Evaluating Northern Hemisphere Growing Season Net Carbon Flux in Climate Models Using Aircraft Observations

Morgan Loechli[1] , Britton B. Stephens[2] , Roisin Commane[3] , Frédéric Chevallier[4] , Kathryn McKain[5,6] , Ralph F. Keeling[7] , Eric J. Morgan[7] , Prabir K. Patra[8] , Maryann R. Sargent[9], Colm Sweeney[6] , and Gretchen Keppel-Aleks[1]

[1]University of Michigan, Ann Arbor, MI, USA, [2]National Center for Atmospheric Research, Boulder, CO, USA, [3]Department of Earth and Environmental Sciences, Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA, [4]LSCE Labratoire des Sciences du Climat ed de l'Environnement, Gif-Sur-Yvette Cedes, France, [5]Global Monitoring Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA, [6]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA, [7]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA, [8]Research Institute for Global Change, JAMSTEC, Yokohama, Japan, [9]Harvard University, Cambridge, MA, USA

Abstract Understanding terrestrial ecosystems and their response to anthropogenic climate change requires quantification of land-atmosphere carbon exchange. However, top-down and bottom-up estimates of large-scale land-atmosphere fluxes, including the northern extratropical growing season net flux (GSNF), show significant discrepancies. We developed a data-driven metric for the GSNF using atmospheric carbon dioxide concentration observations collected during the High-Performance Instrumented Airborne Platform for Environmental Research Pole-to-Pole Observations and Atmospheric Tomography Mission flight campaigns. This aircraft-derived metric is bias-corrected using three independent atmospheric inversion systems. We estimate the northern extratropical GSNF to be 5.7 ± 0.3 Pg C and use it to evaluate net biosphere productivity from the Coupled Model Intercomparison Project phase 5 and 6 (CMIP5 and CMIP6) models. While the model-to-model spread in the GSNF has decreased in the CMIP6 models relative to that of the CMIP5 models, there is still disagreement on the magnitude and timing of seasonal carbon uptake with most models underestimating the GSNF and overestimating the length of the growing season relative to the observations. We also use an emergent constraint approach to estimate annual northern extratropical gross primary productivity to be 56 ± 17 Pg C, heterotrophic respiration to be 25 ± 13 Pg C, and net primary productivity to be 28 ± 12 Pg C. The flux inferred from these aircraft observations provides an additional constraint on large-scale gross fluxes in prognostic Earth system models that may ultimately improve our ability to accurately predict carbon-climate feedbacks.

Plain Language Summary The exchange of carbon between the land and atmosphere is an important part of the Earth's climate, and this exchange might change due to human-caused climate change. However, estimates of land-atmosphere carbon fluxes made using different techniques do not agree with each other. We use atmospheric carbon dioxide observations collected during two flight campaigns to show that 5.7 Pg C is exchanged between the atmosphere and the land in the northern hemisphere during the summer growing season. This estimate is used to evaluate the performance of two generations of climate prediction models. The newer generation of models show less spread than the older generation, but there is still significant disagreement on the magnitude and timing of land-atmosphere carbon exchange among models. Most models underestimate the growing season net flux and overestimate the length of the growing season. We also use our observational estimate to reduce the spread on component fluxes of carbon exchange, namely uptake by photosynthesis and release by respiration.

1. Introduction

Approximately half of the carbon dioxide (CO_2) released annually by the combustion of fossil fuels stays in the atmosphere (Friedlingstein et al., 2022; Keeling et al., 1976; Schimel et al., 2001). The remaining CO_2 is taken up by the terrestrial biosphere and ocean in roughly equal proportions (Keeling & Manning, 2014; Khatiwala et al., 2009; Sabine et al., 2004). The efficiency of the ocean and land sinks varies with both climate and atmospheric CO_2, representing an important feedback in the climate system (e.g., Ballantyne et al., 2012;

Fernández-Martínez et al., 2019; Fung et al., 2005). The strength of the land sink may be related to the amplitude of the seasonal cycle of atmospheric CO_2 (e.g., Keeling et al., 1996; Randerson et al., 1997) via annual and seasonal imbalances between photosynthesis and respiration. However, the magnitude and spatial and temporal distributions of gross primary productivity (GPP) and net primary productivity (NPP) vary noticeably among Earth system models (ESMs) (e.g., Hu et al., 2022). Furthermore, models typically underestimate the change in amplitude of seasonal CO_2 exchange in northern land ecosystems over time (e.g., Graven et al., 2013) or underestimate CO_2 uptake in the Northern Hemisphere mid-high latitudes (e.g., Canadell et al., 2021, Figure 5.24).

Multi-model ensembles of coupled carbon-climate models show large differences in their land sink projections, especially for terrestrial carbon uptake (e.g., Arora et al., 2020; Cadule et al., 2010). For example, Friedlingstein et al. (2014) showed that the Coupled Model Intercomparison Project phase 5 (CMIP5) models range between −173 and 758 Pg C in simulations of cumulative land carbon uptake for 1850 to 2100 when forced by RCP8.5. This uncertainty exists in historical simulations where models both overestimate and underestimate the historical atmospheric CO_2 increase by over 20%. These differences are mainly due to uncertainties in the land carbon cycle response, with differences in their cumulative land flux estimates of 214 Pg C for 1850–2005, more than double the differences in their cumulative ocean flux estimates (Friedlingstein et al., 2014).

Quantifying the exchange of carbon between the atmosphere and the land surface at hemispheric and global scales is challenging because the heterogeneity of the Earth's surface makes it difficult to upscale local flux measurements (e.g., Friend et al., 2007; Kumar et al., 2016). Atmospheric inversion, wherein carbon fluxes are estimated from atmospheric CO_2 observations using atmospheric tracer transport models, provides a method to infer large-scale carbon fluxes (e.g., Ciais et al., 2010; Tans et al., 1990; Thompson et al., 2016); however, this method has been shown to be sensitive to uncertainty due to the simulation of vertical transport (Schuh et al., 2019; Stephens et al., 2007; Verma et al., 2017). Atmospheric inversions that rely only on surface observations must accurately represent vertical mixing to estimate CO_2 concentrations aloft. Uncertainty in atmospheric inversion flux estimates can be characterized through the use of observations of the vertical profile of atmospheric CO_2 (e.g., Peiro et al., 2022; Stephens et al., 2007).

Global-scale aircraft observations, such as those made during the High-Performance Instrumented Airborne Platform for Environmental Research (HIAPER) Pole-to-Pole Observations project (HIPPO, 2009–2011) and the Atmospheric Tomography Mission (ATom, 2016–2018), are representative of large regions and capture the vertical profile of atmospheric CO_2 (Thompson et al., 2022; Wofsy, 2011). These campaigns measured the vertical structure of CO_2 in the atmosphere across a range of latitudes and over the full seasonal cycle and allowed for the analysis of seasonal changes in hemispheric-scale atmospheric CO_2 (e.g., Jin et al., 2021), which are dominated by land exchange. We use the seasonal cycle of atmospheric CO_2 concentrations measured during the HIPPO and ATom flight campaigns to develop a metric for evaluating the simulation of terrestrial CO_2 exchange in prognostic ESMs.

We derive estimates of the northern hemisphere net land flux integrated over the growing season, or growing season net flux (GSNF), as a benchmark for model evaluation (e.g., Collier et al., 2018). The creation of flux benchmarks allows for a direct comparison of observations and model simulations at the flux level rather than at the concentration level (e.g., Keppel-Aleks et al., 2013), which requires either using an atmospheric transport model or emulator (Liptak et al., 2017) to translate fluxes into atmospheric mole fraction variations. This research explores an alternative approach to formal inverse modeling to constrain net land-atmosphere carbon fluxes at the hemispheric scale. We use CO_2 measurements from the HIPPO and ATom flight campaigns to infer the GSNF with only minimal reliance on atmospheric transport models. Thus, our estimated flux is less sensitive to errors in transport simulation and gives a more robust insight into prognostic model inconsistencies.

We describe the data sets and methods used to derive GSNF in Section 2. We discuss the GSNF and compare it to ESM estimates of net biosphere productivity (NBP), GPP, heterotrophic respiration (RH), and NPP using the output from the Coupled Model Intercomparison Project phases 5 and 6 (CMIP5 and CMIP6) in Section 3. This is followed by a discussion of those results in Section 4 and conclusions in Section 5.
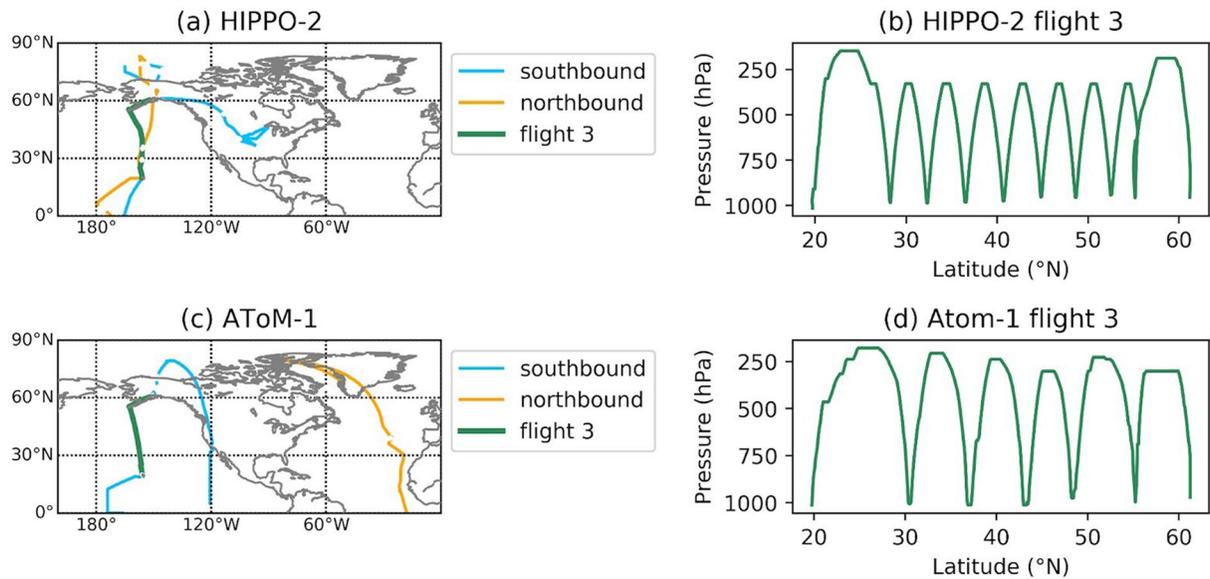
Figure 1. Flight paths for (a) HIPPO-2, which flew over the remote Pacific in November 2009, and (c) ATom-1, which flew over the Pacific and Atlantic in August 2016. All other campaigns followed similar flight paths. Flight path with continuous vertical profiling for flight 3, which flew from Anchorage, AK to Kona, HI for (b) HIPPO-2 and (d) ATom-1. All other flights flew a similar path.

2. Data and Methods

2.1. Aircraft Observations

We used dry air mole fractions of CO_2 measured in the free troposphere during the HIPPO and ATom aircraft campaigns. HIPPO (Wofsy, 2011; Wofsy et al., 2017) used the NSF/NCAR HIAPER Gulfstream V aircraft to make measurements primarily over the remote Pacific from 87°N to 67°S (Figure 1a) during five campaigns that spanned all four seasons between 2009 and 2011 (Table 1). The aircraft flew vertical profiles from near the surface to an altitude of 14 km; typically, a full profile was completed over ~2.2° of latitude (Figure 1b). During these flights, measurements were made of greenhouse gasses and related tracers. CO_2 mole fractions were measured using three different in situ instruments and two whole air samplers: the Harvard Quantum Cascade Laser System (QCLS, Santoni et al., 2014), the Harvard Observations of the Middle Stratosphere (OMS, Daube et al., 2002) instrument, the National Center for Atmospheric Research (NCAR) Airborne Oxygen Instrument (AO2, Stephens et al., 2021), the National Oceanic and Atmospheric Administration (NOAA) Portable Flask Packages (PFP, Sweeney et al., 2015), and the NCAR/Scripps Medusa Whole Air Sampler (Stephens et al., 2021). For our analysis, we used the recommended CO2.X variable, which is derived primarily from QCLS measurements with calibration periods gap-filled using OMS measurements, reported as part per million dry air mole fractions (Wofsy et al., 2017). We used the 10-s merge data product and all CO_2 measurements are reported to be within 0.2 ppm with respect to the WMO X2007 scale (Santoni et al., 2014). The mean bias between QCLS and NOAA flask measurements across all five HIPPO campaigns was 0.11 ppm (Santoni et al., 2014). We used comparisons to the other four systems as a measure of analytical uncertainty. We also used observations of N_2O made by QCLS to identify stratospheric samples.

ATom (Thompson et al., 2022; Wofsy et al., 2021) is a more recent series of flight campaigns that used the NASA DC-8 aircraft to measure atmospheric trace gas concentrations by traveling south over the Pacific and north over the Atlantic (Figure 1c) and which included a much larger scientific payload. As with HIPPO, a full annual cycle was measured, with flights that occurred in each of the four seasons over a three-year period from 2016 to 2018 (Table 1). Flights spanned 83°N to 86°S and sampled vertical profiles

Table 1
Aircraft Data Used in This Study

Deployment	Northern Hemisphere (southbound) dates and ocean basin	Northern Hemisphere (northbound) dates and ocean basin
HIPPO-1	1/8/09–1/16/09, Pacific	1/28/09–1/30/09, Pacific
HIPPO-2	10/31/09–11/7/09, Pacific	11/16/09–11/22/09, Pacific
HIPPO-3	3/24/10–3/31/10, Pacific	4/10/10–4/16/10, Pacific
HIPPO-4	6/14/11–6/22/11, Pacific	7/4/11–7/11/11, Pacific
HIPPO-5	8/9/11–8/24/11, Pacific	9/3/11–9/8/11, Pacific
ATom-1	7/29/16–8/6/16, Pacific	8/17/16–8/23/16, Atlantic
ATom-2	1/26/17–2/3/17, Pacific	2/15/16–2/21/17, Atlantic
ATom-3	9/28/17–10/6/17, Pacific	10/19/17–10/28/17, Atlantic
ATom-4	4/24/18–5/1/18, Pacific	5/14/18–5/21/18, Atlantic

from 0.2 to 12 km in altitude (Figure 1d). ATom measured CO_2 using the QCLS, AO2, Medusa, and PFPs similar to HIPPO but also included a NOAA Picarro instrument. For our analysis, we used the CO2.X variable, which consists of NOAA Picarro measurements gap-filled using QCLS measurements. However, during the first two flights of ATom-1, the NOAA Picarro measurements were not reported due to an inlet problem. Similar to HIPPO, we used the 10-s merge data product (Wofsy et al., 2021). To identify and remove stratospheric samples, we used observations of N_2O from QCLS and the NOAA PAN and Trace Hydrohalocarbon ExpeRiment (PANTHER, ATom-1 only). While the WMO CO_2 scale has been recently updated, both the HIPPO and ATom observations used here have been calibrated with respect to the previous, WMO X2007, scale.

2.2. Curtain Averages From Atmospheric Concentrations

CO_2 observations from all flight campaigns are combined to estimate the average northern extratropical tropospheric CO_2 seasonal cycle (Bent, 2014). We then use a set of transport models to convert the time derivative of this cycle into estimates of northern extratropical terrestrial CO_2 flux. We refer to this process as "bias correction."

To isolate tropospheric CO_2 signals, we define an upper cutoff of 300 hPa and remove any remaining observations with detectable stratospheric influence using the measured concentration of nitrous oxide (N_2O) and a cutoff value of 319 parts per billion (ppb) after detrending the data to 2009; samples whose N_2O concentration falls below this threshold are removed from the observations (Bent, 2014). We also manually removed outlying samples primarily obtained during takeoffs and landings to avoid strong local influences from biospheric exchange or fossil emissions. The flights and times filtered are identified in Data Sets S1 and S2 of Supporting Information S1. We filter output at the same locations and times for the transport model CO_2 mole fractions simulated along the flight tracks as discussed in Section 2.3. This stratospheric and local influence filtering removes 2.8% of the observations within the defined domain from the HIPPO and ATom data sets. We did not use observations from the northbound leg of HIPPO-1 because it only extended to 40°N, and both QCLS and OMS were filtered for altitude-dependent biases on these flights.

We then detrend the filtered data by removing the long-term trend in the NOAA Mauna Loa in situ CO_2 mole fraction record (Thoning et al., 2022), found by Seasonal-Trend decomposition using locally estimated scatterplot smoothing (STL, Cleveland & Cleveland, 1990) with a 2-year smoothing window. By detrending, the overall annual mean level of flux is removed, leaving only the (relative) seasonal cycle.

We calculate the extratropical mean drawdown by first aggregating the detrended data in latitude and pressure bins. We discretized the atmosphere into bins of 5° in latitude and 50 hPa in pressure for the latitude range 20°N to 90°N and the pressure range 300 hPa to 1,000 hPa. Observations at latitudes south of 20°N are excluded because of the differences in the phasing of the tropical seasonal cycle to that north of 20°N, and observations at pressures below 300 hPa were excluded because measurements were sparse and frequently in the stratosphere. Within each bin, we average all data collected for a given day of the year and then fit a second-order harmonic as a function of day of the year with an offset due to the difference in the annual mean relative to Mauna Loa (Figure S1 in Supporting Information S1). We then generate seasonal time series at the daily resolution from the harmonic fits and take the pressure-weighted average of these values for each latitude bin. These partial columns are then integrated over latitudes from 20°N to 90°N (Equation 1), using cosine(lat) weighting to reflect the influence on the zonal volume below 300 hPa at latitudes where observations were made. We call the result of this integration the "curtain average" concentration of atmospheric CO_2 (Bent, 2014).

$$\text{CurtainAvg} = \frac{\int_{20°N}^{90°N} \int_{300\,\text{hPa}}^{1,000\,\text{hPa}} \text{aveCO}_2(\varphi, P)\,\mathrm{d}p\,\cos\varphi\,\mathrm{d}\varphi}{\int_{20°N}^{90°N} \int_{300\,\text{hPa}}^{1,000\,\text{hPa}} \mathrm{d}p\,\cos\varphi\,\mathrm{d}\varphi} \qquad (1)$$

The curtain average is shown in black in Figure 2, and is compared to the northbound and southbound legs of each HIPPO and ATom mission where each point is found by filtering, detrending, interpolating and extrapolating to get a full altitude and latitude slice, then taking a pressure and cosine of latitude-weighted average (Akima, 1978).

The derivative, found as a finite-difference, of the curtain average concentration fit line with respect to time, then gives the rate of change in CO_2 of this atmospheric volume as a function of the day of the year.

We convert from a rate of change in dry air mole fraction to a rate of change in mass balance (MB) by multiplying the mole fraction by the mass of dry air north of 20°N and between the surface and 300 hPa in pressure. This mass
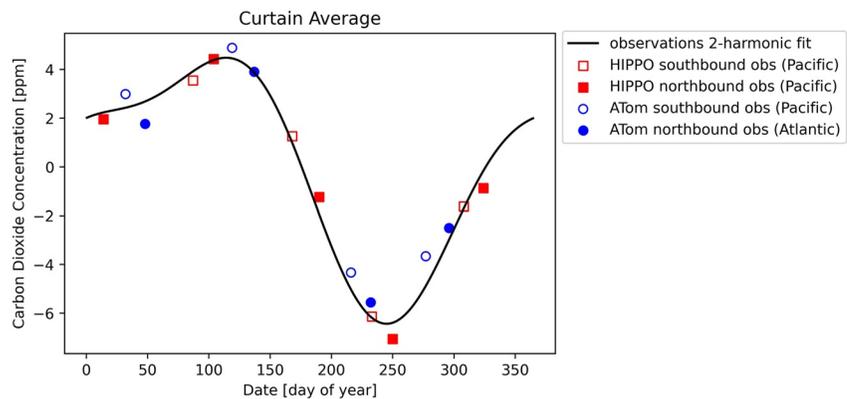
Figure 2. Two-harmonic fit to detrended average carbon dioxide concentration in ppm as a function of day of year for HIPPO and ATom flight campaigns in the atmospheric curtain between 20°N and 90°N in latitude and between 1,000 and 300 hPa in pressure. The points are found by filtering, detrending, interpolating and extrapolating to get a full altitude and latitude slice, then taking a pressure and cosine of latitude-weighted average. The black line is the average of all fits to individual latitude-pressure bins with the annual mean removed. The direction of flight (southbound or northbound) is shown with filled and unfilled symbols because southbound flights occurred 2–3 weeks earlier than northbound flights.

is found using the ERA5 reanalysis fields used by Tracer Model 5 (TM5) within CarbonTracker 2019 (Jacobson et al., 2020; Krol et al., 2005). We use the time mean mass of $1.21087452 \times 10^{18}$ kg within the domain across the HIPPO and ATom time periods, neglecting annual and seasonal variations, which are less than 0.2%.

2.3. Flux Estimates Using Atmospheric Transport Models

Although the HIPPO and ATom observations over the remote ocean provide representative estimates of background values, the sampling and discretization methods, zonal gradients, and mixing out of the domain result in differences between our MB time derivatives and zonal fluxes. Also, fossil fuel emissions and air-sea gas exchange make small contributions to the observed cycles. We use atmospheric transport models to account for the cumulative effects of (a) atmospheric mixing across the southern boundary and above the pressure boundary; (b) spatial sampling biases associated with specific flight tracks; (c) zonal sampling bias; (d) temporal sampling biases associated with synoptic variability, subseasonal sample distribution, and interannual variability; and (e) contributions from fossil-fuel emissions and ocean uptake.

Atmospheric inversions provide optimal estimates of surface-atmosphere CO_2 exchange derived from both atmospheric CO_2 mole fraction data and initial estimates for land-atmosphere and ocean-atmosphere exchange in the context of an atmospheric transport model. We use posterior concentrations from three different inversions (Table 2) to reduce the uncertainty that may arise due to biases present in the choice of transport model as differences in transport have previously been shown to lead to large differences in optimized fluxes (Gurney et al., 2002; Schuh et al., 2019; Stephens et al., 2007).

CarbonTracker is a data assimilation system consisting of the TM5 atmospheric transport model coupled to an ensemble Kalman filter (Jacobson et al., 2020; Peters et al., 2007). TM5 is a global two-way nested transport

Table 2
Inverse Models Used in This Study

	CT2019B	MIROC4-ACTM	CAMS
Years Available	2000–2018	1996–2018	1979–2020
Years Used	2009–2018	2009–2018	2009–2018
Transport	TM5	ACTM	LMDZ
Meteorology	ERA5	JRA55	ERA5
Resolution (lat × lon in degrees)	Glb2 × 3, N America 1 × 1	Glb2.8 × 2.8	Glb1.9 × 3.75
Fossil Fuels	Miller and ODIAC	EDGARv432	GCP-GridFEDv2021.2
Reference	Jacobson et al. (2020)	Chandra et al. (2022)	Chevallier et al. (2005)

model driven by 3-hr meteorological forcing from the ERA5 operational forecast model (Krol et al., 2005). We used the output from the most recent Carbon Tracker release (CT2019B, Jacobson et al., 2020), which includes optimized carbon fluxes through the HIPPO and ATom time period, and CO_2 mole fractions simulated along the flight paths for the HIPPO and ATom campaigns, which match the dates, times, and locations for the HIPPO and ATom data included in the GLOBALVIEWplus v5.0 ObsPack product (Cooperative Global Atmospheric Data Integration Project, 2019). CT2019B assimilated 460 time series data sets including data from the HIPPO and ATom campaigns. The data sets assimilated in CT2019B were mostly surface in situ, surface flask, and tower in situ observations from sites around the world.

The Model for Interdisciplinary Research on Climate version 4 atmospheric general circulation model based chemistry transport model (MIROC4-ACTM) provides posterior 4-D CO_2 fields and optimized surface fluxes through the HIPPO and ATom periods (Chandra et al., 2022). Atmospheric CO_2 transport in MIROC4-ACTM was simulated by the Japan Agency for Marine-Earth Science and Technology's ACTM, a transport model driven by meteorological parameters from the Japanese 55-year Reanalysis (JRA55, Patra et al., 2018). We used the 2020 version of MIROC4-ACTM output and MIROC CO_2 mole fractions simulated along the flight paths for the HIPPO and ATom campaigns matching the dates, times, and locations for the HIPPO and ATom data included in the GLOBALVIEWplus v5.0 ObsPack product. MIROC4-ACTM assimilated surface flask data from 50 sites around the world (Chandra et al., 2022).

We used a third set of inverse modeling output from the Copernicus Atmosphere Monitoring Service (CAMS; Chevallier et al., 2005). Within CAMS, transport of atmospheric CO_2 is simulated by the global climate model of the Laboratoire de Météorologie Dynamique, zoom capacity (LMDZ) driven by meteorological parameters from ECMWF (Chevallier et al., 2005). We used CO_2 mole fractions simulated along the HIPPO and ATom flight paths matching the dates, times, and locations for the HIPPO and ATom data included in the GLOBALVIEWplus v5.0 ObsPack product, and posterior carbon fluxes from CAMS v20r1, which contains output through the HIPPO and ATom period. CAMS v20r1 assimilated surface air-sample data from 159 sites around the world.

For each model, we calculate the annual cycle of the northern extratropical net land flux by removing the long-term annual mean of the posterior land flux, which excludes fossil fuel emissions and ocean fluxes, from 2000 to 2018 at each grid cell and then taking an area-weighted average north of 20°N. We linearly interpolate between monthly means to obtain an annual cycle at daily resolution to allow direct comparison to the aircraft observation-derived MB.

To correct for bias, we match the model ObsPack output date, time, and location to the 10-s merge files for HIPPO and ATom, then repeat the analysis described in Section 2.2 using posterior CO_2 mole fractions simulated along the HIPPO and ATom flight tracks to calculate the curtain average and MB for each atmospheric inversion system. The averaged posterior land flux is then subtracted from the MB to derive a seasonal correction. The MB found using posterior CO_2 mole fractions simulated along the HIPPO and ATom flight tracks is the MB that would be observed if transport and fluxes in our world perfectly matched transport and fluxes in the model, with a time delay between the two curves due to the time taken for the signal of land fluxes to reach the location where measurements are made. Thus, in each model, we assume that the difference between the MB for each model (solid lines in Figure 3a) and northern extratropical average posterior land flux (dashed lines in Figure 3a) is primarily due to mixing outside the domain, model fluxes, and time delay, with additional influences listed above. We determine the correction for each model and subtract it from the observationally derived MB (dotted black line in Figure 3b), resulting in transport model-specific flux estimates (solid color lines in Figure 3b). The average difference across the three models (dashed black line in Figure 3b) is subtracted from the observationally derived MB to estimate the seasonal cycle of the average net flux, hereafter referred to as the "flux cycle", into the atmosphere (solid black line in Figure 3b).

We find that the contribution of atmospheric transport uncertainty for the large spatial scale over which we average is small. In particular, noted variations in representations of vertical mixing (Schuh et al., 2019; Stephens et al., 2007) may change the distribution of CO_2 within our domain but not the domain average.

2.4. Growing Season Net Flux From Seasonal Flux Cycles

We then calculate the net atmospheric carbon exchange during the growing season, or GSNF, as the integral of the flux cycle during the growing season, defined to be when the bias-corrected flux cycle is negative, which
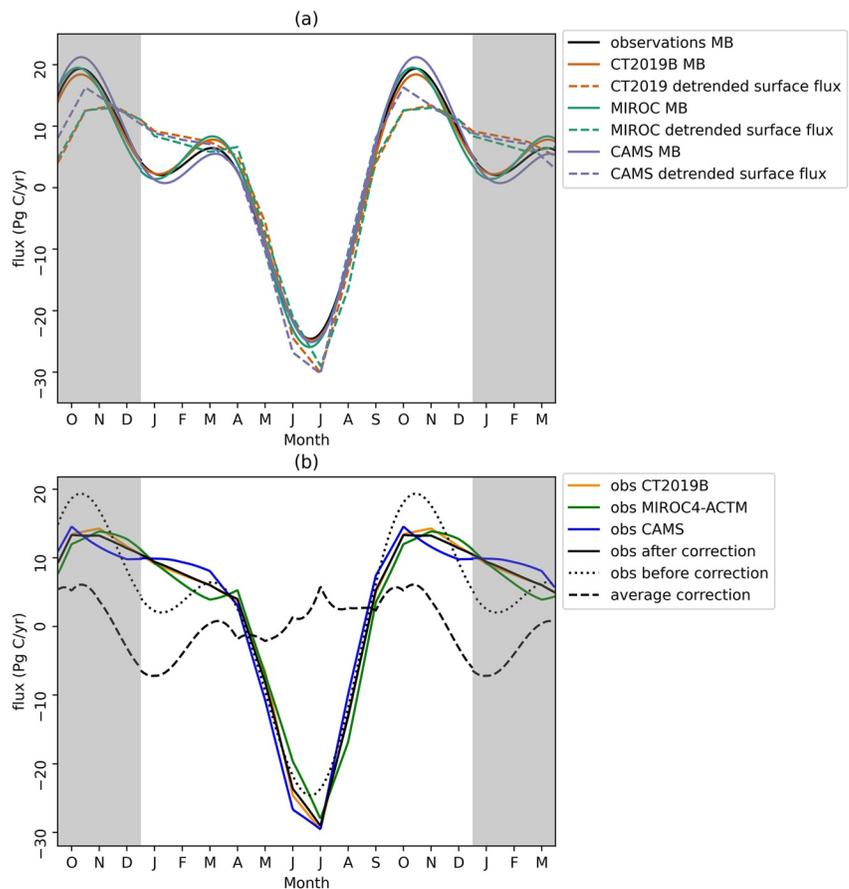
Figure 3. (a) Time derivative of concentrations from observations and inverse models along with model fluxes. The dashed lines show the area weighted average of posterior land fluxes from each inversion system in the domain 20°N–90°N. The solid lines are found by using the carbon dioxide mole-fractions along the flight track for each model to calculate the MB as described in Section 2.2. The solid black line is the time derivative of concentrations using the HIPPO and ATom observations. The estimated flux for the observations is bias corrected by finding the difference between the dotted and solid lines for a given model and applying that difference to the time derivative of the concentration. (b) Estimated flux after bias correction. The colored lines are found by calibrating using only the model indicated and the solid black line is found using the average correction. The dotted black line is the time derivative of concentration before the correction.

is nominally equivalent to the period when detrended atmospheric CO_2 is declining, primarily due to additional uptake by the biosphere as GPP outpaces respiration.

By detrending the observations, this estimate excludes the annual mean flux of CO_2, which itself includes fossil fuel emissions and terrestrial and oceanic sinks. Thus, our estimate of GSNF reflects the seasonal-only component of terrestrial exchange; the actual net uptake by the terrestrial biosphere during the growing season is larger when the annual component (long-term sink) is included. This approach is consistent with prior use of GSNF (e.g., Fung et al., 1983; Yang et al., 2007). Seasonal variations in fossil-fuel emissions and air-sea exchange contribute to seasonal variations in atmospheric CO_2, but these influences are small at about 3% and 5% of land exchange, respectively, on average for the three inversion systems, and have been removed by our use of the land flux in our model-based bias correction.

2.5. Earth System Models (CMIP5 and CMIP6)

The CMIP is an international, multi-model research intercomparison project whose purpose is to compare a coordinated set of simulations from ESMs in order to gain a better understanding of our ability to model climate change and associated feedback (Friedlingstein et al., 2006). The ESMs that participate in the CMIP simulate relevant physical, chemical, and biological processes within the coupled Earth system (Eyring et al., 2016) using

Table 3
Earth System Models Used in This Study

Model	Generation	Land component	Ocean component	Institution ID	Reference
ACCESS-ESM1-5	**CMIP6**	**CABLE2.4**	**ACCESS-OM2 (MOM5)**	**CSIRO**	**Ziehn et al. (2019)**
CanESM5	**CMIP6**	**CLASS3.6/CTEM1.2**	**NEMO3.4.1**	**CCCma**	**Swart et al. (2019)**
CESM2	**CMIP6**	**CLM5**	**POP2**	**NCAR**	**Danabasoglu (2019a)**
CESM2-FV2	CMIP6	CLM5	POP2	NCAR	Danabasoglu (2019b)
CESM2-WACCM	**CMIP6**	**CLM5**	**POP2**	**NCAR**	**Danabasoglu (2019c)**
CESM2-WACCM-FV2	CMIP6	CLM5	POP2	NCAR	Danabasoglu (2019d)
CMCC-CM2-SR5	**CMIP6**	**CLM4.5**	**NEMO3.6**	**CMCC**	**Lovato and Peano (2020)**
CMCC-ESM2	**CMIP6**	**CLM4.5**	**NEMO3.6**	**CMCC**	**Lovato et al. (2021)**
GISS-E2-1-G	CMIP6	GISS LSM	GISS	NASA-GISS	NASA/GISS (2018)
GISS-E2-1-H	CMIP6	GISS LSM	HYCOM	NASA-GISS	NASA/GISS (2019)
IPSL-CM6A-LR	**CMIP6**	**ORCHIDEE**	**NEMO-OPA**	**IPSL**	**Boucher et al. (2018)**
MPI-ESM-1-2-HAM	**CMIP6**	**JSBACH3.2**	**MPIOM1.63**	**HAMMOZ-Consortium**	**Neubauer et al. (2019)**
MPI-ESM1-2-LR	**CMIP6**	**JSBACH3.2**	**MPIOM1.63**	**MPI-M**	**Wieners et al. (2019)**
NorCPM1	**CMIP6**	**CLM4**	**MICOM1.1**	**NCC**	**Bethke et al. (2019)**
NorESM2-LM	**CMIP6**	**CLM**	**MICOM**	**NCC**	**Seland et al. (2019)**
NorESM2-MM	**CMIP6**	**CLM**	**MICOM**	**NCC**	**Bentsen et al. (2019)**
TaiESM1	**CMIP6**	**CLM4**	**POP2**	**AS-RCEC**	**Lee and Liang (2020), Lee et al. (2020)**
CanESM2	**CMIP5**	**CLASS2.7 and CTEM1**	**CanOM4 and CMOC1.2**	**CCCma**	**Arora et al. (2011)**
CCSM4	CMIP5	CLM4	POP2	NCAR	Gent et al. (2011)
CESM1-BGC	CMIP5	CLM	BEC	NSF-DOE-NCAR	Long et al. (2013)
GFDL-ESM2G	**CMIP5**	**LM3**	**TOPAZ**	**NOAA GFDL**	**Dunne et al. (2013)**
HadGEM2-CC	**CMIP5**	**MOSES2 and TRIFFID**	**HadGOM2**	**MOHC**	**Bellouin et al. (2011)**
HadGEM2-ES	**CMIP5**	**MOSES2 and TRIFFID**	**HadGOM2**	**MOHC**	**Bellouin et al. (2011)**
INM-CM4.0	CMIP5	-	-	INM	Volodin et al. (2010)
IPSL-CM5A-LR	**CMIP5**	**ORCHIDEE**	**ORCA2**	**IPSL**	**Dufresne et al. (2013)**
IPSL-CM5A-MR	**CMIP5**	**ORCHIDEE**	**ORCA2**	**IPSL**	**Dufresne et al. (2013)**
MIROC-ESM	**CMIP5**	**MATSIRO**	**COCO**	**MIROC**	**Watanabe et al. (2011)**
MIROC-ESM-CHEM	**CMIP5**	**MATSIRO**	**COCO**	**MIROC**	**Watanabe et al. (2011)**
NorESM1-M	**CMIP5**	**CLM**	**MICOM**	**NCC**	**Tjiputra et al. (2013)**

Note. Models in bold are included in the subset used to analyze gross primary productivity, heterotrophic respiration, and net primary productivity.

models developed by individual modeling centers worldwide, with the goal of including the most important processes that feed back into the climate system.

Here, we analyzed the historical simulations for CMIP5 and CMIP6 (Table 3) that span the period from 1850 to 2005 for CMIP5 and 2014 for CMIP6. We analyzed CO_2 concentration-driven historical simulations in which environmental forcing, such as greenhouse gas concentrations and solar forcing, are prescribed. Land and ocean fluxes are allowed to evolve prognostically in response to greenhouse gasses and other forcings (Eyring et al., 2016).

The northern extratropical seasonal land flux for the CMIP5 and CMIP6 models is calculated by removing the long-term annual mean and taking the area-weighted average of each model's NBP output north of 20°N, linearly interpolating between monthly mean values, similar to the method used for the inversion posterior fluxes (Section 2.3). In addition, we fit a second-order harmonic to the NBP seasonal cycle to find seasonal timing within the models. We define the growing season in each model as the period for which the model simulates net terrestrial uptake, which allows us to evaluate each model's growing season start and end dates against those inferred from the aircraft observations. In ESMs, NBP reflects the balance of gross photosynthetic uptake, ecosystem
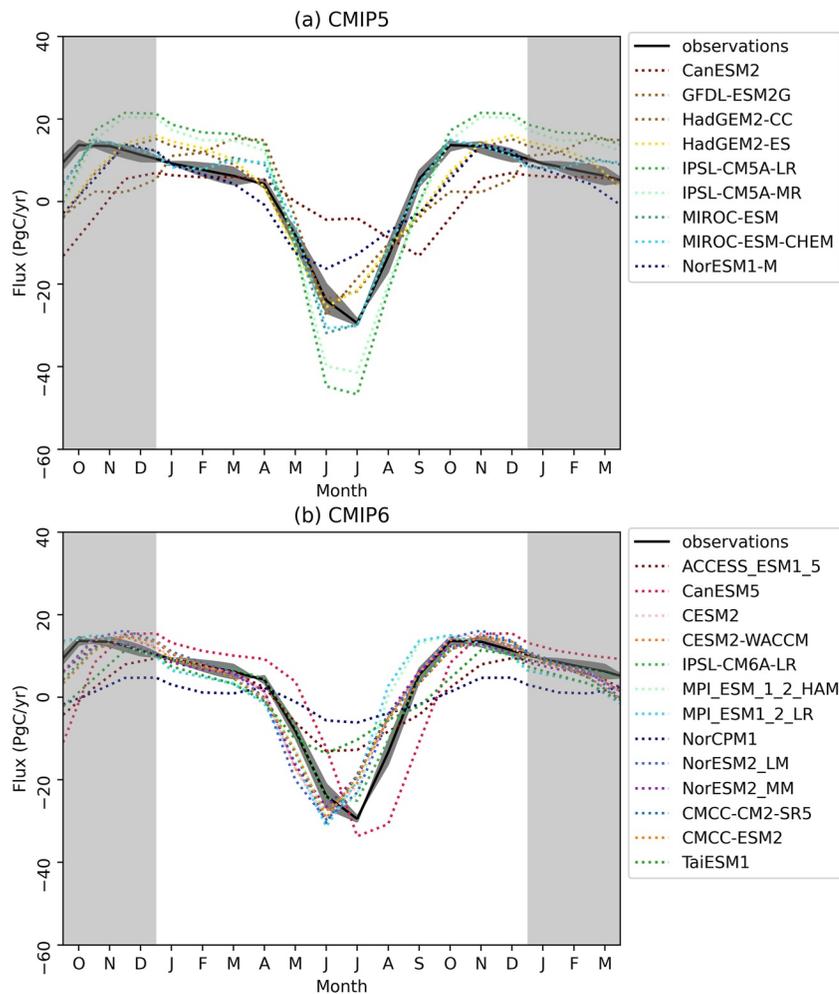
Figure 4. Corrected flux estimated from the HIPPO and ATom campaigns in comparison to area-weighted average NBP in the same domain from the (a) Coupled Model Intercomparison Project phase 5 (CMIP5) and (b) Coupled Model Intercomparison Project phase 6 (CMIP6) models. The bias corrected observation error is the standard deviation between correction using the three different inverse models. While the spread in magnitude and timing of the flux in the CMIP6 models is smaller than that of the CMIP5 models, there is still disagreement between the models.

respiration, and disturbance and harvest fluxes and corresponds to the land-atmosphere carbon exchange, making it comparable with our observationally derived flux. We note that ESMs generally do not represent lateral carbon fluxes in rivers, but we expect these to have a relatively minor contribution to our observed seasonal variations. We average multiple years of NBP output from the CMIP models to derive climatology; for CMIP6, we average over 2009–2014, and for CMIP5, we average over 2000–2005 because the output is not available through the HIPPO and ATom timeframes (Figure 4). We evaluate the ensembles as a whole by taking the median of each ensemble of models. We also evaluated area-weighted averages of three major component fluxes: GPP, RH, and NPP. Instead of analyzing the fluxes integrated over the growing season, we analyze the fluxes integrated over the entire year for GPP, RH, and NPP, which still correlate with the GSNF and are more useful than the seasonal fluxes when analyzing the carbon budget (e.g., Ballantyne et al., 2015; Tans et al., 1990). For the flux analysis, we use a subset of the models (9 of 12 CMIP5 models and 13 of 17 CMIP6 models, Table 3) for which historical NBP, GPP, RH, and NPP are available.

3. Results

Observations of atmospheric carbon dioxide from the HIPPO and ATom aircraft campaigns were used to estimate a GSNF of 5.7 ± 0.3 Pg C out of the atmosphere north of 20°N averaged over the period 2009–2018 (Figure 5).
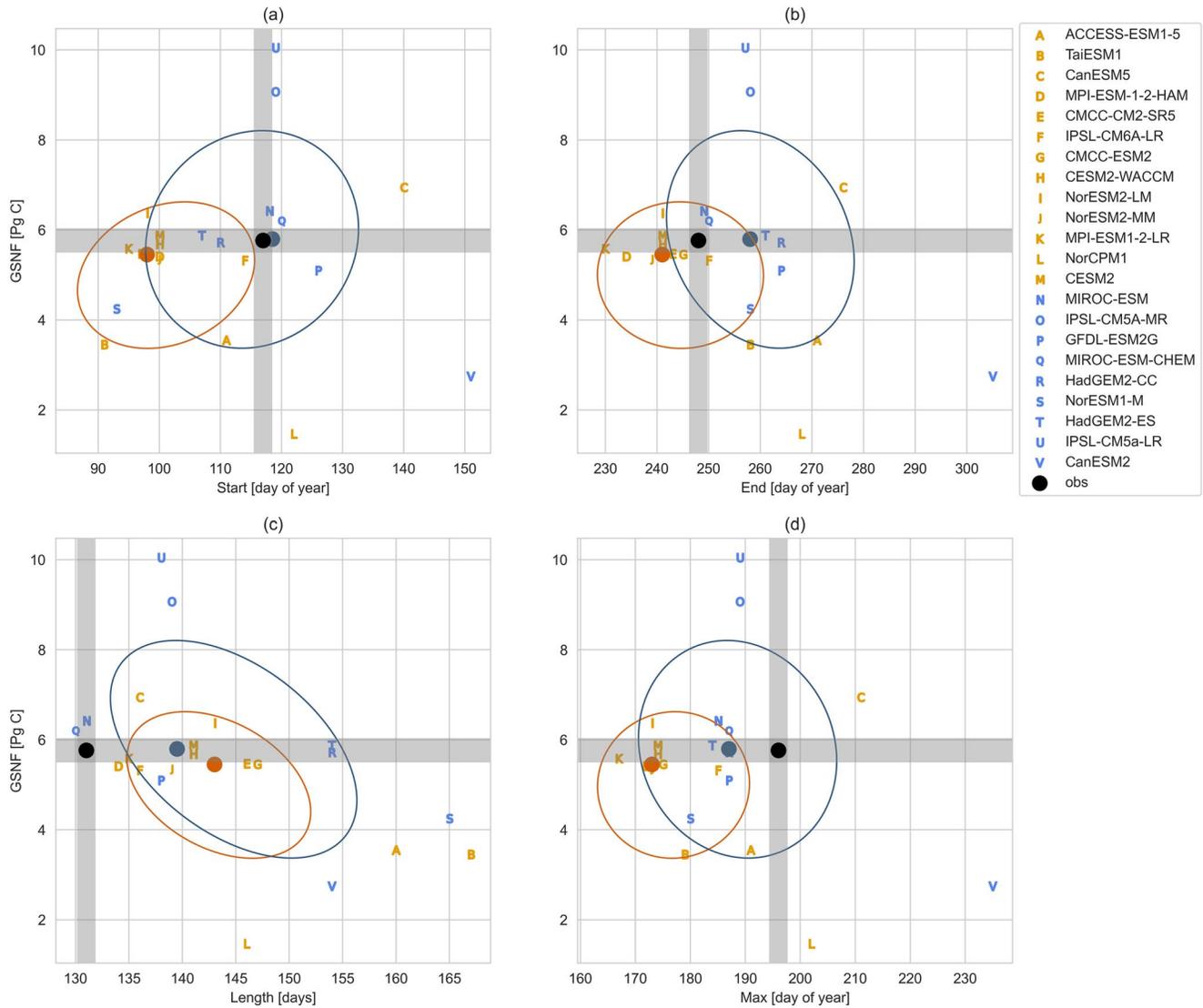
Figure 5. Growing season net flux plotted against the (a) start of the season, defined to be the first day when the seasonal component of atmospheric CO_2 is decreasing (seasonal component of flux changes from positive to negative), (b) end of the season, defined to be the last day when the seasonal component of atmospheric CO_2 is decreasing (seasonal component of flux changes from negative to positive), (c) length of the season, and (d) max of season, defined to be the day when flux is most negative. The black point is the number inferred from the observations with gray lines showing uncertainty. Coupled Model Intercomparison Project phase 5 (CMIP5) models are shown in blue and Coupled Model Intercomparison Project phase 6 (CMIP6) models are shown in orange. The blue and orange points are the multi-model mean for the CMIP5 and CMIP6 ensembles, respectively. The surrounding ellipses show the covariance to one standard deviation. Only models where gross primary productivity, heterotrophic respiration and net primary productivity output were available are included.

This value is equivalent to the net CO_2 exchange between the land and atmosphere during the growing season after removing the annual mean. The growing season is defined as the period when seasonal fluxes are negative (net uptake by land greater than the annual mean) and occurs between day 117 in late April and day 248 in early September. This corresponds to the day when the curtain average is maximum to the day when the curtain average is minimum (Figure 2). The flux cycle shows maximum uptake on day 196 (Figure 3b).

We conducted sensitivity tests to ensure that our choices for the latitudinal boundary of the domain, vertical boundary of the domain, and bin size did not have a large influence on the calculated GSNF. Expanding the region north (south) of 20°N by 5°N resulted in a decrease (increase) in the strength of the GSNF of just under 1%. The GSNF was also generally robust to the choice of pressure ceiling for the aircraft observations, increasing by just over 1% when we instead used 350 hPa as a ceiling. The relative standard deviation for all boundary combinations tested (all possible combinations of pressure cutoffs of 300, 325, 350, 375, and 400 hPa with
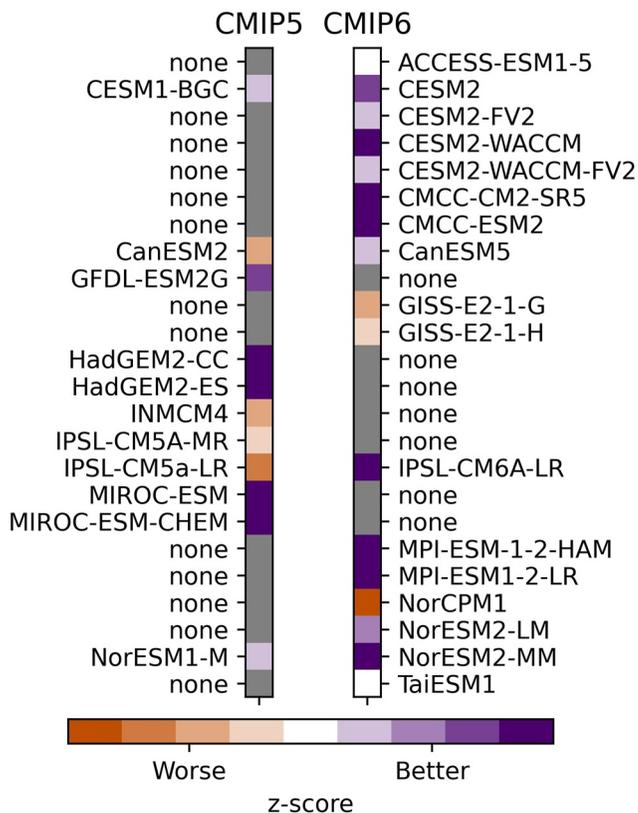
Figure 6. Coupled Model Intercomparison Project phase 5 and Coupled Model Intercomparison Project phase 6 model absolute values of *z*-score calculated for all models. The color gray and the label, none have been used when one generation of a model is not used or not existent.

latitude cutoffs of 20°N and 25°N) was 2%. Similarly, we saw just over a 1% increase in GSNF magnitude for a doubling of latitude or pressure bin size. The relative standard deviation for all bin sizes tested (all possible combinations of pressure bin sizes of 25, 50, and 100 hPa with latitude bin sizes of 5° and 10°) was 1%.

Given the small differences that the choice of boundary and bin size make on the magnitude of the calculated GSNF, most of the uncertainty in GSNF results from the transport model bias correction process. We see a spread of 0.2 Pg C or 4% when using CT2019B alone to bias correct versus using CAMS alone to bias correct. Additionally, the interannual variability, calculated as the standard deviation of the GSNF estimated from the average of the three inversion fluxes for each year over the period 2009–2018, is less than 0.2 Pg C or just over 3%. When adding the uncertainty from bin size, boundary choice, bias correction, and interannual variability in quadrature, assuming uncorrelated errors, the 1-sigma error of GSNF is 0.3 Pg C and on the start and end of the growing season is 2 days.

Considering the extensive altitude-latitude coverage of the aircraft observations, the inferred flux represents a unique and robust hemispheric estimate of terrestrial biosphere exchange and its seasonal phasing. We used the observationally inferred GSNF metric to evaluate NBP and its seasonal phasing from the CMIP5 and CMIP6 ensembles (Figure 5). CMIP6 models (mean GSNF: 5.3 ± 1.6 Pg C, range: 1.5–6.9 Pg C) on the whole have less spread than CMIP5 models (mean GSNF: 5.7 ± 2.4 Pg C, range: 2.5–10.0 Pg C) (Figure 5). Three of the 13 CMIP6 ensemble members evaluated, CESM2, CESM2-WACCM, and MPI-ESM1-2-LR, were within 0.3 Pg C of the observed value, and most of the models with a large bias underestimated the GSNF.

We note that some modeling centers showed substantial improvement in capturing GSNF between CMIP5 and CMIP6 (Figure 6) by decreasing the absolute value of their *z*-score, which indicates how many standard deviations away from the observed value a model falls. For example, the two versions of the IPSL model overestimate GSNF by more than 4 Pg C in CMIP5, which improved substantially in CMIP6 to underestimate the GSNF by less than 1 Pg C. In both CMIP5 and CMIP6, the majority of ensemble members underestimate the seasonal flux, falling below the horizontal gray bar in Figure 5.

A decrease in GSNF model spread in the newer generation has not necessarily led to an improved agreement between models and observations on the phasing of the seasonal cycle. For example, The CMIP5 ensemble median start day (Julian day 119 ± 17) is closer to the observed start day (Julian day 117 ± 2) than the CMIP6 ensemble mean start day (Julian day 98 ± 14) (Figure 5a). None of the 13 CMIP6 models evaluated fell within 2 days of the observed start day, while three of the 9 CMIP5 models evaluated fell within the 2-day uncertainty range (Figure 5a). These observations suggest that the growing season onset has become more biased in CMIP6. However, the CMIP6 ensemble did show a smaller bias than CMIP5 for the model-median end day, which was Julian day 241 ± 16 for CMIP6 and 258 ± 18 for CMIP5, in contrast to day 248 ± 2 in the observations (Figure 5b). At 143 ± 9 days, the CMIP6 ensemble median growing season length is 12 days longer than the observed length of 131 ± 2. In comparison, the CMIP5 ensemble median growing season length is 11 days longer at 140 ± 12 days (Figure 5c). While the CMIP6 ensemble median end day compares more favorably with the observed end day than does the CMIP5 ensemble median end day, in both cases, the simulated growing season is longer than what is observed (Figure 5c).

Seasonal phasing in general does not appear to be correlated with GSNF in the model ensembles, suggesting that phasing is not a dominant driver of GSNF spread among models. No correlation was seen ($r^2 < 0.2$, $p > 0.05$) between GSNF magnitude and the start day, end day, length, or max day across models (Figure 5). This suggests that factors other than the phasing of the growing season may explain inter-model differences.
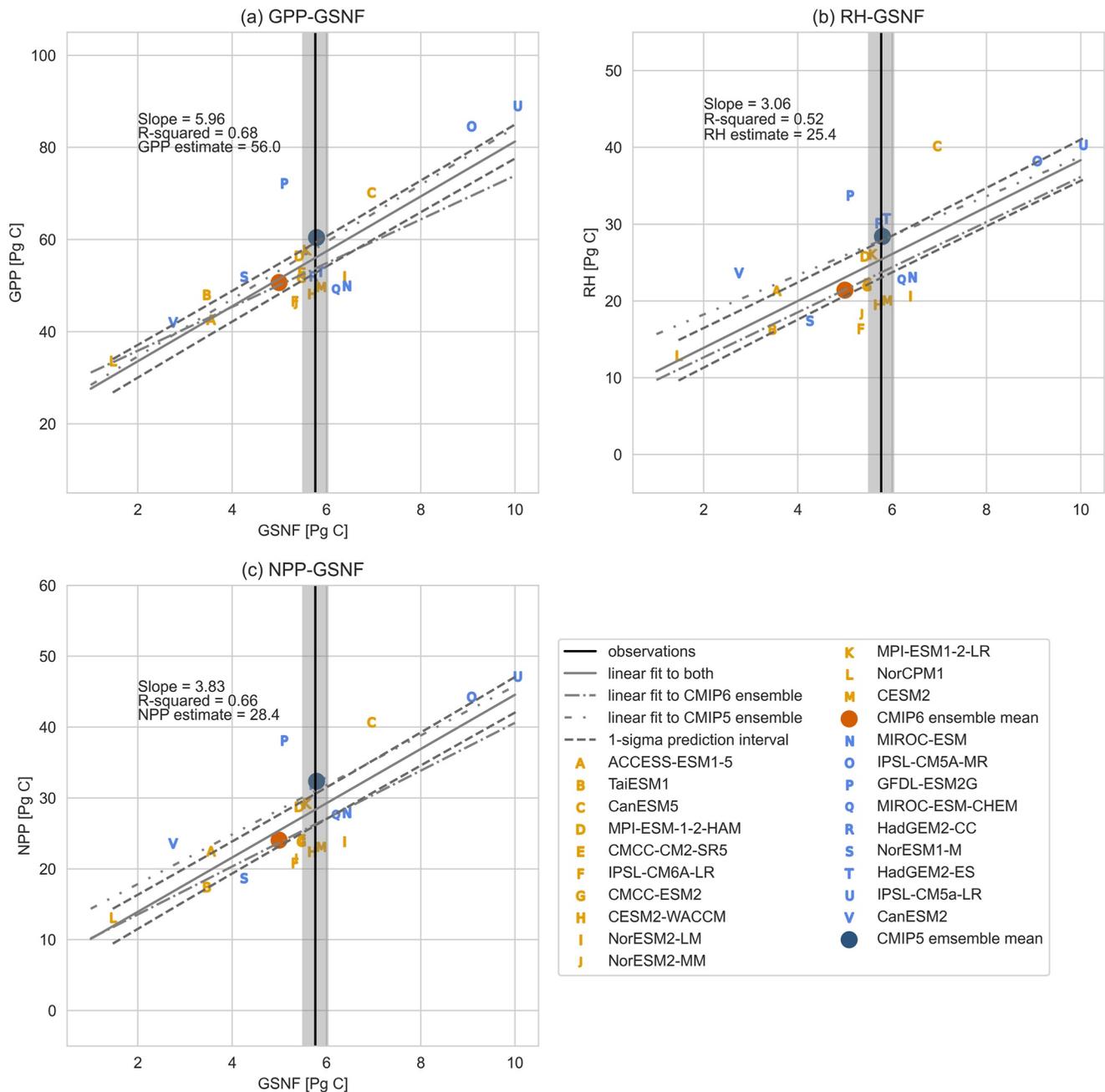
Figure 7. Growing season net flux (GSNF) plotted as a function of (a) integrated gross primary productivity (GPP) and (b) integrated heterotrophic respiration (RH), and (c) integrated net primary productivity (NPP) for the Coupled Model Intercomparison Project phase 5 (CMIP5) and Coupled Model Intercomparison Project phase 6 (CMIP6) models. The estimated GSNF from the HIPPO and ATom observations is shown in gray, CMIP5 models are shown in blue, and CMIP6 models are shown in orange. Only models where GPP, RH, and NPP data were available were included.

We analyzed GPP, RH, and NPP for the subset of CMIP5 and CMIP6 ensemble members, which include these outputs to determine if these component fluxes might explain model disagreement on GSNF (Figure 7). We found that in both the CMIP5 and CMIP6 ensembles, GSNF was correlated ($r^2 = 0.68$, $p < 0.05$) with GPP, where models with larger GPP generally had larger GSNF. Models with large GPP also tend toward higher respiration values with an r^2 value of 0.77 between GPP and RH, as GPP provides the inputs to support RH. RH showed a weaker correlation with GSNF than did GPP; however, the correlation is still moderately strong ($r^2 = 0.52$, $p < 0.05$). As expected, models with higher GPP values also tend toward higher NPP values, and the correlation between NPP and GSNF is moderately strong with an r^2 value of 0.66 and a p-value less than 0.05.

We see a large range for GPP, RH, and NPP across the CMIP5 and CMIP6 ensembles. The range in GPP is smaller for the CMIP6 ensemble (median = 51 $\pm$ 9 Pg C, spread = 37 Pg C) than for the CMIP5 ensemble (median = 60 $\pm$ 17 Pg C, spread = 47 Pg C) where the error in the median is one standard deviation. However, this decrease in the model range between ensemble generations is not seen for RH or NPP. The spreads for RH were 23 Pg C (median = 27 + 8 Pg C) for CMIP5 and 27 Pg C (median = 21 $\pm$ 7 Pg C) for CMIP6. NPP from CMIP5 and CMIP6 had median values of 32 $\pm$ 9 and 24 $\pm$ 7 Pg C, respectively, and a spread of 28 Pg C for both CMIP5 and CMIP6. We note that two models in the CMIP6 ensemble have GSNF values consistent with the observational constraint, MPI-ESM1-2-LR and IPSL-CM6A-LR, but the GPP values spanned by these models are over 10 Pg C, the RH values spanned by these models are 10 Pg C, and the NPP spanned by these models are 9 Pg C.

The strong correlations and large ensemble spread enabled us to indirectly constrain northern extratropical GPP, RH, and NPP through an "emergent constraint" (EC) approach (e.g., Eyring et al., 2019; Simpson et al., 2021; M. S. Williamson et al., 2021). ECs are correlations between some observable element X that varies across the ESM ensemble and some important variable Y assuming a physically meaningful relationship exists between X and Y. Here, we assumed GSNF to be X and assumed a physically meaningful relationship between the net flux and its component fluxes GPP, RH, and NPP (Figure 7). The CMIP5 and CMIP6 ensembles can be analyzed separately to estimate GPP, RH, and NPP. The CMIP5 ensemble EC estimate for GPP is 58 $\pm$ 23 Pg C, for RH is 28 $\pm$ 13 Pg C, and for NPP is 31 $\pm$ 12 Pg C where the error is the 95% prediction interval from a hypothetical sample generated using a Monte Carlo simulation of $n = 10,000$ and assuming a Gaussian distribution (D. B. Williamson & Sansom, 2019). Similarly, the CMIP6 ensemble EC estimate for GPP is 54 $\pm$ 13 Pg C, for RH is 24 $\pm$ 13 Pg C, and for NPP is 26 $\pm$ 10 Pg C. However, we can also combine the two ensembles to create a larger sample size as the estimates agree within error and the modeled relationship does not change substantially between the two generations. The combined EC estimate for GPP is 56 $\pm$ 17 Pg C, for RH is 25 $\pm$ 13 Pg C, and for NPP is 28 $\pm$ 11 Pg C. Although we did not constrain autotrophic respiration (RA) explicitly, the constraint on GPP and NPP imply RA is 28 $\pm$ 20 Pg C where the error has been propagated by summing the errors in quadrature. Applying an emergent constraint explicitly to RA would likely give smaller errors.

4. Discussion

Here we derive an observational constraint on northern extratropical GSNF from two novel aircraft campaigns that measured the atmospheric CO_2 curtain over remote oceans. Our estimated GSNF of 5.7 $\pm$ 0.3 Pg C is significantly smaller in magnitude than the GSNF of 7.9 Pg C yr^{-1} out of the atmosphere estimated north of 30°N by Yang et al. (2007) based on total column observations and spatially sparse aircraft profiles in North America. These differences are most likely due to differences in methodology, as Yang et al. used a combination of total column and aircraft measurements at 8 locations between 30°N and 70°N to scale fluxes from a terrestrial ecosystem model (TEM), while we used aircraft data that were sampled nearly continuously between our cutoff of 20°N and roughly 87°N to estimate GSNF without using a specific TEM. When our cutoff was changed from 20°N to 30°N, our estimate increased to 5.8 Pg C. Fung et al. (1983) quantified GSNF in the Northern Hemisphere using a three-dimensional tracer transport model. When our cutoff is changed to the Equator, our estimate increases to 6.1 Pg C, which falls within the 3.4–10.7 Pg C range given by Fung et al. (1983). However, the range given by Fung et al. (1983) is large. The uncertainty on northern hemisphere GSNF could likely be reduced with the application of multiple transport models to existing records.

We may expect to see an increase in GSNF over time due to the observed increase in the CO_2 seasonal cycle amplitude (SCA). Graven et al. (2013) saw an increase of 32%–59% in the NH SCA in the 50-year period from 1958 to 1963 to 2009–2011. This equates to an increase of 0.56%–0.93% yr^{-1}. If this trend continued through the HIPPO and ATom time periods, we would expect to see a 4.0%–6.7% increase across the 7 years between these two missions. Applying the methodology to HIPPO alone, which covers 2009–2011, gives a value of 11.2 ppm for the SCA of the curtain average versus 10.7 ppm when using ATom alone, which covers 2016–2018. This corresponds to a more than 1% decrease contrary to the increase expected from Graven. However, this may not be accurate as there may not be enough data in either mission alone to fully constrain the seasonal cycle (Figure S3 in Supporting Information S1) and account for interannual variability in CO_2 fluxes, which were shown by Jin et al. (2021) to be non-negligible.

Despite any interannual variability, the expected increase in GSNF is not apparent when comparing the flight-based estimates to previous estimates. Yang et al. (2007) estimated a larger GSNF despite covering an earlier time

period. The large range given by Fung et al. (1983) means that the increase in GSNF may be seen if the true GSNF in 1982 fell within the range of 4.6–5.2 Pg C. Looking at surface records, Keeling and Graven (2021) saw an increase in the amplitudes at Mauna Loa and Barrow between 2010 and 2017 equating to 10.9% at Mauna Loa and 1.1% at Barrow using 5-year running means, suggesting possibly a shift to less SCA growth at high latitudes. These results should be interpreted with caution as large interannual variability may dominate when looking at the short time period of 2010–2017.

The HIPPO and ATom observations reflect the atmospheric mass of carbon integrated over a large latitudinal and altitudinal extent; thus, our estimate is less sensitive to specific representations of atmospheric transport than are other estimates. For example, the GSNF estimated over our domain, bias corrected with the three inverse models individually, only varied by just 0.2 Pg C or less than 4% (Figure 3b), even while previous research has reported the relative spread of 10 simulations for the posterior annual mean northern extratropical land flux to be 13% (Gaubert et al., 2019). Additionally, inverse models seem to be converging on the land flux as the GSNF estimated from the posterior inversion land fluxes averaged 5.8 Pg C, slightly larger than the observationally based estimate of 5.7 Pg C and varied by just 0.2 Pg C. The small differences among the inverse models at the hemispheric scale mean that, even though it represented the largest source of uncertainty, the bias correction process imparted minimal uncertainty on our ultimate GSNF value.

Use of the transformed coordinate, $M_{\theta e}$, introduced by Jin et al. (2021) may further reduce uncertainty. $M_{\theta e}$ is defined as the dry air mass under a given equivalent potential temperature surface within a hemisphere and its relationship with latitude and altitude is nearly fixed over the seasonal cycle. When integrating to find the curtain average, Jin et al. (2021) show that using the transformed coordinate as an alternative to latitude reduces error in the curtain average due to sparse sampling and synoptic variability.

The low uncertainty on the hemispheric integral makes GSNF and its phasing robust targets for evaluating TEMs and the land components of ESMs. Although direct comparison with atmospheric CO_2 mole fraction has been used to evaluate ESMs previously, these comparisons rely on simulation of the three-dimensional atmospheric CO_2 mole fraction within the ESM (e.g., Keppel-Aleks et al., 2013) or require estimating CO_2 using an offline transport model or operator (e.g., Liptak et al., 2017). In contrast, our GSNF constraint can be used to evaluate ESMs at the flux level rather than relying on comparisons at the concentration level. Compared to our metric, the CMIP6 ensemble has shown some improvements relative to the CMIP5 ensemble, namely a reduction in the spread between models (5.5 Pg C vs. 7.3 Pg C) and a more favorable simulation of the timing of the end of the growing season (7 days early vs. 10 days late) when considering the ensemble median values (Figures 5b and 5c). However, there are still large disagreements, and in some ways, the CMIP5 models perform better in relation to the observationally inferred flux than the CMIP6 models. For example, CMIP5 models outperform CMIP6 models in simulating the start of the growing season (2 days late vs. 19 days early) and the GSNF magnitude (5.7 ± 2.4 vs. 5.3 ± 1.6 Pg C compared to observational value of 5.7 ± 0.3 Pg C) when considering the ensemble median values (Figure 5). The models tend to underestimate the magnitude of GSNF on the whole with 3 of the 9 CMIP5 models and 8 of the 13 CMIP6 models underestimating the flux (falling below the horizontal gray lines in Figure 5), only 2 CMIP5 and 3 CMIP6 models falling within the uncertainty range, and 4 CMIP5 models and 2 CMIP6 models overestimating the net flux. This is consistent with previous results from Keppel-Aleks et al. (2013) suggesting that the Community Land Model version 4 (CLM4), a version of which is used as the land model for over one-third of the ESMs evaluated, underestimated GSNF. Three of the models with the largest underestimate of GSNF (NorCPM1, NorESM1-M, and TaiESM1) utilize CLM4 as their land model; however, we note that two of the best performing models (CESM2 and CESM2-WACCM) both utilize the Community Land Model version 5 (CLM5) as their land component, reflecting major improvement in CLM5 compared to CLM4 in simulating CO_2 seasonality (Lawrence et al., 2019). Correlations between GSNF and seasonal phasing were weak or nonexistent, meaning that the improvements in CLM5 are more likely due to improvements in simulations of the overall magnitude of photosynthesis and respiration than to improvements in simulations of the timing of the growing season.

The methods chosen to define the growing season have some effect on model-observation comparisons. We could have chosen to prescribe the growing season dates for calculating model metrics from the observations rather than allowing the dates to vary from model to model. The primary difference in the results using this approach is a slight reduction in the magnitude of the GSNF for most of the CMIP5 and CMIP6 models, a less strong correlation between GSNF and GPP, RH, and NPP, and a slight increase in the inferred values of northern GPP, RH, and NPP (Figure S2 in Supporting Information S1).

Although the aircraft CO_2 data provide a constraint on GSNF, this benchmark is also useful for constraining component fluxes that are difficult to infer observationally at large spatial scales. We examine the GPP, RH, and NPP of the ESM ensembles in relation to GSNF. GPP, RH, and NPP are components of the net flux, and we see moderate correlations between simulated GSNF and simulated productivity (both GPP and NPP) and RH. It is noteworthy that correlations between GSNF and the various annual fluxes are stronger than those between GSNF and seasonal timing, which may indicate that the magnitudes of photosynthesis and respiration are more dominant drivers of GSNF than seasonal phasing (Baldocchi et al., 2018; Valentini et al., 2000). Cadule et al. (2010) analyzed three CMIP models and concluded that models generally underestimate the seasonal amplitude due to shortcomings in vegetation phenology and RH response to climate. Our results generally support this role for discrepancies in component fluxes as driving a discrepancy in the resulting GSNF.

These correlations between GSNF and its component fluxes provide an opportunity to constrain GPP, RH, and NPP using an emergent constraint approach. Previous estimates of the northern extratropical GPP are highly uncertain, with large disagreements between estimates. For example, Mao et al. (2012) used the Moderate Resolution Imaging Spectroradiometer (MODIS) GPP product to estimate Northern Hemisphere GPP averaged over 2000–2009 as 64.75 ± 0.97 Pg C yr^{-1}. In contrast, the FLUXCOM product, based on upscaled observations from FLUXNET sites using various machine learning approaches (Jung et al., 2019), estimated GPP north of 20°N averaged over 2009–2014 to be 48 Pg C yr^{-1}. Our GSNF-constrained value of 56 ± 15 Pg C for GPP is consistent with both of these estimates, given the larger error bars on the emergent constraint approach, and indicates that GPP falls between these two estimates. The large value for GPP relative to FLUXCOM is interesting in light of the uncertainty in global GPP. The range in global mean GPP magnitudes for 2008–2010 from FLUXCOM members is 106–130 Pg C yr^{-1} (Jung et al., 2020). This range covers the observation-based estimate of global mean GPP of 123 ± 8 Pg C yr^{-1} found using eddy covariance flux data and various diagnostic models (Beer et al., 2010) but is smaller than the GPP magnitudes of 150–175 Pg C yr^{-1} derived from an isotope-based study (Welp et al., 2011). Our results may indicate a higher global mean GPP than the flux tower upscaling yields; however, we are constraining extratropical GPP, which may be affected by different factors than tropical GPP. Therefore, these results need to be interpreted with caution.

5. Conclusions

ESMs disagree in their simulation of large-scale carbon fluxes, making it crucial to evaluate models to contextualize their climate predictions. We have presented an approach to constrain the seasonal land flux using aircraft data from the HIPPO and ATom flight campaigns. The northern extratropical GSNF of 5.7 ± 0.3 Pg C derived from these observations can be used to evaluate prognostic model fluxes for net flux directly and for component fluxes via an emergent constraint approach. We note that this constraint is tied to atmospheric transport models because the flux is bias corrected by comparing posterior mole fractions from inverse models with their fluxes. However, we found that at the hemispheric scale, the constraint is robust to the choice of atmospheric transport model since common transport errors tend to cancel out at this scale.

When compared to the CMIP5 and CMIP6 models, the inferred GSNF suggests a larger net flux and shorter growing season than those simulated in both model ensembles. This gives modelers an additional observation to target during model development and could be added to a benchmarking system such as the International Land Model Benchmarking (ILAMB) System (Collier et al., 2018). While there is decreased model spread between CMIP6 models, this benchmark also highlights some of the ways in which CMIP6 models have not improved from CMIP5, such as in simulating the start of the growing season.

Correlations within the CMIP5 and CMIP6 ensembles allowed us to apply an emergent constraint approach to estimate northern hemisphere GPP, RH, and NPP. We found that the GSNF-constrained value for GPP is at the higher end of a range of estimates from FLUXCOM, a commonly used ensemble of data products of upscaled biosphere-atmosphere fluxes.

Overall, the HIPPO and ATom inferred GSNF provides a robust metric that allows for the evaluation of large-scale fluxes in flux space and sheds light on component fluxes, filling a need highlighted by Collier et al. (2018) for land model benchmarking. More regular global scale airborne tomography could resolve GSNF at higher time resolution and leverage carbon cycle interannual variability for improved tests of ESM process representations.

Data Availability Statement

HIPPO data are available at the Earth Observing Laboratory (EOL) HIPPO data archive at http://www.eol.ucar.edu/projects/hippo/. (http://doi.org/10.3334/CDIAC/HIPPO_010). ATom data are available at the Oak Ridge National Laboratory DAAC at https://daa.ornl.gov/ATOM/campaign/ (https://doi.org/10.3334/ORNLDAAC/1925). CarbonTracker CT2019B results provided by NOAA ESRL, Boulder, Colorado, USA from the website at http://carbontracker.noaa.gov (https://doi.org/10.25925/20201008). MIROC4-ACTM inversion results are available on Zenodo: https://doi.org/10.5281/zenodo.5776197. CAMS inversion results were provided by the Copernicus Atmosphere Monitoring Service, part of the European Union's Earth observation programme, from the website at https://ads.atmosphere.copernicus.eu/. CMIP5 and CMIP6 model output files are freely available from the Earth System Grid Federation data replication centers and used under creative commons license CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0/.

References

Akima, H. (1978). A method of Bivariate interpolation and smooth surface fitting for irregularly distributed data points. *ACM Transactions on Mathematical Software*, *4*(2), 148–159. Copyright 1978, Association for Computing Machinery, Inc., reprinted by permission. https://doi.org/10.1145/355780.355786

Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., et al. (2020). Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models. *Biogeosciences*, *17*(16), 4173–4222. https://doi.org/10.5194/bg-17-4173-2020

Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., et al. (2011). Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophysical Research Letters*, *38*(5), 3–8. https://doi.org/10.1029/2010GL046270

Baldocchi, D., Chu, H., & Reichstein, M. (2018). Inter-annual variability of net and gross ecosystem carbon fluxes: A review. *Agricultural and Forest Meteorology*, *249*, 520–533. https://doi.org/10.1016/j.agrformet.2017.05.015

Ballantyne, A. P., Alden, C. B., Miller, J. B., Tans, P. P., & White, J. W. C. (2012). Increase in observed net carbon dioxide uptake by land and oceans during the past 50 years. *Nature*, *488*(7409), 70–72. https://doi.org/10.1038/nature11299

Ballantyne, A. P., Andres, R., Houghton, R., Stocker, B. D., Wanninkhof, R., Anderegg, W., et al. (2015). Audit of the global carbon budget: Estimate errors and their impact on uptake uncertainty. *Biogeosciences*, *12*(8), 2565–2584. https://doi.org/10.5194/bg-12-2565-2015

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., et al. (2010). Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science*, *329*(5993), 834–838. https://doi.org/10.1126/science.1184984

Bellouin, N., Collins, W. J., Culverwell, I. D., Halloran, P. R., Hardiman, S. C., Hinton, T. J., et al. (2011). The HadGEM2 family of met office unified model climate configurations. *Geoscientific Model Development*, *4*(3), 723–757. https://doi.org/10.5194/gmd-4-723-2011

Bent, J. (2014). *Airborne oxygen measurements over the Southern Ocean as an integrated constraint of seasonal biogeochemical processes, PhD thesis*. University of California.

Bentsen, M., Oliviè, D. J. L., Seland, Ø., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2019). NCC NorESM2-MM model output prepared for CMIP6 CMIP historical, version 20191126 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.8040

Bethke, I., Wang, Y., Counillon, F., Kimmritz, M., Fransner, F., Samuelsen, A., et al. (2019). NCC NorCPM1 model output prepared for CMIP6 CMIP historical, version 20190921 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.10894

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M. A., et al. (2018). IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP historical, version 20180711 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.5195

Cadule, P., Friedlingstein, P., Bopp, L., Sitch, S., Jones, C. D., Ciais, P., et al. (2010). Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements. *Global Biogeochemical Cycles*, *24*(2), GB2016. https://doi.org/10.1029/2009GB003556

Canadell, J. G., Monteiro, P. M. S., Costa, M. H., Da Cunha, L. C., Cox, P. M., Alexey, V., et al. (2021). Global carbon and other biogeochemical cycles and feedbacks. In L. G. Masson-Delmotte (Ed.), *Climate change 2021: The physical science basis. Contribution of WG I to the sixth assessment report of the IPCC*. Cambridge University Press.

Chandra, N., Patra, P. K., Niwa, Y., Ito, A., Iida, Y., Goto, D., et al. (2022). Estimated regional CO_2 flux and uncertainty based on an ensemble of atmospheric CO_2 inversions. *Atmospheric Chemistry and Physics*, *22*(14), 9215–9243. https://doi.org/10.5194/acp-22-9215-2022

Chevallier, F., Fisher, M., Peylin, P., Serrar, S., Bousquet, P., Bréon, F. M., et al. (2005). Inferring CO_2 sources and sinks from satellite observations: Method and application to TOVS data. *Journal of Geophysical Research*, *110*(24), 1–13. https://doi.org/10.1029/2005JD006390

Ciais, P., Rayner, P., Chevallier, F., Bousquet, P., Logan, M., Peylin, P., & Ramonet, M. (2010). Atmospheric inversions for estimating CO_2 fluxes: Methods and perspectives. *Climatic Change*, *103*(1–2), 69–92. https://doi.org/10.1007/s10584-010-9909-3

Cleveland, R. B., & Cleveland, W. S. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, *6*, 3–33.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The international land model benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018MS001354

Cooperative Global Atmospheric Data Integration Project. (2019). Multi-laboratory compilation of atmospheric carbon dioxide data for the period 1957–2018; obspack_co2_1_GLOBALVIEWplus_v5.0_2019_08_12 [Dataset]. NOAA Earth System R. https://doi.org/10.25925/20190812

Danabasoglu, G. (2019a). NCAR CESM2 model output prepared for CMIP6 CMIP historical, version 20190116 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.7627

Danabasoglu, G. (2019b). NCAR CESM2-FV2 model output prepared for CMIP6 CMIP historical, version 20190731 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.11297

Danabasoglu, G. (2019c). NCAR CESM2-WACCM model output prepared for CMIP6 CMIP historical, version 20190131 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.10071

Danabasoglu, G. (2019d). NCAR CESM2-WACCM-FV2 model output prepared for CMIP6 CMIP historical, version 20190831 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.11298

Daube, B. C., Boering, K. A., Andrews, A. E., & Wofsy, S. C. (2002). A high-precision fast-response airborne CO_2 analyzer for in situ sampling from the surface to the Middle stratosphere. *Journal of Atmospheric and Oceanic Technology*, *19*(10), 1532–1543. https://doi.org/10.1175/1520-0426(2002)019<1532:AHPFRA>2.0.CO;2

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., et al. (2013). Climate change projections using the IPSL-CM5 Earth system model: From CMIP3 to CMIP5. *Climate Dynamics*, *40*(9–10), 2123–2165. https://doi.org/10.1007/s00382-012-1636-1

Dunne, J. P., John, J. G., Shevliakova, E., Stouffer, R. J., Krasting, J. P., Malyshev, S. L., et al. (2013). GFDL's ESM2 global coupled climate–carbon Earth system models. Part II: Carbon system formulation and baseline simulation characteristics. *Journal of Climate*, *26*(7), 2247–2267. https://doi.org/10.1175/JCLI-D-12-00150.1

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Fernández-Martínez, M., Sardans, J., Chevallier, F., Ciais, P., Obersteiner, M., Vicca, S., et al. (2019). Global trends in carbon sinks and their relationships with CO_2 and temperature. *Nature Climate Change*, *9*(1), 73–79. https://doi.org/10.1038/s41558-018-0367-7

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., et al. (2006). Climate-carbon cycle feedback analysis: Results from the C4MIP model intercomparison. *Journal of Climate*, *19*(14), 3337–3353. https://doi.org/10.1175/JCLI3800.1

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., et al. (2022). Global carbon budget 2021. *Earth System Science Data*, *14*(4), 1917–2005. https://doi.org/10.5194/essd-14-1917-2022

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, *27*(2), 511–526. https://doi.org/10.1175/JCLI-D-12-00579.1

Friend, A. D., Arneth, A., Kiang, N. Y., Lomas, M., Ogée, J., Rödenbeck, C., et al. (2007). FLUXNET and modelling the global carbon cycle. *Global Change Biology*, *13*(3), 610–633. https://doi.org/10.1111/j.1365-2486.2006.01223.x

Fung, I., Prentice, K., Matthews, E., Lerner, J., & Russell, G. (1983). Three-dimensional tracer model study of atmospheric CO_2: Response to seasonal exchanges with the terrestrial biosphere. *Journal of Geophysical Research*, *88*(C2), 1281–1294. https://doi.org/10.1029/JC088iC02p01281

Fung, I. Y., Doney, S. C., Lindsay, K., & John, J. (2005). Evolution of carbon sinks in a changing climate. *Proceedings of the National Academy of Sciences*, *102*(32), 11201–11206. https://doi.org/10.1073/pnas.0504949102

Gaubert, B., Stephens, B. B., Basu, S., Chevallier, F., Deng, F., Kort, E. A., et al. (2019). Global atmospheric CO_2 inverse models converging on neutral tropical land exchange but diverging on fossil fuel and atmospheric growth rate. *Biogeosciences Discussions*, 1–25. https://doi.org/10.5194/bg-2018-384

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., et al. (2011). The community climate system model version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011JCLI4083.1

Graven, H. D., Keeling, R. F., Piper, S. C., Patra, P. K., Stephens, B. B., Wofsy, S. C., et al. (2013). Enhanced seasonal exchange of CO_2 by northern ecosystems since 1960. *Science*, *341*(6150), 1085–1089. https://doi.org/10.1126/science.1239207

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., et al. (2002). Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models. *Nature*, *415*(6872), 626–630. https://doi.org/10.1038/415626a

Hu, Q., Li, T., Deng, X., Wu, T., Zhai, P., Huang, D., et al. (2022). Intercomparison of global terrestrial carbon fluxes estimated by MODIS and Earth system models. *Science of the Total Environment*, *810*, 152231. https://doi.org/10.1016/j.scitotenv.2021.152231

Jacobson, A. R., Schuldt, K. N., Miller, J. B., Oda, T., Tans, P., Andrews, A., et al. (2020). CarbonTracker CT2019B [Dataset]. NOAA Global Monitoring Laboratory. https://doi.org/10.25925/20201008

Jin, Y., Keeling, R. F., Morgan, E. J., Ray, E., Parazoo, N. C., & Stephens, B. B. (2021). A mass-weighted isentropic coordinate for mapping chemical tracers and computing atmospheric inventories. *Atmospheric Chemistry and Physics*, *21*(1), 217–238. https://doi.org/10.5194/acp-21-217-202

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., et al. (2019). The FLUXCOM ensemble of global land-atmosphere energy fluxes [Dataset]. Scientific Data, *6*(1), 74. https://doi.org/10.1038/s41597-019-0076-8

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., et al. (2020). Scaling carbon fluxes from eddy covariance sites to globe: Synthesis and evaluation of the FLUXCOM approach. *Biogeosciences*, *17*(5), 1343–1365. https://doi.org/10.5194/bg-17-1343-2020

Keeling, C. D., Adams, J. A., Ekdahl, C. A., & Guenther, P. R. (1976). Atmospheric carbon dioxide variations at the South Pole. *Tellus*, *28*(6), 552–564. https://doi.org/10.1111/j.2153-3490.1976.tb00702.x

Keeling, C. D., Chin, J. F. S., & Whorf, T. P. (1996). Increased activity of northern vegetation inferred from atmospheric CO_2 measurements. *Nature*, *382*(6587), 146–149. https://doi.org/10.1038/382146a0

Keeling, R. F., & Graven, H. D. (2021). Insights from time series of atmospheric carbon dioxide and related tracers. *Annual Review of Environment and Resources*, *46*(1), 85–110. https://doi.org/10.1146/annurev-environ-012220-125406

Keeling, R. F., & Manning, A. C. (2014). Studies of recent changes in atmospheric O_2 content. In *Treatise on geochemistry*. Elsevier. https://doi.org/10.1016/B978-0-08-095975-7.00420-4

Keppel-Aleks, G., Randerson, J. T., Lindsay, K., Stephens, B. B., Keith Moore, J., Doney, S. C., et al. (2013). Atmospheric carbon dioxide variability in the community Earth system model: Evaluation and transient dynamics during the twentieth and twenty-first centuries. *Journal of Climate*, *26*(13), 4447–4475. https://doi.org/10.1175/JCLI-D-12-00589.1

Khatiwala, S., Primeau, F., & Hall, T. (2009). Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean. *Nature*, *462*(7271), 346–349. https://doi.org/10.1038/nature08526

Krol, M., Houweling, S., Bregman, B., van den Broek, M., Segers, A., van Velthoven, P., et al. (2005). The two-way nested global chemistry-transport zoom model TM5: Algorithm and applications. *Atmospheric Chemistry and Physics*, *5*(2), 417–432. https://doi.org/10.5194/acp-5-417-2005

Kumar, J., Hoffman, F., Hargrove, W., & Collier, N. (2016). Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements. In *Earth system science data discussions* (pp. 1–25). https://doi.org/10.5194/essd-2016-36

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., et al. (2019). The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4245–4287. https://doi.org/10.1029/2018MS001583

Lee, W.-L., & Liang, H.-C. (2020). AS-RCEC TaiESM1.0 model output prepared for CMIP6 CMIP historical, version 20200616 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.9755

Lee, W.-L., Wang, Y.-C., Shiu, C.-J., Tsai, I., Tu, C.-Y., Lan, Y.-Y., et al. (2020). Taiwan Earth system model version 1: Description and evaluation of mean state. *Geoscientific Model Development*, 13(9), 3887–3904. https://doi.org/10.5194/gmd-13-3887-2020

Liptak, J., Keppel-Aleks, G., & Lindsay, K. (2017). Drivers of multi-century trends in the atmospheric CO_2 mean annual cycle in a prognostic ESM. *Biogeosciences*, 14(6), 1383–1401. https://doi.org/10.5194/bg-14-1383-2017

Long, M. C., Lindsay, K., Peacock, S., Moore, J. K., & Doney, S. C. (2013). Twentieth-century oceanic carbon uptake and storage in CESM1(BGC)*. *Journal of Climate*, 26(18), 6775–6800. https://doi.org/10.1175/JCLI-D-12-00184.1

Lovato, T., & Peano, D. (2020). CMCC CMCC-CM2-SR5 model output prepared for CMIP6 CMIP historical, version 20200527 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.3825

Lovato, T., Peano, D., & Butenschön, M. (2021). CMCC CMCC-ESM2 model output prepared for CMIP6 CMIP historical, version 20201225 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.13195

Mao, J., Thornton, P. E., Shi, X., Zhao, M., & Post, W. M. (2012). Remote sensing evaluation of CLM4 GPP for the period 2000–09. *Journal of Climate*, 25(15), 5327–5342. https://doi.org/10.1175/JCLI-D-11-00401.1

NASA Goddard Institute for Space Studies (NASA/GISS). (2018). NASA-GISS GISS-E2.1G model output prepared for CMIP6 CMIP historical, version 20181015 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.7127

NASA Goddard Institute for Space Studies (NASA/GISS). (2019). NASA-GISS GISS-E2.1H model output prepared for CMIP6 CMIP historical, version 20190328 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.7128

Neubauer, D., Ferrachat, S., Siegenthaler-Le Drian, C., Stoll, J., Folini, D. S., Tegen, I., et al. (2019). HAMMOZ-Consortium MPI-ESM1.2-HAM model output prepared for CMIP6 CMIP historical, version 20191124 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.5016

Patra, P. K., Takigawa, M., Watanabe, S., Chandra, N., Ishijima, K., & Yamashita, Y. (2018). Improved chemical tracer simulation by MIROC4.0-based atmospheric chemistry-transport model (MIROC4-ACTM). *Scientific Online Letters on the Atmosphere*, 14(0), 91–96. https://doi.org/10.2151/SOLA.2018-016

Peiro, H., Crowell, S., Schuh, A., Baker, D. F., O'Dell, C., Jacobson, A. R., et al. (2022). Four years of global carbon cycle observed from the Orbiting Carbon Observatory 2 (OCO-2) version 9 and in situ data and comparison to OCO-2 version 7. *Atmospheric Chemistry and Physics*, 22(2), 1097–1130. https://doi.org/10.5194/acp-22-1097-2022

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, T. J., Conway, T. J., Masarie, K., et al. (2007). An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker. *Proceedings of the National Academy of Sciences*, 104(48), 18925–18930. https://doi.org/10.1073/pnas.0708986104

Randerson, J. T., Thompson, M. V., Conway, T. J., Fung, I. Y., & Field, C. B. (1997). The contribution of terrestrial sources and sinks to trends in the seasonal cycle of atmospheric carbon dioxide. *Global Biogeochemical Cycles*, 11(4), 535–560. https://doi.org/10.1029/97GB02268

Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., et al. (2004). The oceanic sink for anthropogenic CO_2. *Science*, 305(5682), 367–371. https://doi.org/10.1126/science.1097403

Santoni, G. W., Daube, B. C., Kort, E. A., Jiménez, R., Park, S., Pittman, J. V., et al. (2014). Evaluation of the airborne quantum cascade laser spectrometer (QCLS) measurements of the carbon and greenhouse gas suite-CO_2, CH_4, N_2O, and CO-during the CalNex and HIPPO campaigns. *Atmospheric Measurement Techniques*, 7(6), 1509–1526. https://doi.org/10.5194/amt-7-1509-2014

Schimel, D. S., House, J. I., Hibbard, K. A., Bousquet, P., Ciais, P., Peylin, P., et al. (2001). Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature*, 414(6860), 169–172. https://doi.org/10.1038/35102500

Schuh, A. E., Jacobson, A. R., Basu, S., Weir, B., Baker, D., Bowman, K., et al. (2019). Quantifying the impact of atmospheric transport uncertainty on CO_2 surface flux estimates. *Global Biogeochemical Cycles*, 33(4), 484–500. https://doi.org/10.1029/2018GB006086

Seland, Ø., Bentsen, M., Oliviè, D. J. L., Toniazzo, A., Graff, L. S., et al. (2019). NCC NorESM2-LM model output prepared for CMIP6 CMIP historical, version 20190917 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.8036

Simpson, I. R., McKinnon, K. A., Davenport, F. v., Tingley, M., Lehner, F., al Fahad, A., & Chen, D. (2021). Emergent constraints on the large scale atmospheric circulation and regional hydroclimate: Do they still work in CMIP6 and how much can they actually constrain the future? *Journal of Climate*, 34(15), 6355–6377. https://doi.org/10.1175/JCLI-D-21-0055.1

Stephens, B. B., Gurney, K. R., Tans, P. P., Sweeney, C., Peters, W., Bruhwiler, L., et al. (2007). Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric CO_2. *Science*, 316(5832), 1732–1735. https://doi.org/10.1126/science.1137004

Stephens, B. B., Morgan, E. J., Bent, J. D., Keeling, R. F., Watt, A. S., Shertz, S. R., & Daube, B. C. (2021). Airborne measurements of oxygen concentration from the surface to the lower stratosphere and pole to pole. *Atmospheric Measurement Techniques*, 14(3), 2543–2574. https://doi.org/10.5194/AMT-14-2543-2021

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). CCCma CanESM5 model output prepared for CMIP6 CMIP historical, version 20190501 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.3610

Sweeney, C., Karion, A., Wolter, S., Newberger, T., Guenther, D., Higgs, J. A., et al. (2015). Seasonal climatology of CO_2 across North America from aircraft measurements in the NOAA/ESRL global greenhouse gas reference network. *Journal of Geophysical Research: Atmospheres*, 120(10), 5155–5190. https://doi.org/10.1002/2014JD022591

Tans, P. P., Fung, I. Y., & Takahashi, T. (1990). Observational constraints on the global atmospheric CO_2 budget. *Science*, 247(4949), 1431–1438. https://doi.org/10.1126/science.247.4949.1431

Thompson, C. R., Wofsy, S. C., Prather, M. J., Newman, P. A., Hanisco, T. F., Ryerson, T. B., et al. (2022). The NASA atmospheric tomography (ATom) mission: Imaging the chemistry of the global atmosphere. *Bulletin of the American Meteorological Society*, 103(3), E761–E790. https://doi.org/10.1175/BAMS-D-20-0315.1

Thompson, R. L., Patra, P. K., Chevallier, F., Maksyutov, S., Law, R. M., Ziehn, T., et al. (2016). Top–down assessment of the Asian carbon budget since the mid 1990s. *Nature Communications*, 7(1), 10724. https://doi.org/10.1038/ncomms10724

Thoning, K. W., Crotwell, A. M., & Mund, J. W. (2022). *Atmospheric carbon dioxide dry air mole fractions from continuous measurements at Mauna Loa, Hawaii, Barrow, Alaska, American Samoa and south Pole. 1973–2021, version 2022–05 National Oceanic and Atmospheric Administration (NOAA)*. Global Monitoring Laboratory (GML). https://doi.org/10.15138/yaf1-bk21

Tjiputra, J. F., Roelandt, C., Bentsen, M., Lawrence, D. M., Lorentzen, T., Schwinger, J., et al. (2013). Evaluation of the carbon cycle components in the Norwegian Earth system model (NorESM). *Geoscientific Model Development*, 6(2), 301–325. https://doi.org/10.5194/gmd-6-301-2013

Valentini, R., Matteucci, G., Dolman, A. J., Schulze, E. D., Rebmann, C., Moors, E. J., et al. (2000). Respiration as the main determinant of carbon balance in European forests. *Nature*, 404(6780), 861–865. https://doi.org/10.1038/35009084

Verma, S., Marshall, J., Gerbig, C., Rödenbeck, C., & Uwe Totsche, K. (2017). The constraint of CO_2 measurements made onboard passenger aircraft on surface-atmosphere fluxes: The impact of transport model errors in vertical mixing. *Atmospheric Chemistry and Physics*, 17(9), 5665–5675. https://doi.org/10.5194/acp-17-5665-2017

Volodin, E. M., Dianskii, N. A., & Gusev, A. V. (2010). Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations. *Izvestiya: Atmospheric and Oceanic Physics*, *46*(4), 414–431. https://doi.org/10.1134/S000143381004002X

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., et al. (2011). MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, *4*(4), 845–872. https://doi.org/10.5194/gmd-4-845-2011

Welp, L. R., Keeling, R. F., Meijer, H. A. J., Bollenbacher, A. F., Piper, S. C., Yoshimura, K., et al. (2011). Interannual variability in the oxygen isotopes of atmospheric CO_2 driven by El Niño. *Nature*, *477*(7366), 579–582. https://doi.org/10.1038/nature10421

Wieners, K.-H., Giorgetta, M., Jungclaus, J., Reick, C., Esch, M., Bittner, M., et al. (2019). MPI-M MPI-ESM1.2-LR model output prepared for CMIP6 CMIP historical, version 20190929 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.6595

Williamson, D. B., & Sansom, P. G. (2019). How are emergent constraints quantifying uncertainty and what do they leave behind? *Bulletin of the American Meteorological Society*, *100*(12), 2571–2588. https://doi.org/10.1175/BAMS-D-19-0131.1

Williamson, M. S., Thackeray, C. W., Cox, P. M., Hall, A., Huntingford, C., & Nijsse, F. J. M. M. (2021). Emergent constraints on climate sensitivities. *Reviews of Modern Physics*, *93*(2), 025004. https://doi.org/10.1103/RevModPhys.93.025004

Wofsy, S. C. (2011). HIAPER Pole-to-Pole Observations (HIPPO): Fine-grained, global-scale measurements of climatically important atmospheric gases and aerosols. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *369*(1943), 2073–2086. https://doi.org/10.1098/rsta.2010.0313

Wofsy, S. C., Afshar, S., Allen, H. M., Apel, E. C., Asher, E. C., Barletta, B., et al.(2021). ATom: Merged atmospheric chemistry, trace gases, and aerosols, version 2 [Dataset]. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1925

Wofsy, S. C., Daube, B. C., Jimenez, R., Kort, E., Pittman, J. V., Park, S., et al. (2017). HIPPO merged 10-second meteorology, atmospheric chemistry, and aerosol data. Version 1.0 [Dataset]. UCAR/NCAR: Earth Observing Laboratory. https://doi.org/10.3334/CDIAC/HIPPO_010

Yang, Z., Washenfelder, R. A., Keppel-Aleks, G., Krakauer, N. Y., Randerson, J. T., Tans, P. P., et al. (2007). New constraints on Northern Hemisphere growing season net flux. *Geophysical Research Letters*, *34*(12), L12807. https://doi.org/10.1029/2007GL029742

Ziehn, T., Chamberlain, M., Lenton, A., Law, R., Bodman, R., Dix, M., et al. (2019). CSIRO ACCESS-ESM1.5 model output prepared for CMIP6 CMIP historical, version 20191115 [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4272