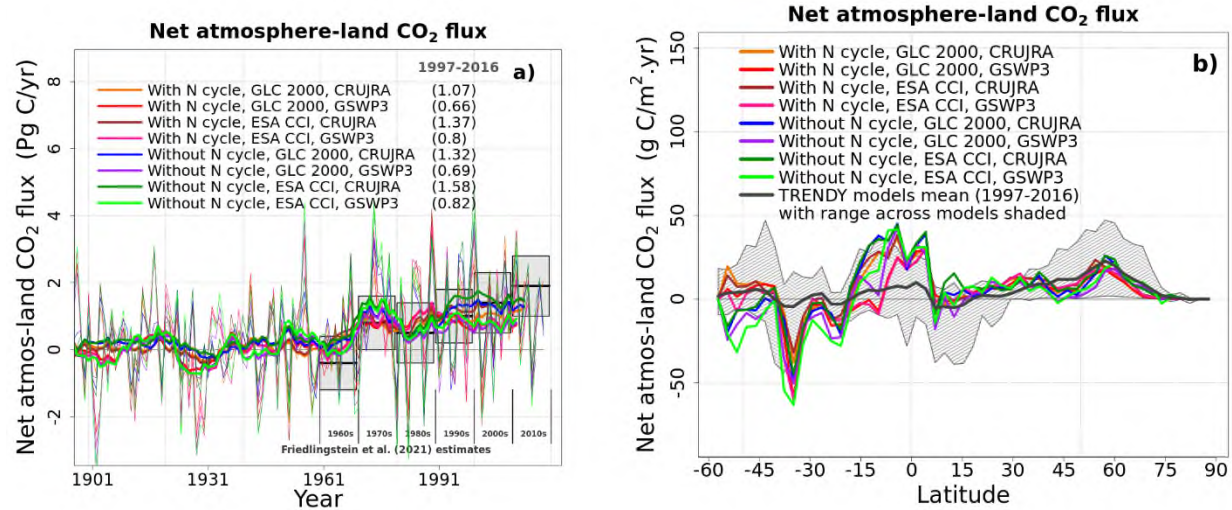
Figure A8: Time series of global heterotrophic and autotrophic respiration (over all land area excluding Greenland and Antarctica) averaged over the four ensemble members each that are driven with and without an interactive N cycle (panels a, b), driven with the GLC 2000 and ESA CCI based land cover (panels c, d), and driven the 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.

1453  Figure A9: Time series of simulated global annual area burned (a) and fire $CO_2$ emissions (b)
1454  from the eight simulations summarized in Table 1. Panels (c) and (d) show the zonally-averaged
1455  area burned and fire $CO_2$ emissions from the eight simulations averaged over the 1997-2016
1456  period. The thin lines for the time series show the individual years and the thick lines show their
1457  11-year moving average. Model values averaged over the pre-industrial (1851-1860) and
1458  present-day (1997-2016) time periods, and their difference, are also shown for panels (a) and
1459  (b).

1463

Figure A10: Time series of global fire $CO_2$ emissions (over all land area excluding Greenland and Antarctica) (panels a, c, and e) and their zonally-averaged values (panels b, d, and f) averaged over the four ensemble members each that 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 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 in panels (a), (c), and (e). 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).

1472

1473

1474

1475

1476    Figure A11: Time series of simulated global annual net atmosphere-land $CO_2$ flux (a) and its
1477    zonally-averaged values from the eight simulations summarized in Table 1 averaged over the
1478    1997-2016 period. In panel (a) simulated annual net atmosphere-land $CO_2$ flux values are
1479    compared to the estimates from the Global Carbon Project (Friedlingstein et al., 2022). The thin
1480    lines for the time series in panel (a) show the individual years and the thick lines show their 11-
1481    year moving average. In panel (b) the simulated zonally-averaged values are compared to the
1482    range from 11 models that contributed to the TRENDY 2020 intercomparison.

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

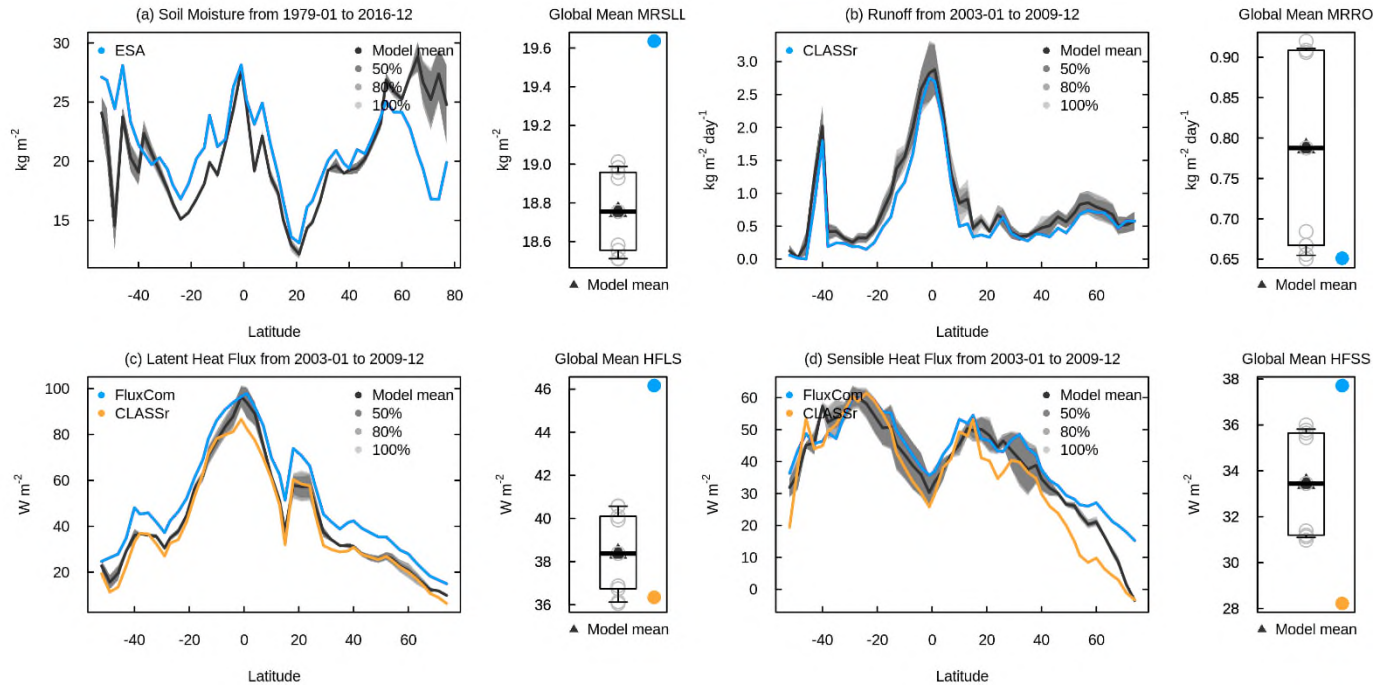



Figure A12: Zonally-averaged values of soil moisture (a), runoff (b), latent heat flux (c), and sensible heat flux (d) from the eight simulations summarized in Table 1. The model results are shown as their mean (black) and the spread across the eight simulations indicated by 50%, 80%, and 100% ranges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

1501

1502

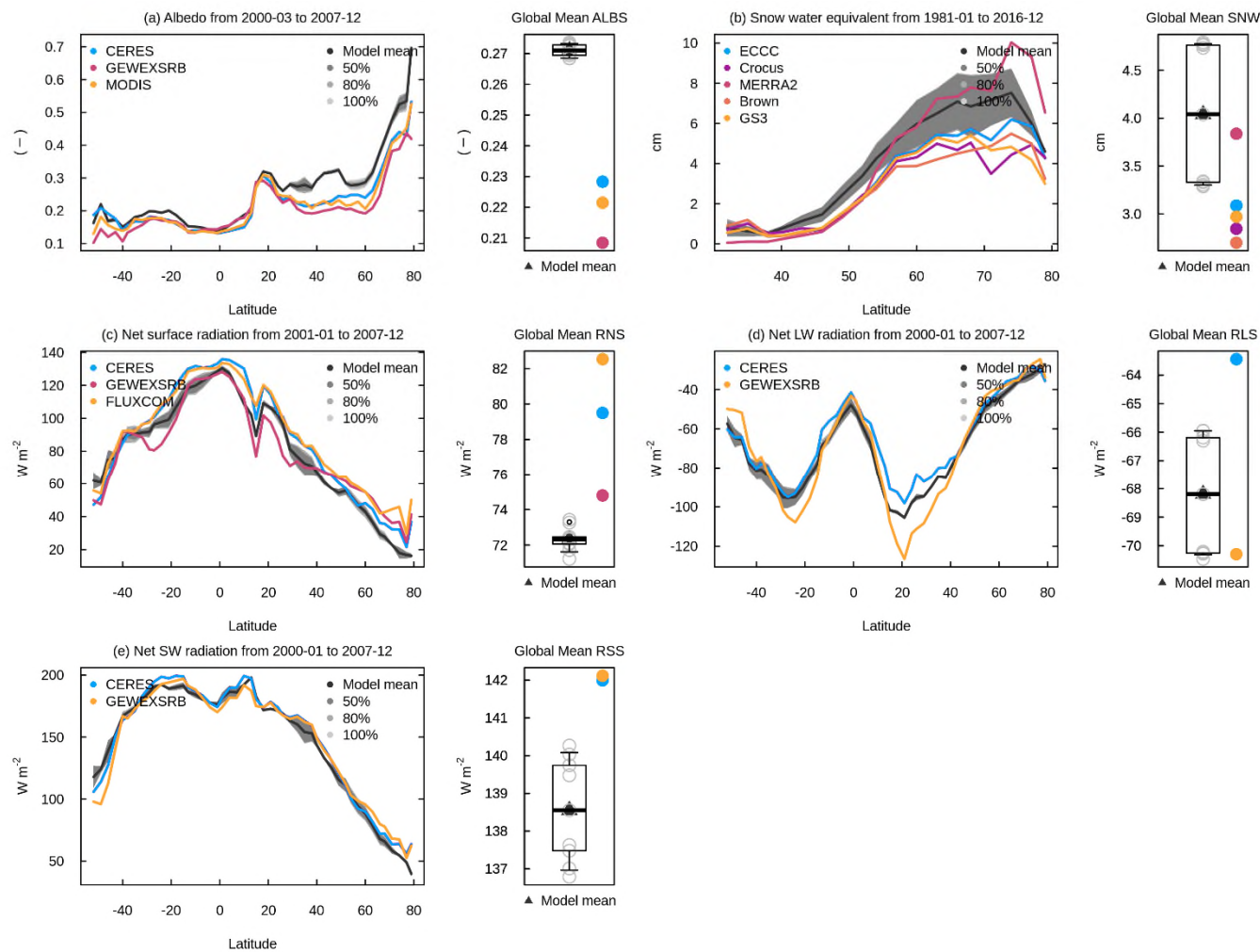
1503

1504

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.

1511

1512