Quantifying errors in observationally-based estimates of ocean carbon sink variability

Lucas Gloege[1*], Galen A. McKinley[1], Peter Landschützer[2], Amanda R. Fay[1], Thomas L. Frölicher[3,4], John C. Fyfe[5], Tatiana Ilyina[2], Steve Jones[6], Nicole S. Lovenduski[7], Christian Rödenbeck[8], Keith Rodgers[9,10], Sarah Schlunegger[11], Yohei Takano[2]

[1] Lamont-Doherty Earth Observatory, Palisades, NY 10964 USA

[2] Max Planck Institute for Meteorology, Hamburg, Germany

[3] Climate and Environmental Physics, University of Bern, Bern, Switzerland

[4] Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

[5] Environment and Climate Change Canada, Victoria, Canada

[6] University of Bergen, Bergen, Norway

[7] University of Colorado, Boulder, CO 80309 USA

[8] Max Planck Institute for Biogeochemistry, Jena, Germany

[9] Center for Climate Physics, Institute for Basic Science, Busan, South Korea

[10] Pusan National University, Busan, South Korea

[11] Princeton University, Princeton, NJ 08544 USA

*Corresponding author: Contact Lucas Gloege at gloege@ldeo.columbia.edu

Abstract

Reducing uncertainty in the global carbon budget requires better quantification of ocean CO_2 uptake and its temporal variability. Several methodologies for reconstructing air-sea CO_2 exchange from sparse pCO_2 observations indicate larger decadal variability than estimated using ocean models. We assess these reconstructions' ability to estimate spatiotemporal variability, by creating the Large Ensemble Testbed using four independent Earth system models. Model pCO_2 fields are subsampled as the observations, for each of 100 ensemble members, and the reconstruction is performed as is done with real-world observations. The power of a testbed is that the perfect reconstruction is known from the original model fields; thus, reconstruction skill can be comprehensively assessed. We find that a commonly used neural-network approach can skillfully reconstruct air-sea CO_2 fluxes when and where it is trained with sufficient data. Flux bias is low for the global mean and Northern Hemisphere, but can be regionally high in the Southern Hemisphere. The phase and amplitude of the seasonal cycle are accurately reconstructed outside of the tropics, but longer-term variations are reconstructed with only moderate skill. For Southern Ocean decadal variability, insufficient sampling leads to a 39% [15%:58%, interquartile range] overestimation of amplitude, and phasing is only moderately correlated with known truth (r=0.54 [0.46:0.63]). Globally, the amplitude of decadal variability is overestimated by 21% [3%:34%]. Machine learning, when supplied with sufficient data, can skillfully reconstruct ocean properties. However, data sparsity remains a fundamental limitation to quantification of decadal variability in the ocean carbon sink.

Significance statement

The ocean is a significant sink for anthropogenic CO_2. Accurate estimates of air-sea CO_2 exchange are needed to track if we are on target to satisfy internationally agreed climate targets. Observations remain very sparse, thus statistical approaches have been proposed to fill data gaps. However, the uncertainties in these gap-filling reconstructions are unclear. We use Earth system model simulations to evaluate one commonly used machine learning approach, and find that data sparsity in the Southern Ocean causes decadal variability to be significantly overestimated. Data coverage needs to be increased to improve quantification of the ocean carbon sink. Our approach can be used to evaluate other existing reconstructions, and to support the development of new approaches.

Introduction

The ocean significantly modulates atmospheric CO_2, having absorbed about 39% of industrial-age fossil carbon emissions (1). Under high emission scenarios, the ocean sink is projected to grow and become the primary sink for anthropogenic carbon emissions over the next several centuries (2). Under low emission scenarios, such as those that would limit global warming to 2°C, the ocean carbon sink will decline rapidly as the near-surface waters that hold the bulk of anthropogenic carbon (3) come into equilibrium with the atmosphere (4, 5). As the long-term response to the changing atmospheric pCO_2 unfolds, the ocean sink will continue to be modified on seasonal to decadal timescales by climate variability and change. Ultimately, our ability to accurately monitor the fate of anthropogenic carbon in the Earth system requires a quantification of the spatially-resolved variability of the ocean carbon sink on timescales from seasonal to multi-decadal. To achieve this goal, global maps of surface ocean pCO_2 are required, from which air-sea CO_2 exchange can be derived.

The direction of the air-sea CO_2 flux is set by the gradient in pCO_2 across the air-sea interface with additional controls from the gas transfer velocity and CO_2 solubility setting the magnitude. Satellites cannot directly measure surface ocean pCO_2; therefore, hindcast simulations with ocean models (1) and observation-based gap-filling techniques (6) are integral to providing a global picture of the evolving ocean carbon sink. Essential to these techniques are high quality *in-situ* pCO_2 measurements, such as those annually compiled in the Surface Ocean CO_2 ATlas (SOCAT) (7, 8). But these data are too sparse to directly constrain global air-sea CO_2 exchange. The latest SOCAT database release covers only 1.5% of all possible monthly 1°x1° points from 1982 to 2019, which poses challenges to an accurate global CO_2 flux estimate. Current gap-filling techniques, such as the self-organizing map feed-forward neural-network (SOM-FFN)(9), provide continuous monthly mean estimates. However, these results lack a comprehensive, spatially-resolved assessment of uncertainties. Understanding these uncertainties is important for understanding the mechanisms of variability (10, 11), to compare model output to observation-based data products (12), to benchmark Earth system model based prediction systems (13), and to assess impacts on the global carbon budget (1, 14). Here, we present a comprehensive, spatially-resolved assessment of uncertainty in the SOM-FFN.

Our Large Ensemble Testbed uses 100 members from four Large Ensemble Earth system models, 25 members each, to evaluate the performance of the SOM-FFN over 1982-2016 given real-world pCO_2 sampling (Figure 1A). For each ensemble member, the pCO_2 reconstruction is performed in the same manner as in the SOM-FFN application to SOCAT pCO_2 data (see Methods). We sample the pCO_2 field of each testbed ensemble member as the SOCATv5 database (step 1) and use co-located driver data (see Methods) from the same ensemble member output to train, evaluate, and test the SOM-FFN (step 2). We then reconstruct full-field pCO_2 from the full-field driver data (step 3). CO_2 flux is then calculated using the reconstructed and original climate model pCO_2 field

(step 4). This is repeated for each ensemble member, providing a total of 100 unique reconstruction and model-truth pairs. To assess the performance across various timescales, we deconstruct the flux into seasonal, decadal, and sub-decadal components (Figure 1B) (see Methods). Performance on decadal time scales is of particular interest, since the reconstruction techniques indicate greater decadal variability than ocean models, especially in the Southern Ocean (15–18).

We emphasize that the goal of this work is not to provide an estimate of real-world air-sea CO_2 exchange, but instead to assess the statistical fidelity of SOM-FFN given real-world sampling. Fidelity is quantified by three metrics: the method's ability to capture the long-term mean, and the phase and amplitude of seasonal to decadal time-scale variability. Our approach allows assessment of the reconstruction's fidelity across a wide range of potential states of ocean internal variability as estimated by 25 ensembles each from 4 independent Earth System models.

Results

Reconstruction bias

Regionally, the 1982-2016 mean CO_2 flux from SOM-FFN can be biased high or low by more than 0.50 mol C m^{-2} yr^{-1} (Figure 2A), but these patches average out such that the global average bias is small (-0.01 mol C m^{-2} yr^{-1}). Regional biases are smaller in the Northern Hemisphere where data are more dense, and larger in the Indian Ocean and Southern Hemisphere where data are more sparse (Figure 2B,C). The mean and interquartile range of biases in the Northern and Southern Hemisphere is [0.01, -0.05:0.06] and [-0.04, -0.13:0.06] mol C m^{-2} yr^{-1}, respectively. Grid cells with at least 48 months of data have a mean bias that does not exceed 0.14 mol C m^{-2} yr^{-1} 90% of the time (Figure 2C).

Reconstruction phasing

Temporal correlation of the reconstruction to the original model field for each ensemble member indicates the ability of SOM-FFN to accurately capture phasing of variability at seasonal, sub-decadal, and decadal time scales (Figure 3A-C). The standard deviation of the correlations indicates the degree to which correlations are consistent across ensemble members (Figure 3D-F). Spatial coincidence of low standard deviations and high correlations indicates that the reconstruction performs well across all the climate states represented by the ensemble members.

Reconstructed CO_2 flux, for the seasonal cycle, has the highest correlation to its original model field in the subtropics (Figure 3A). The large seasonal amplitude provides a prominent signal that the neural-network can identify (supplemental Figure 1). Higher data density in the Northern Hemisphere (Figure 2B) leads to a marginally better reconstruction which leads to better constraints on the seasonal cycle here. The lack of a prominent seasonal cycle in the tropics (19) leads to a limited signal for an SOM-FFN reconstruction. The ability of the SOM-FFN to capture monthly variations is patchy in the Southern Ocean and Indian Ocean; two regions that have been

previously identified as having the largest mismatch towards observations and the expected seasonal amplitude increase (10, 20). Despite smaller correlations around the equator and in the Southern Ocean and Indian Ocean, the global average correlation is 0.89. Additionally, regions of high correlation have low spread across the ensemble members (Figure 3D). The pattern correlation between the mean correlation (Figure 3A) and the spread of the correlations (Figure 3D) is -0.88, indicating a tight consistency between the mean result and the 100 ensemble members.

The SOM-FFN methodology, when combined with the available observations, is less capable in reconstructing variability at sub-decadal (Figure 3B) and decadal (Figure 3C) time-scales; contrasting the seasonal signal. Global average correlation values are 0.75 and 0.58, respectively. Correlations are lower on decadal timescales (Figure 3C) than on sub-decadal timescales (Figure 3B) in the subtropics. The decadal signal is best reconstructed in the Western Pacific warm pool. The pattern correlations between the mean and standard deviation across ensemble members are moderate (r=-0.77 for sub-decadal, and r=-0.66 for decadal); indicating a wide spread of correlations where the mean correlations are moderate. This suggests that in some ensemble members at specific locations, even the very sparse sampling that occurred was sufficient to capture the dominant modes of variation. However, this is not generally true across the ensemble, indicating a lack of robustness to the particular realization of oceanic variability.

Reconstruction of the amplitude

Percent error of the standard deviation quantifies how well the reconstruction captures the true amplitude of variability. SOM-FFN, for the global average, overestimates the amplitude of the seasonal cycle by 7%(Figure 4A). Regionally, the reconstruction is accurate north of 35°N, but in the tropics and Southern Hemisphere, the seasonal amplitude is overestimated by a median value of 10% [3%:12%] (Figure 4A,D). The amplitude of sub-decadal variability is slightly underestimated at most locations, with a global average of -1% (Figure 4B).

On decadal timescales, SOM-FFN overestimates the amplitude of variability at most locations and for both the regional and global means. Globally, the overestimate is 21% (Figure 4C). In the Southern Ocean (<35°S), the median is a 39% overestimation, with a large interquartile range across ensemble members (Figure 4F).

The percent overestimation, by definition, is inversely proportional to the model standard deviation, and the four climate models of the Large Ensemble have different inherent amplitudes of decadal variability (Figure 5A) (21, 22). It is quite promising that the amplitude of the reconstructed decadal variability is close to its appropriate original model, as indicated by the small spread in average absolute error (AAE) (0.03-0.06 molC m^{-2} yr^{-1}). AAE is defined as the mean of the absolute difference between the standard deviation of the reconstruction and of the original

model field. Thus, SOM-FFN is skillful in capturing the broad range of decadal variability simulated by the different climate models, even with the very sparse sampling. However, this broad range of underlying decadal variability influences the percent error. MPI-GE has a large decadal variability, and a low percent error (5%); conversely, GFDL-ESM2M has a small decadal variability and thus a high percent error (84%). Since we do not know which of these models best represent the true decadal variability of the Southern Ocean, the median across all four Large Ensembles (39%) is our best estimate.

Influence of additional Southern Ocean sampling

In recent years, the sampling density in the Southern Ocean has substantially increased through the launch of the fleet of drifters and Bio-Argo floats (23, 24)**.** To assess the future impact of this new data source on our results, we test the potential impact that this additional Southern Ocean sampling would have on the reconstruction. We additionally supplement the sampling in the Southern Ocean (Figure 6A) for a subset of ensemble members within the Large Ensemble Testbed (Figure 1). Specifically, the historical sample locations of all SOCCOM and CARIOCA measurements are collapsed to a monthly climatology, and then assumed to have occurred at the same locations every year from 1982-2016. This adds 114,972 additional samples at 592 locations, equivalent to increasing data density from 1.4% with only SOCAT to 2.1% with the artificially persistent floats.

As expected, the additional sampling substantially improves the fidelity of the Southern Ocean reconstruction on all timescales (Figure 6). Enhanced sampling in the Southern Ocean also improves the reconstruction outside of the region because the biogeographic provinces of SOM-FFN are constrained by physical and biogeochemical properties; not by geography. Focusing on the Southern Ocean, the phasing of the decadal variability is improved, as indicated by higher mean correlations (Figure 6B vs. 3C). Error in the amplitude is much reduced at most locations (Figure 6C vs. 4C). The simulated additional sampling also reduces the spread of amplitude error across the ensemble members on seasonal (Figure 6D), sub-decadal (Figure 6E), and decadal time scales (Figure 6F). The interquartile range, for the decadal time-scale, across the 28 member subset is [-11.6%, 0.0%] with a median of -6.9%. A combination of ship and float data has recently led to smaller Southern Ocean fluxes in recent years (25). However, if SOCAT sampling had been supplemented by continuous drifters and floats in the Southern Ocean for the last 3 decades, giving only 2.1% sampling coverage, we would now be able to reconstruct the amplitude of real-world decadal variations in the Southern Ocean carbon sink to within 20% (Figure 5B,6F) and globally to within 2% (Figure 6C). This provides evidence that Southern Ocean observations are key to improving the reconstruction's ability to capture decadal variability.

Discussion

These results offer the first spatially-resolved quantification of the uncertainty of observation-based CO_2 flux reconstruction on seasonal to decadal timescales. We address reconstruction fidelity for the ocean CO_2 flux given real-world pCO_2 data sparsity across a range of simulated realizations of the ocean's internal variability. We do not account for uncertainties in measurements, in the representativity of one or small number of instantaneous pCO_2 observations for a full month and a 1°x1° grid cell, nor in the full-field driver data. System lags beyond the instantaneous non-linear response to changes in the driver fields are also ignored. Model output has previously been used to assess performance of this or similar statistical approaches for pCO_2 reconstruction either using a single model (26, 27) or an ensemble of hindcast models (28), and should continue to serve as a method benchmark (29). The advantage of the Large Ensemble Testbed is that it is much less dependent on the particular model simulation used as "truth" and allows for a statistically robust assessment of reconstruction performance across a range of climate states and model structures/representations. This testbed can be used to test other reconstruction approaches, for development of new approaches, and for evaluating new sampling strategies (26), and is now publicly available (see Methods). Here, we tested the ability of the SOM-FFN method to accurately reconstruct pCO_2 across the global ocean. We illustrate that the reconstruction method itself can be fairly accurate across timescales, but that data sparsity remains a fundamental limitation.

The SOM-FFN has previously been used as a reference field to assess the performance of model simulations over the historical period (12, 14, 30–33). We find that SOM-FFN provides a robust global estimate of the mean CO_2 uptake by the ocean, but regionally and locally, its performance is dependent on the location and the density of observations. If there are at least 48 months of data for a 35 year timeframe, the mean bias in the long-term mean is under 0.14 mol C m^{-2} y^{-1} 90% of the time (Figure 2C). Mean bias can locally be much larger, particularly in poorly sampled regions such as the Southern Hemisphere. Similarly, the ability of the reconstruction to accurately capture the phase (Figure 3) and amplitude (Figure 4) of variability on sub-decadal and decadal time scales varies regionally. To improve observation-based reconstructions of the ocean carbon sink in the future, additional sampling will be critical (Figure 6).

When driven with real-world SOCAT observations and driver data, SOM-FFN indicates large amplitude decadal variability in the Southern Ocean carbon sink, with a significant slowdown in uptake over the 1990s, reaching a minimum in 2001, and then a recovery (11, 15, 16, 34) until around 2011 (17). Here, we demonstrate that the SOM-FFN method overestimates the amplitude of the decadal variability in the Southern Ocean by a median of 39% across all ensemble members. A reduction of the amplitude of decadal variability would bring SOM-FFN more in line with other observation-based products (18), ocean circulation inverse models (15) and with ocean models (1, 14, 15).

Though this work strongly indicates that SOM-FFN overestimates decadal variability of the Southern Ocean and of the globe, it does not provide a clear basis for a direct rescaling of the SOM-FFN for comparison to other estimates (1, 11, 16, 34). First, correlations indicate that decadal variability (Figure 3C, F) is only reconstructed with moderate skill in terms of phasing. Second, with respect to amplitude, one could note that the magnitude of Southern Ocean reconstructed variability from real data using SOM-FFN is 0.17 mol C m^{-2} yr^{-1} and from (35), 0.16 mol C m^{-2} yr^{-1} (16). A re-scaling could be derived from the mean AAE, implying a reduction of 0.04 mol C m^{-2} yr^{-1} to arrive at 0.12-0.13 mol C m^{-2} yr^{-1}. This would be a downscaling of approximately 25%. However, a percent scaling has been shown to be strongly dependent on the background variability of the real ocean (Figure 5), and we do not know which of the four climate models best represents this. Directly from the median percent error, 39% would be the best choice for a rescaling of Southern Ocean amplitude, leading to 0.06 mol C m^{-2} yr^{-1}. To restate, we do not have a clear basis for a direct rescaling. One way to constrain this range in the future could be to reconstruct pCO_2 within a suite of hindcast models that have less spread in their underlying variability due to their forcing with realistic meteorology.

We use Large Ensemble model output to provide the first detailed statistical assessment of the uncertainty in a reconstruction of air-sea CO_2 fluxes based on sparse *in-situ* ocean pCO_2 data. Flux bias is low for the global mean, and at most locations in the Northern Hemisphere. However, bias can be regionally high in the data-poor Southern Hemisphere. The seasonal cycle is well-captured in phase and amplitude outside of the tropics. Interannual phase and amplitude are better captured in the Northern Hemisphere and the tropics than in the Southern Hemisphere. In the Southern Ocean, insufficient sampling leads to a 39% [15%:58%] overestimation of decadal variability. Globally averaged, the amplitude of decadal variability is overestimated by 21% [3%:34%]. To improve observation-based reconstructions of the ocean carbon sink, extension of sampling to include the Southern Ocean and other data-poor regions is required.

Methods

SOM-FFN pCO2 interpolation

Self-organizing map feed-forward neural-network (SOM-FFN) (11, 20, 36) is a non-linear regression using a combination of self-organizing maps (SOM) and feed-forward neural-networks (FFN) to extrapolate from sparse pCO_2 observations to a global 1°x1° grid at a monthly resolution. To estimate pCO_2 at each spatial location, SOM-FFN relies on auxiliary datasets with full, or approximately full, global coverage: Sea Surface Temperature (SST) and Surface Chlorophyll-a (Chl-a) from satellite; Sea Surface Salinity (SSS) from a compilation of *in-situ* data sources; Mixed layer depth (MLD) climatology from argo floats; and atmospheric CO_2 mixing ratio (xCO_2). These variables serve as proxies for known processes affecting pCO_2. The long-term growth of pCO_2 is driven by atmospheric CO_2 (xCO_2). Solubility is set by SSS and SST. Biological

uptake of dissolved inorganic carbon (DIC) is indicated by Chl-a. Biological productivity and entrainment of DIC are influenced by MLDs.

The first step uses a self-organizing map (SOM) to cluster the global ocean into 16 biogeochemical provinces based on climatological variables (surface ocean pCO_2 from (37), SST, SSS, MLD, and Chl-a). This allows for neural-network algorithms specific to each province to be developed in the second step, taking advantage of regional coherence in the dominant drivers of pCO_2 variability (eg. SST in subtropics, DIC in subpolar).

The second step develops a non-linear regression to estimate pCO_2 given the aforementioned environmental driver variables (SST, SSS, MLD, Chl-a, xCO_2). All driver variables are monthly varying from 1982 through 2016, with the exception of climatological MLD. Any gaps in the driver data are either replaced with climatology or removed from the estimation. Within each province, a unique feed-forward neural-network (FFN) is developed to link the driver variables to pCO_2 observations from SOCAT. This approach does not impose mechanistic relationships. Once the FFN algorithm is trained, tested, and evaluated on SOCAT pCO_2 in each province, the relationship is applied to continuous fields of driver variables to estimate pCO_2 at all 1°x1° locations and all months from 1982-2016. Finally, air-sea CO_2 exchange is calculated following (38).

The large ensemble testbed

Our 100-member Large Ensemble Testbed includes 25 randomly selected members from each of four independent initial-condition ensemble models:

- CanESM2: Second Generation Canadian Earth-System Model (RCP8.5) (39)
- CESM-LENS: Community Earth System Model – Large Ensemble (RCP8.5) (40)
- GFDL-ESM2M: Geophysical Fluid Dynamics Laboratory Earth-System Model (RCP8.5) (41)
- MPI-GE: Max Planck Institute for Meteorology Grand Ensemble (RCP8.5) (42)

Each individual climate model is an imperfect representation of the actual Earth system, thus we use multiple Large Ensembles to span across the different model structures and their representation of internal variability. Each large ensemble member uses the same external forcing of historical atmospheric CO_2 before 2005 and Representative Concentration Pathway 8.5 (RCP8.5) afterwards. Spread in the ensemble members is generated by perturbing the initial state of the Earth system at the start of each simulation. This is accomplished either by changing the seed value that goes into a random number generator as part of the cloud parameterization (CanESM2), perturbing the initial air-temperature field with round-off level differences (CESM-LENS), or branching off from snap-shots of the historical simulation (GFDL-ESM) or pre-industrial simulation (MPI-GE). These initial perturbations cause each ensemble member to have a unique atmosphere and ocean state at each point in time, i.e. a different state of internal variability. By using many ensemble

members it is possible to test the methods ability to capture the full range of pCO_2 variability potential in the system under any possible climate state, not only that which occurred in the real ocean. As a specific example, the real ocean experienced an El Niño in 1997-1998. In the testbed, ensembles may have had a La Niña, El Niño or been neutral at this time. We expect only that Southern Oscillation statistics be consistent with the real world in the climate model ensembles.

To create the testbed, we retrieve monthly averaged SST, SSS, Chl-a, MLD, xCO_2, and pCO_2 from each member. A bilinear interpolation scheme is used to transform each field to a 1°x1° rectilinear grid, the same resolution as the SOCATv5 gridded product (8). Each member's monthly varying ocean pCO_2 is then sampled at the resolution of the SOCATv5 data product, with the other variables remaining un-sampled. The sampled pCO_2 field and co-located driver data for each of the 100 members constitutes the Large Ensemble Testbed capable of evaluating pCO_2 interpolation methods. The intention is to create fields that mimic the environmental driver variables and SOCATv5 data used in the real-world application of the SOM-FFN interpolation. After the monthly varying pCO_2 field is reconstructed for each member, air-sea CO_2 exchange is calculated.

Air-sea CO_2 exchange

Air-sea CO_2 flux is calculated in mol C m^{-2} yr^{-1} for each month at each 1°x1° spatial location using the (38) parameterization with the (43) scale factor of 0.27. High-frequency output is not available for all large ensemble members thus to be consistent with the flux calculation used in the real-world application of the SOM-FFN flux product, we use ERA-interim 6-hourly global atmospheric reanalysis (44) as an estimate for the wind-speed variance. Saturation vapor pressure is removed from the total pressure when calculating the atmospheric partial pressure of CO_2 (45). See Supplemental Text 1 for more details.

Temporal decomposition

To evaluate the performance of the SOM-FFN on various time scales, an approach similar to (46) is used to temporally decompose the air-sea CO_2 flux into additive components at each grid point (see Figure 1B for an illustration).

We first eliminate the influence of increasing atmospheric CO_2 by removing a linear-trend at each 1°x1° grid cell from the reconstructed air-sea CO_2 flux and the model truth. Then, a repeating seasonal cycle is calculated from the detrended time series. After removing the seasonal component, the decadal signal is isolated by applying a locally weighted regression (loess) smoother (47) with a 10-year window. Finally, the remaining signal not explained by a linear trend, seasonal cycle, or decadal trend is here termed the sub-decadal component. This decomposition was done for both the reconstructed and model truth air-sea CO_2 flux for each of the 100 ensemble members. Statistical metrics were applied across each time scale.

Statistical metrics

The fidelity of the reconstruction is based on a suite of statistical metrics (48). Our focus is on bias, correlation, percent error in standard deviation, and average absolute error, chosen to assess if the reconstruction captures the long-term mean, temporal phasing of the signal, and variability observed in the model. Each ensemble member is treated as an equally likely climate state, thus statistical metrics are averaged across the 100 ensemble members. Spread in each metric across ensemble members is quantified by the standard deviation.

Bias is calculated as the long-term mean of the reconstruction (R) minus the model truth (M), $bias = \bar{R} - \bar{M}$,with the overbar representing the mean over 1982-2016. Bias is a measure of the systematic discrepancy between the reconstruction and model over the long term. It is important to note that values near zero may be misleading as positive and negative discrepancies can cancel out.

Pearson correlation coefficient, r, is defined as the covariance between the reconstruction and the model divided by the product of their standard deviations, $r = \frac{cov(R,M)}{\sigma_R \sigma_M}$. Correlation is used to quantify the synchrony between the reconstruction and model truth. Values are bounded between $-1 \leq r \leq 1$, which quantifies the degree to which reconstruction captures the phasing observed in the model. Values near 1 and -1 indicate that the reconstruction and model are perfectly in or out of phase, respectively. Intermediate values indicate a phase shift between the two signals, with values closer to zero indicating a larger phase shift between signals.

Percent error ($\%error = (\frac{\sigma_R - \sigma_M}{\sigma_M}) * 100$) in the standard deviation quantifies the degree to which the reconstruction correctly captures the amplitude of CO_2 flux variability as observed in the ensemble member. This metric indicates whether the reconstruction overestimates ($\%error>0$), underestimates (%error<0), or perfectly captures ($\%error=0$) the variability of the model truth. This metric is sensitive to the model standard deviation.

Average Absolute Error (AAE) quantifies how well the magnitude of variability is reconstructed in units of mol C m^{-1} yr^{-1} . It is defined as the absolute difference between the standard deviation of the reconstruction and of the original model field averaged across all ensemble members ($AAE = |\sigma_R - \sigma_M|$).

Data Availability

The 100 member large ensemble testbed is publicly available at https://figshare.com/collections/Large_ensemble_pCO2_testbed/4568555. Data analysis scripts are contained in GitHub repository https://github.com/lgloege/large_ensemble_testbed. SOCATv5 is available at https://www.socat.info/index.php/previous-versions/. ERA-interm 6

hourly output is available at https://apps.ecmwf.int/datasets/. Any other inquiries should be addressed to L.G.

Acknowledgments
We acknowledge support from Columbia University and the Center for Climate and Life at Lamont-Doherty Earth Observatory. We acknowledge NCAR for use of the Cheyenne high performance computer under project code UWIS0028. LG, GAM, ARF, and NSL acknowledge funding from the National Science Foundation (OCE-1558225). TLF acknowledges support from the Swiss National Science Foundation under grant PP00P2-170687 and from the European Union's Horizon 2020 Research and Innovative Programme under grant agreement number 821003 (CCiCC). Support for K.B.R. was provided by the Institute for Basic Science project code IBS-R028-D1. The Surface Ocean CO_2 Atlas (SOCAT) is an international effort, endorsed by the International Ocean Carbon Coordination Project (IOCCP), the Surface Ocean Lower Atmosphere Study (SOLAS) and the Integrated Marine Biosphere Research (IMBeR) program, to deliver a uniformly quality-controlled surface ocean CO_2 database. The many researchers and funding agencies responsible for the collection of data and quality control are thanked for their contributions to SOCAT. We acknowledge Data were collected and made freely available by the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Project funded by the National Science Foundation, Division of Polar Programs (NSF PLR -1425989), supplemented by NASA, and by the International Argo Program and the NOAA programs that contribute to it. (http://www.argo.ucsd.edu, http://argo.jcommops.org). The Argo Program is part of the Global Ocean Observing System.

Contributions
L.G., G.A.M., P.L. and N.S.L. designed the study. P.L. performed all reconstructions with the SOM-FFN. L.G. drafted the manuscript and performed all analyses. All authors (L.G., G.A.M., P.L., N.S.L., K.R., A.R.F., T.L.F., J.C.F., T.I., S.J., C.R., S.R., and Y.T.) participated in the interpretation of the analysis, discussed results, and refined the manuscript.

Competing interests
The authors are not aware of any competing interests

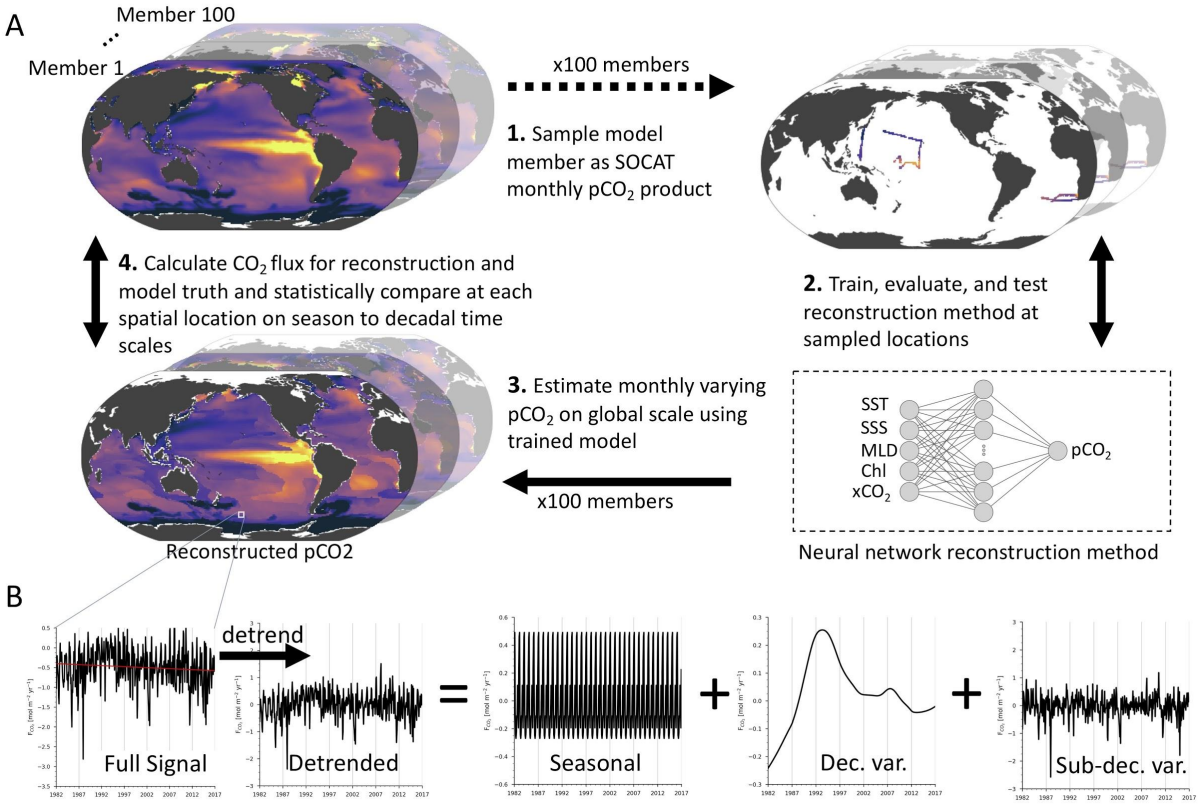


Figure 1: The Large Ensemble Testbed. A) Schematic of the testbed; oceanic pCO_2 from each of the 100 members is sampled in space and time like the SOCAT gridded product (Step 1). The sampled output is used with auxiliary model output variables to reconstruct pCO_2 in the same way as the real-world application of the SOM-FFN (Landschützer, Gruber, Bakker, & Schuster, 2014) (Step 2). pCO_2 is reconstructed everywhere using full-field auxiliary datasets (Step 3). Finally, CO_2 flux is calculated for the model truth and reconstruction for each of the 100 ensemble members and then statistically compared across seasonal to decadal time scales (Step 4). Maps in the schematic are pCO_2. B) Illustrated breakdown of CO_2 flux time series at a single point into seasonal, decadal, and sub-decadal variability.

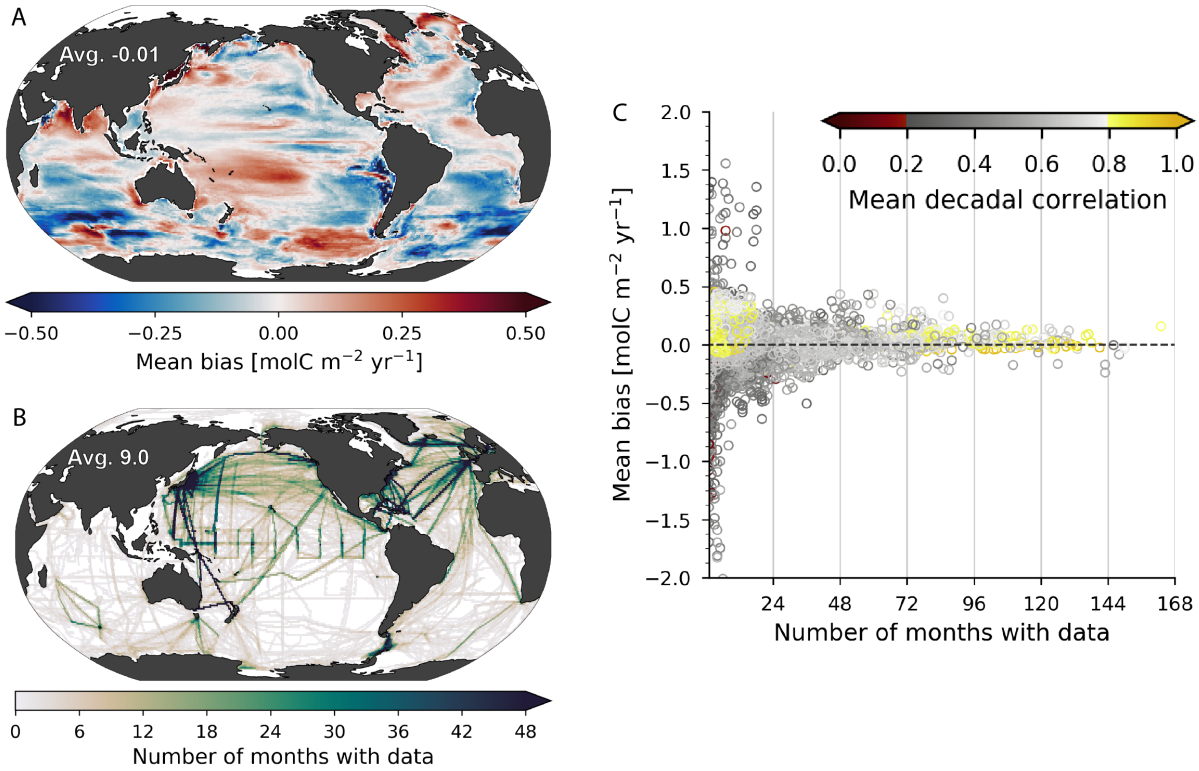


Figure 2: **Reconstruction bias and sampling density.** A) Bias between reconstruction and model truth, averaged over the 100 ensemble members. Red and blue shading indicates regions where the reconstruction is biased high or low, respectively. B) Number of months with observations in each grid cell. C) Cross plot of bias with number of months with data, by 1°x1° grid cell. Color indicates correlation between the reconstruction and model truth on decadal time scale.

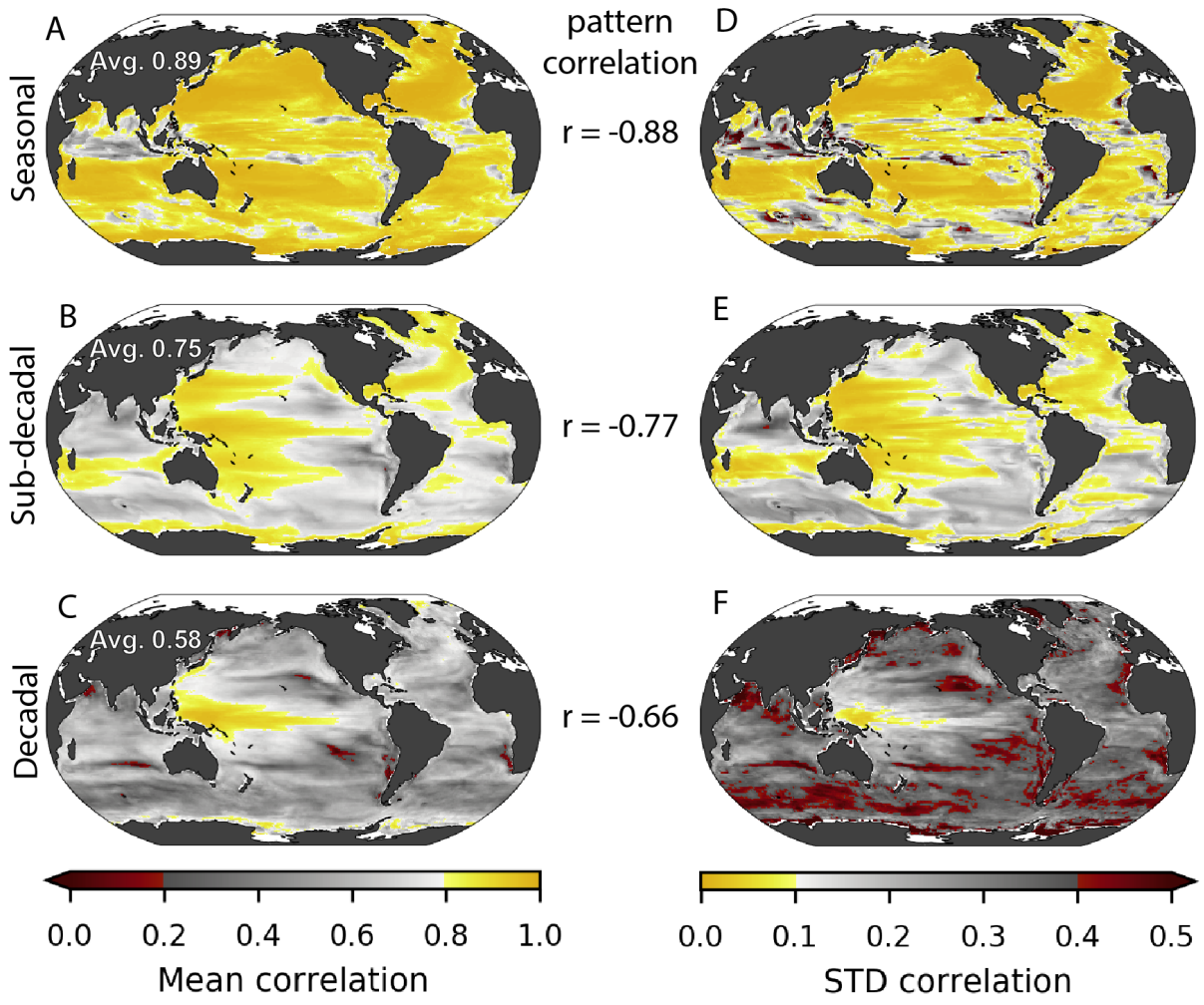


Figure 3: **Phasing of SOM-FFN reconstructed variability on seasonal, sub-decadal and decadal, compared to original model.** Correlation between reconstruction and original model on A) seasonal, B) sub-decadal, and C) decadal time scales, averaged across the 100 ensemble members. The global average is displayed atop each plot. The standard deviation of the correlation across the 100 ensemble members is shown on D) seasonal, E) sub-decadal, and F) decadal time scales. The pattern correlation between the mean and standard deviation is displayed between each pair of maps, with values close to -1 signifying high correlations are consistent across ensemble members. Note the reversed scale such that high mean correlation and low standard deviation, together indicating a robust reconstruction, have the same coloration.

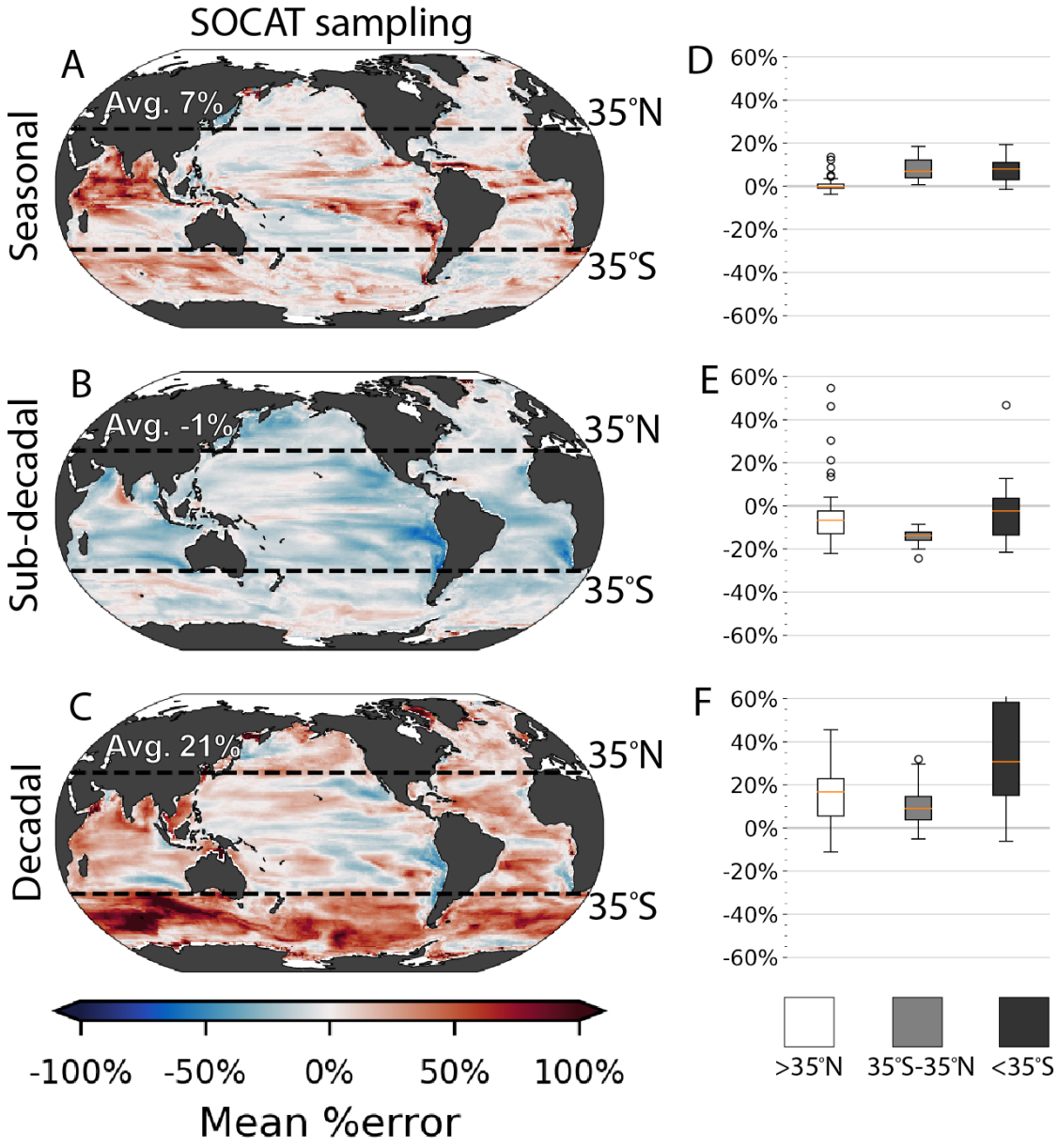


Figure 4: **Error of amplitude in SOM-FFN reconstructed variability on seasonal, sub-decadal and decadal.** Percent error of CO_2 flux standard deviation on A) seasonal, B) sub-decadal, and C) decadal time scales, averaged across the 100 ensemble members. Global average is shown in white text. Color indicates the percentage by which the reconstruction over or under-estimates the variability. (D-F) Percent error as shown in A-C, averaged within three regions delineated by latitude for each of the 100 ensemble members and displayed as box plots on D) seasonal, E) sub-decadal, and F) decadal time scales. Boxes indicate the interquartile range (IQR), the orange line indicates the median, and circles indicate points greater than 1.5*(IQR).

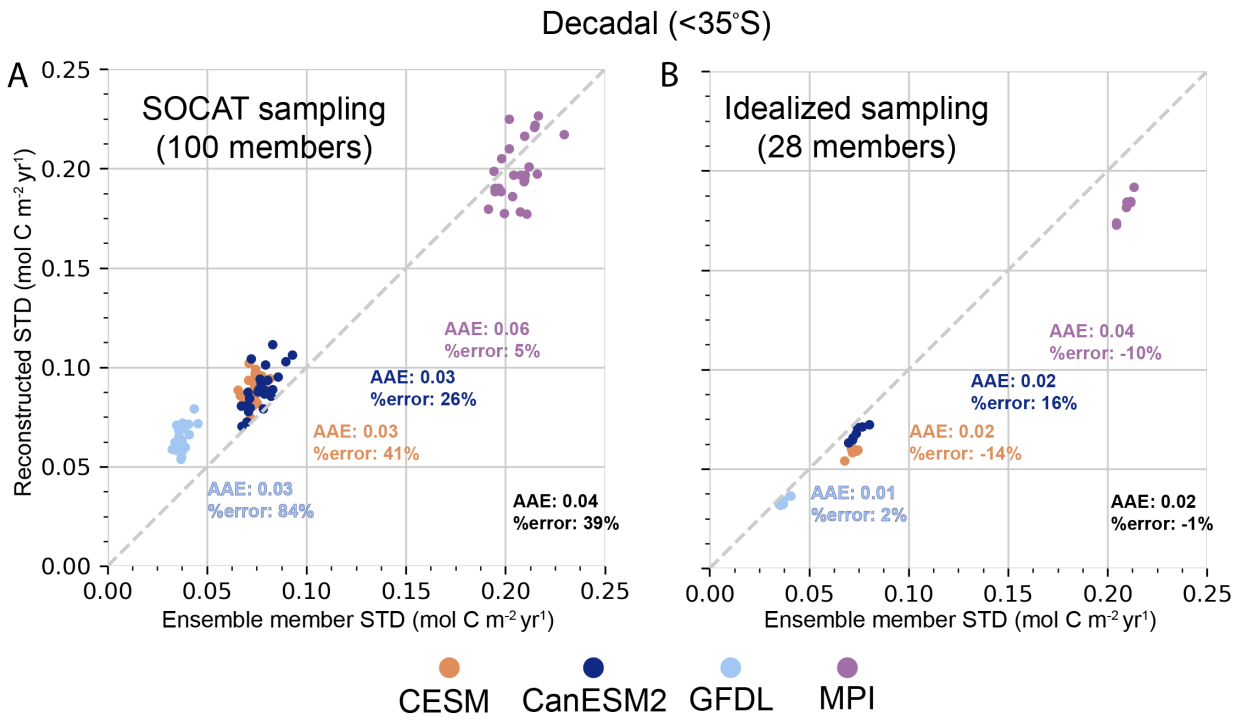


Figure 5: Cross plot of decadal standard deviation in the Southern Ocean. The reconstructed and ensemble member decadal standard deviation averaged across the Southern Ocean (<35°S), separated by model. Colored text indicates average absolute error (AAE), and the percent error averaged across members from each model with A) the SOCAT sampling and B) idealized sampling. Black text indicates statistics averaged across all the ensemble members.

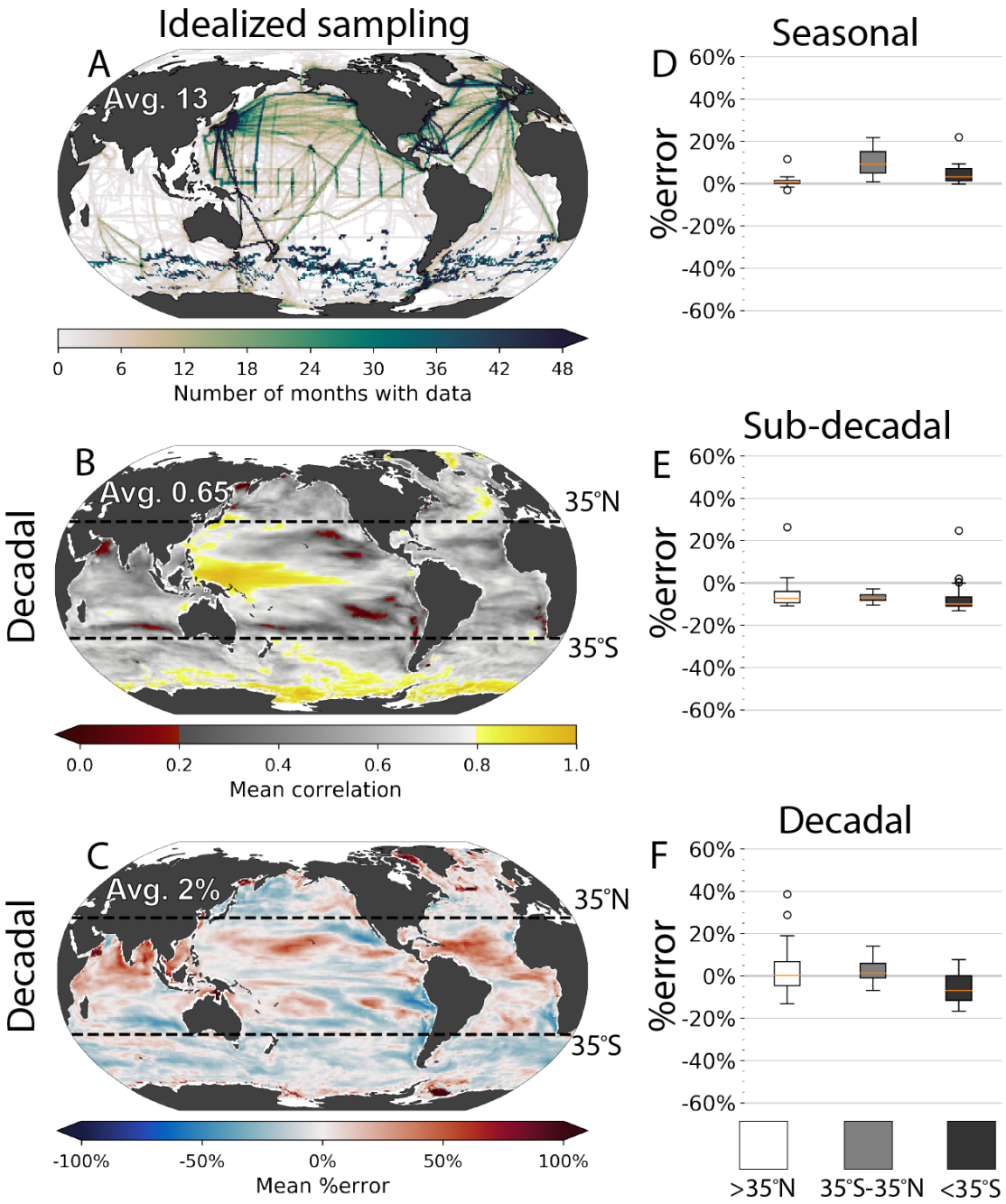
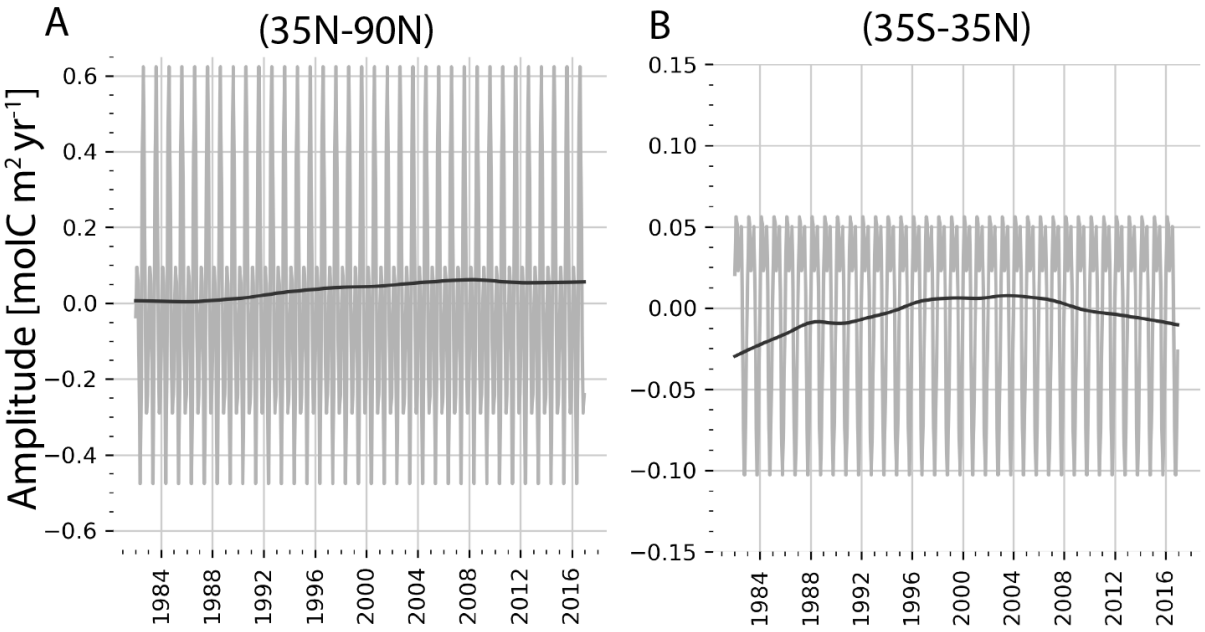


Figure 6: Potential fidelity (phasing and amplitude) of SOM-FFN decadal reconstruction, had there been persistent drifters and floats in the Southern Ocean since 1982. A) Number of months with data, with SOCAT plus idealized float sampling in the Southern Ocean; the mean B) correlation and C) percent error of CO_2 flux standard deviation on decadal time scales across the 28 members using SOCAT plus idealized float sampling, similar to Figure 4C but with additional sampling. Box plots of percent error indicate spread among members within three regions delineated by latitude are shown on D) seasonal, E) sub-decadal, and F) decadal time scales.

Supplemental material



Supplemental Figure 1: Average seasonal and decadal amplitude in Northern Hemisphere and tropics. The average seasonal cycle (gray) and decadal component (black) is displayed across A) 35°N - 90°N and B) 35°S - 35°N. Note different y-axis scales.

Supplemental text 1

Air-sea CO_2 flux (F_{CO_2}) is calculated in mol C m^{-2} yr^{-1} for each month at each 1°x1° spatial location using the (38) parameterization (Equation 1).

$$\bar{F_{CO_2}} = \bar{k_w} \bar{S_{CO_w}} (1 - \bar{f_{ice}})(\bar{pCO_{2,atm-moist}} - \bar{pCO_{2,ocean}}) \qquad Equation\ 1$$

which parameterizes F_{CO_2} as a function of the gas transfer velocity (k_w), CO_2 solubility (S_{CO_2}), ice fraction (f_{ice}), and partial pressure of CO_2 in moist air ($pCO_{2,atm-moist}$) and surface ocean ($pCO_{2,ocean}$). Overbars denote monthly averages. We use the (38) gas transfer velocity with the (43) scale factor of 0.27 (Equation 2).

$$\bar{k_w} = 0.27(\bar{u}^2 + \bar{u'^2})\ (\bar{Sc}/660)^{-0.5} \qquad Equation\ 2$$

Because high-frequency output is not available for all large ensemble members, and to be consistent with the flux calculation used in the real-world application of the SOM-FFN flux product, we use ERA-interim 6-hourly global atmospheric reanalysis (44) as an estimate for the wind-speed variance ($\bar{u'^2}$).

Solubility is calculated following (49) with the (38) Schmidt number (Sc). Partial pressure of moist air ($pCO_{2,atm-moist}$) is calculated following Equation 3.

$$pCO_{2,atm-moist} = xCO_2(P_{atm} - pH_2O) \qquad Equation\ 3$$

Where xCO_2 is the dry air mixing ratio of atmospheric CO_2, P_{atm} is the total atmospheric pressure, and pH_2O is the saturation vapor pressure (45).

References

1. P. Friedlingstein, *et al.*, Global Carbon Budget 2019. *Earth Syst. Sci. Data* **11**, 1783–1838 (2019).
2. J. T. Randerson, *et al.*, Multicentury changes in ocean and land contributions to the climate-carbon feedback. *Global Biogeochemical Cycles* **29**, 744–759 (2015).
3. N. Gruber, *et al.*, The oceanic sink for anthropogenic CO2 from 1994 to 2007. *Science* **363**, 1193–1199 (2019).
4. P. M. Cox, Emergent Constraints on Climate-Carbon Cycle Feedbacks. *Curr Clim Change Rep* (2019).
5. C. D. Jones, *et al.*, Simulating the Earth system response to negative emissions. *Environ. Res. Lett.* **11**, 095012 (2016).
6. C. Rödenbeck, *et al.*, Data-based estimates of the ocean carbon sink variability–first results of the Surface Ocean pCO2 Mapping intercomparison (SOCOM). *Biogeosciences* **12**, 7251–7278 (2015).
7. D. C. Bakker, *et al.*, A multi-decade record of high-quality fCO2 data in version 3 of the Surface Ocean CO2 Atlas (SOCAT) (2016).
8. C. L. Sabine, *et al.*, Surface Ocean CO2 Atlas (SOCAT) gridded data products. 9 (2013).
9. P. Landschützer, N. Gruber, D. C. Bakker, Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles* **30**, 1396–1417 (2016).
10. P. Landschützer, N. Gruber, D. C. Bakker, I. Stemmler, K. D. Six, Strengthening seasonal marine CO 2 variations due to increasing atmospheric CO 2. *Nature Climate Change* **8**, 146 (2018).
11. P. Landschützer, *et al.*, The reinvigoration of the Southern Ocean carbon sink. *Science* **349**, 1221–1224 (2015).
12. N. Mongwe, M. Vichi, P. Monteiro, The seasonal cycle of pCO 2 and CO 2 fluxes in the Southern Ocean: diagnosing anomalies in CMIP5 Earth system models. *Biogeosciences* **15**, 2851–2872 (2018).
13. H. Li, T. Ilyina, W. A. Müller, P. Landschützer, Predicting the variable ocean carbon sink. *Sci. Adv.* **5**, eaav6471 (2019).
14. C. Le Quéré, *et al.*, Global carbon budget 2018. *Earth System Science Data* **10**, 2141–2194 (2018).
15. T. DeVries, *et al.*, Decadal trends in the ocean carbon sink. *Proceedings of the National Academy of Sciences* **116**, 11646–11651 (2019).
16. N. Gruber, P. Landschützer, N. S. Lovenduski, The Variable Southern Ocean Carbon Sink. *Annual Review of Marine Science* **11**, 159–186 (2019).
17. L. Keppler, P. Landschützer, Regional Wind Variability Modulates the Southern Ocean Carbon Sink. *Scientific Reports* **9** (2019).
18. R. Ritter, *et al.*, Observation-Based Trends of the Southern Ocean Carbon Sink. *Geophysical Research Letters* **44**, 12,339-12,348 (2017).
19. U. Schuster, *et al.*, An assessment of the Atlantic and Arctic sea–air CO_2 fluxes, 1990–2009. *Biogeosciences* **10**, 607–627 (2013).
20. P. Landschützer, N. Gruber, D. Bakker, U. Schuster, Recent variability of the global ocean carbon sink. *Global Biogeochemical Cycles* **28**, 927–949 (2014).
21. L. Resplandy, R. Séférian, L. Bopp, Natural variability of CO2 and O2 fluxes: What can we learn from centuries-long climate models simulations? *Journal of Geophysical Research: Oceans* **120**, 384–404 (2015).
22. S. Schlunegger, *et al.*, Time of emergence and large ensemble inter comparison for ocean biogeochemical trends. *submitted*.
23. J. Boutin, L. Merlivat, C. Hénocq, N. Martin, J. B. Sallée, Air-sea CO2 flux variability in frontal regions of the Southern Ocean from CARbon Interface OCean Atmosphere drifters. *Limnology and Oceanography* **53**, 2062–2079 (2008).
24. S. C. Riser, D. Swift, R. Drucker, Profiling Floats in SOCCOM: Technical Capabilities for Studying the Southern Ocean. *Journal of Geophysical Research: Oceans* **123**, 4055–4073 (2018).

25. S. M. Bushinsky, *et al.*, Reassessing Southern Ocean Air-Sea CO2 Flux Estimates With the Addition of Biogeochemical Float Observations. *Global Biogeochemical Cycles* **33**, 1370–1388 (2019).
26. L. Gregor, S. Kok, P. M. S. Monteiro, Empirical methods for the estimation of Southern Ocean CO2: support vector and random forest regression. *Biogeosciences* **14**, 5551–5569 (2017).
27. S. D. Jones, C. L. Quéré, C. Rödenbeck, A. C. Manning, A. Olsen, A statistical gap-filling method to interpolate global monthly surface ocean carbon dioxide data. *Journal of Advances in Modeling Earth Systems* **7**, 1554–1575 (2015).
28. A. D. Lebehot, *et al.*, Reconciling Observation and Model Trends in North Atlantic Surface CO2 (2019) https:/doi.org/10.1029/2019gb006186.
29. M. Reichstein, *et al.*, Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
30. R. Arruda, *et al.*, Air–sea CO2 fluxes and the controls on ocean surface pCO2 variability in coastal and open-ocean southwestern Atlantic Ocean: a modeling study. *Biogeosciences Discussions* **12**, 7369–7409 (2015).
31. T. Bourgeois, *et al.*, Coastal-ocean uptake of anthropogenic carbon. *Biogeosciences* **13**, 4167–4185 (2016).
32. T. L. Frölicher, *et al.*, Dominance of the Southern Ocean in Anthropogenic Carbon and Heat Uptake in CMIP5 Models. *Journal of Climate* **28**, 862–886 (2015).
33. A. Kessler, J. Tjiputra, The Southern Ocean as a constraint to reduce uncertainty in future ocean carbon sinks. *Earth System Dynamics* **7**, 295–312 (2016).
34. Galen A McKinley, Amanda R Fay, Yassir A Eddibar, Lucas Gloege, Nicole S Lovenduski, External forcing explains recent decadal variability of the ocean carbon sink (2020).
35. C. Rödenbeck, *et al.*, Interannual sea–air CO2 flux variability from an observation-driven ocean mixed-layer scheme. *Biogeosciences* **11**, 4599–4613 (2014).
36. P. Landschützer, *et al.*, A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. *Biogeosciences* **10**, 7793–7815 (2013).
37. T. Takahashi, *et al.*, Climatological mean and decadal change in surface ocean pCO2 and net sea–air CO2 flux over the global oceans. *Deep Sea Research Part II: Topical Studies in Oceanography* **56**, 554–577 (2009).
38. R. Wanninkhof, Relationship between wind speed and gas exchange over the ocean. *Journal of Geophysical Research: Oceans* **97**, 7373–7382 (1992).
39. J. C. Fyfe, *et al.*, Large near-term projected snowpack loss over the western United States. *Nature communications* **8**, 14996 (2017).
40. J. E. Kay, *et al.*, The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society* **96**, 1333–1349 (2015).
41. K. B. Rodgers, J. Lin, T. L. Frölicher, Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences* **12**, 3301–3320 (2015).
42. N. Maher, *et al.*, The Max Planck Institute Grand Ensemble-Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems* **11**, 2050–2069 (2019).
43. C. Sweeney, *et al.*, Constraining global air-sea gas exchange for CO2 with recent bomb 14C measurements. *Global Biogeochemical Cycles* **21** (2007).
44. D. P. Dee, *et al.*, The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* **137**, 553–597 (2011).
45. A. G. Dickson, C. L. Sabine, J. R. Christian, C. P. Bargeron, North Pacific Marine Science Organization, Eds., *Guide to best practices for ocean CO2 measurements* (North Pacific Marine Science Organization, 2007).
46. R. B. Cleveland, W. S. Cleveland, J. E. McRae, I. Terpenning, STL: a seasonal-trend decomposition. *Journal of official statistics* **6**, 3–73 (1990).
47. W. S. Cleveland, S. J. Devlin, Locally Weighted Regression: An Approach to Regression Analysis

by Local Fitting. *Journal of the American Statistical Association* **83**, 596–610 (1988).
48. C. A. Stow, *et al.*, Skill assessment for coupled biological/physical models of marine systems. *Journal of Marine Systems* **76**, 4–15 (2009).
49. R. F. Weiss, Carbon dioxide in water and seawater: the solubility of a non-ideal gas. *Marine Chemistry* **2**, 203–215 (1974).