The "Ocean Carbon States" Database: a proof-of-concept application of cluster analysis in the ocean carbon cycle

Rebecca Latto[1,2], Anastasia Romanou[1,2]
[1]Applied Physics and Applied Math, Columbia University, New York, USA
[2]NASA-GISS, New York, NY, USA

Correspondence to: rl2797@columbia.edu

Abstract. In this paper, we present a database of the basic regimes of the carbon cycle in the ocean as obtained using a data mining/pattern recognition technique in observation-based as well as model data. Advanced data mining techniques are becoming widely used in Climate and Earth Sciences with the purpose of extracting new meaningful information from large and complex datasets. Such techniques need to be rigorously tested, however, in simple, well-understood cases to better assess their utility. This is particularly important for studies of the global carbon cycle, where the interaction of physical and biogeochemical drivers confounds our ability to accurately describe, understand, and predict CO_2 concentrations and their changes in the major planetary carbon reservoirs. In this proof-of-concept study, we focus on using well-understood data that are based on observations, as well as model results from the NASA Goddard Institute for Space Studies (GISS) climate model. The clustering algorithm organizes the observation-based data into various regimes, the "ocean carbon states", which are shown to be associated with the subtropical-subpolar gyre during the colder months of the year and the tropics during the warmer season in the North Atlantic basin. Conversely, in the Southern Ocean, the ocean carbon states can be associated with the subtropical and Antarctic convergence zones in the warmer season and the coastal Antarctic divergence zone in the colder season. Additionally, the carbon states are used to assess physical and biogeochemical processes in the model. We find that the GISS model reproduces the cold and warm season regimes more skillfully in the North Atlantic than in the Southern Ocean and matches the observed seasonality better than the spatial distribution of the regimes. The ocean carbon states can be further used to illustrate and characterize the model biases in the air-sea flux of CO_2. In the North Atlantic, GISS model biases are dominated by the wind speed and salinity biases in the subpolar region due to misrepresentation of the cyclonic wind stress curl and the sea-ice retreat. We also find nitrate and wind speed biases in the subtropics and tropics that are explained, respectively, by inaccurate representations of nitrate utilization and the biases associated with equatorial atmosphere-ocean interaction processes. The Southern Ocean flux bias is dominated by nitrate because of a mischaracterization of the nitrogen cycle in the model. The goal of this proof-of-concept study is to introduce the means by which well-known characteristics of the ocean carbon cycle in the North Atlantic and Southern Ocean basins can be reproduced, and thereby establish a methodology for using cluster analysis with larger, higher frequency datasets. All data and analysis scripts are available at data.giss.nasa.gov/oceans/carbonstates/ (DOI: 10.5281/zenodo.996891)

1 Introduction

The ocean carbon cycle plays an important role in controlling the airborne fraction of CO_2 in the atmosphere, thereby regulating the rate of global warming, i.e. the rising temperatures in the Earth's troposphere. In this paper, we seek to present and assess a data mining/pattern recognition technique, namely cluster analysis, for the purpose of defining the basic regimes, or "ocean carbon states", that describe the oceanic carbon cycle variability.

For geophysical applications, climate datasets have inherent complexities that are not easily identifiable in the age of "big" data. Cluster analysis is a highly effective uni- or multivariate classification method for large, high frequency data sets because it can find structure in a body of complex, geophysical data (Anderberg, 1973; Peron et al., 2014). Clustering seeks to identify the critical modes and natural patterns of a dataset without any training or predetermined spatial-temporal guidelines, therefore it is an "unsupervised" graph theory method. The merit of a novel, unsupervised method such as clustering is that it can recognize connectivity between multiple variables. This can be understood as connectivity in a temporal sense where cluster analysis can identify joint interannual or seasonal patterns and in a spatial sense where clustering has the power to identify patterns that relate different regions or basins (Jain, 2010; Phillips, 2015).

Traditional methods of univariate analysis, such as principal component analysis or spectral decomposition, cannot fully describe important physical states of the climate system or adequately detect change (Hoffman et al. 2011) because these methods neglect interactions between state variables as well as spatial and temporal co-variability. In contrast, cluster analysis has been successfully applied to various dynamical systems in order to extract the organized states and detect change as well as in novel applications of model-data intercomparison (Hoffman et al, 2008). For example, this technique has been used to define atmospheric weather states by identifying cloud regimes (Jakob and Tselioudis, 2003; Rossow et al., 2005; Williams and Webb, 2009; Tselioudis et al., 2013; Bodas-Salcedo et al., 2014; Oreopoulos et al., 2016). Bankert and Solbrig (2015) were able to extract a three-dimensional cloud representation using cluster analysis. This technique has also been used to characterize water types in lakes (Trochta et al., 2014), hydraulic habitat composition in rivers (Hugue et al., 2015), phenology patterns in forests (Trans Mills et al., 2011), solar variability (Zagouras et al., 2012), ENSO phenomena (Radebach et al., 2013), and regions with characteristic hydrological response (Halverson and Fleming, 2014), among many other applications.

Beyond identifying regimes, cluster analysis can be useful in model assessment applications, like that of Wood et al. 2015, which used weather states derived from cluster analysis for process studies, satellite calibration, and model evaluation. Both regime identification and model evaluation are the focus of the cluster analysis presented in this paper as well.

Elsewhere in ocean carbon cycle science, clustering-type methods (self-organizing maps and neural networks) have been used to build reconstructions or as regression analysis alternatives for surface ocean pCO_2 (Lefevre et al., 2005; Telszewski et al.,

2009; Sasse et al., 2013; Landschützer et al., 2013, 2014; Nakaoka et al., 2013). Unlike these studies, here we seek to obtain the co-variability maps and conditions of different ocean related variables and understand where, why and how they change. Other non-statistical studies, but similar in concept to multivariate regime identification, have focused more on larger scale geographic variations (Fay and McKinley, 2014; Trochta et al., 2015) than on the regional aspects of the ocean biogeochemistry and its interaction with physical circulation like in the western boundary current regions, in the upwelling zones on the eastern boundaries, and in the eddying field.

The structure of the paper is as follows. Section 2 describes the data sets used in this study, both the observation-based sources as well as the model experiments. Section 3 presents the k-means cluster analysis methodology and application, including discussion of the k-means clustering technique and sensitivity to number of clusters chosen, to binning, and to data normalization. The results of the methodology are provided in section 4. Section 4.1 focuses on how the methodology is applied in observations from the North Atlantic basin. The observed ocean carbon states are then characterized temporally and spatially in order to reveal their physical meaning. Next, the model carbon states are computed and characterized in a similar way as the observations. Using the ocean carbon states, model biases are also discussed and evaluated. Section 4.2 repeats the analysis presented in Section 4.1, but now applied to the Southern Ocean. Finally, general discussion and conclusions are provided in Section 5. A note about the figures in the paper: some interesting but non-critical figures are offered in the Supplemental Material and are denoted as Fig. S#. All data and analysis scripts are available at data.giss.nasa.gov/oceans/carbonstates/ website (DOI: 10.5281/zenodo.996891)

2. Data

Choice of variables to represent ocean carbon regimes

One critical question to answer at the onset of any clustering analysis is what key geophysical variables should be used to base the analysis on. For the purposes of this study, we picked sea surface temperature (SST) and partial pressure of CO_2 in the ocean surface water (pCO_{2SW}). The rationale for this choice will be explained now. There are two main pathways that determine the ability of the ocean to take up CO_2 (Sarmiento and Gruber, 2006): the chemical disequilibrium, expressed by pCO_2, dissolved inorganic carbon (DIC = the sum of all inorganic carbon species) and nutrients, and the physical processes, such as air-sea interaction (expressed by the wind speed) and ocean circulation (expressed by sea surface temperature and salinity). Greater insight into the ocean's biogeochemical processes that control these pathways can inform the improved use of field measurements, the development of better metrics for model evaluation, and the selection of more suitable parameterizations in climate models in order to provide more accurate predictions. We select pCO_{2SW} and SST because they are able to represent a broad range of biogeochemical and physical processes. We use them in cluster analysis to find temporal and spatial patterns

in their joint parameter space that can be used to understand CO_2 flux distributions and its fluctuations. Other variables pairs can be alternatively used here; a comparison between choices is set aside for future work.

This study will focus on two oceanic basins, namely the North Atlantic (defined as 80°W to 45°E, 0° to 90°N) and the Southern Ocean (defined as 180°W to 180°E, 90°S to 40°S), because of their importance in the global carbon cycle (Takahashi et al., 2009).

2.1 Observation-based data

Air-Sea flux of CO_2 and pCO_2, surface wind speed, sea surface temperature and salinity

The 12-month climatology of the air-sea flux is obtained from the Carbon Dioxide Information Analysis Center (LDEO database (NDP-088); Takahashi et al., 2009). It is derived from the difference between surface water pCO_2 (pCO_{2SW}), air pCO_2, and the air-sea gas transfer rate. Surface water pCO_2 climatological mean distribution was obtained from 3 million measurements from 1970 to 2007, and normalized to a reference year 2000. The pCO_2 of the air is computed from the GlobalView CO_2 concentration zonal mean, NCAR monthly mean barometric pressure, SST, and salinity. Other variables in the data set pertinent to this analysis are wind speed (derived from the 1979-2005 climatological mean NCEP-DOE AMIP-II Reanalysis wind speed field), climatological sea surface temperature (from NOAA Climate Diagnostic Center Objective Interpolation), and salinity (from the NODC World Ocean Database 1998). All variables are available as a 12-month climatology at a 4º x 5º resolution.

Nitrate

The nitrate monthly climatology at 1 degree horizontal resolution is obtained from the World Ocean Atlas 2013 version 2 (Boyer et al., 2013). It is collected from in situ measurements at standard depth levels and is available as annual, seasonal, and monthly climatologies. Nitrate is an essential nutrient that limits the growth of phytoplankton, which is responsible for fixating carbon dioxide from the atmosphere. Therefore, pCO_2 levels in the surface ocean depend partially on the abundance of nitrate.

2.2 Numerical Simulations

The NASA-GISS modelE2.1 output used for this analysis comes from 5 ensemble coupled model simulations of the 20th Century with realistic greenhouse gas, aerosol, land use and solar forcing, as used in CMIP5 experiments. The model physics is somewhat different than the modelE2 used in the CMIP5 experiments mostly due to improved representation of the ocean mesoscale mixing. The physical ocean and the biogeochemistry modules are described in detail in Romanou et al (2013; 2017). Briefly here we note that the ocean model is a non-Boussinesq mass-conserving ocean model with 32 vertical levels and 1°x1.25° horizontal resolution. The vertical coordinate is a stretched z-level coordinate and has a free surface and natural

surface boundary fluxes of freshwater and heat that are obtained by the atmospheric model. In addition to advection and turbulent mixing, it also includes a scheme for isopycnal eddy fluxes and isopycnal thickness diffusion. The interactive ocean carbon cycle model consists of a biogeochemical model (NASA Ocean Biogeochemistry Model (NOBM) Gregg and Casey, 2007) and a gas exchange parameterization for the computation of the CO_2 flux between the ocean and the atmosphere (Romanou et al., 2013). Specifically, the air-sea exchange of CO_2 (Sarmiento and Gruber, 2006; Takahashi et al., 2009) is described by Eq. (1):

$$F = kwK_0(pCO_{2atm} - pCO_{2sw}) \quad (1)$$

where kw is the piston velocity for CO_2 (in ms^{-1}) that depends on the wind speed, K_0 is the solubility coefficient- dependent on sea surface temperature (SST) and sea surface salinity (SSS) (expressed in mole,$CO_2 kg^{-1} atm^{-1}$)- and pCO_2 is the partial pressure of CO_2 (Wanninkhof et al., 2013) in the atmosphere (atm) and the surface ocean (sw). Eq. (1) describes the chemical disequilibrium of CO_2 in the oceanic and atmospheric reservoirs due to the solubility and biological pumps. As discussed in Sarmiento and Gruber (2006), the pCO_{2SW} in Eq. (1) is a function of temperature and salinity, wind speed, DIC, nutrients, and alkalinity (a measure of the excess of bases over acids) which can be expressed as follows:

$$pCO_{2sw} = f(SST, SSS, DIC, windspeed, nutrients, alkalinity) \quad (2)$$

NOBM utilizes ocean temperature and salinity, mixed layer depth and the ocean circulation fields, and the horizontal advection and vertical mixing schemes obtained from the host ocean model as well as shortwave radiation (direct and diffuse) and surface wind speed obtained from the atmospheric model to produce horizontal and vertical distributions of several biogeochemical constituents. The carbon submodel parameterizes the cycling of carbon through the phytoplankton, herbivore and detrital components, affecting the dissolved inorganic and organic carbon in the ocean and interacting with the atmosphere. Alkalinity is assumed analogous to surface salinity, which is an acceptable approximation for the sea surface but does not take into account changes in the carbonate pump. Temperature and salinity are affected only by physical processes such as circulation, advection, eddy mixing and stirring, and local upwelling/downwelling while DIC distributions are influenced by all these physical processes and also several biogeochemical processes such as air-sea gas exchange, production by organisms, biological export to depth and remineralization there and nutrient availability in the water column. Atmospheric pCO_2 (pCO_{2atm}) is the saturation concentration of CO_2 in equilibrium with a water–vapor-saturated atmosphere at a total atmospheric pressure P and a given atmospheric pCO_2 level:

$$\mathrm{pCO2}_{atm} = \frac{P}{P^0}\, CO2^0 \quad (3)$$

where P_0 = 1 atm and $[CO_2]^0$ is the saturation concentration at 1 atm total pressure.

The gas transfer velocity is given by

$$kw = c\left(\frac{Sc}{660}\right)^{-1/2} wspd^2 \quad (4)$$

where wspd is the surface wind speed and c is the piston velocity coefficient taken here equal to $0.337/(3.6\text{x}10^5)$. The value of c has been agreed upon by the Ocean Carbon Model Intercomparison Project, phase II (OCMIP-II) so that the global, annual mean gas transfer coefficient for carbon dioxide (kw, K_0) is equal to 0.061 mol/m^2/yr/l atm for preindustrial times. Sc, the

Schmidt number, is computed using the temperature of the host ocean model following Wanninkhof (1992). The gas transfer velocity kw is computed only over open water. The solubility of CO_2 in the water K_0 is also parameterized based on OCMIP using prognostic temperature, salinity and sea level pressure. In these model runs, the global average of the atmospheric concentration of CO_2 follows the Mauna Loa measurements (Dlugokencky and Tans, 2014), although regionally atmospheric CO_2 is allowed to vary due to the distributions of the ocean sources and sinks.

The five ensemble member runs were averaged into one ensemble mean to account for the intrinsic climate variability that is not adequately resolved in climate models of low spatial resolution. The model output for the years 1995 – 2005 was then averaged again to produce a 12-month climatology for the purpose of direct comparison with the observationally based data in the Takahashi database.

The model output and the observational data were interpolated onto the same grid, which is the Takahashi ocean grid at 4x5 degree resolution, with no Arctic Ocean, and the ocean mask was conformed across all observational and model data sets.

In the rest of the paper, some conventions with regards to nomenclature should be noted. Firstly, the Takahashi carbon flux, pCO_2 and ancillary data as well as the nitrate climatology will be referred to as "observations", for brevity, keeping in mind that they are really observation-based estimates and not direct observations. Secondly, "model" will exclusively refer to the numerical simulations using the NASA-GISS climate model and by "algorithm", "method" or "technique" we will refer to the clustering technique.

All data products are available in the Ocean Carbon States Database (data.giss.nasa.gov/oceans/carbonstates).

3. Methodology

A schematic diagram of the methodology is presented in Fig. 1. First, the 2D histograms pCO_2-SST are computed from the climatological data, then the histograms are clustered using a statistical method, the k-means clustering method, and finally the regimes or "ocean carbon states" are obtained. The methodology steps will be explained in detail below, using as an example the North Atlantic basin data.

(Figure 1)

pCO_2-SST 2D histograms

pCO_{2SW} values in the North Atlantic span the range 50-450 uatm while sea surface temperatures range between -2 and 30° C. The 2D histograms (Fig. 2) show highest frequency of occurrence for pCO2sw values in the range of 300 to 400 uatm and temperatures in the range of 10 to 30^0C. Certain months (December, January, February and March) show a higher frequency of occurrence of cold temperatures (-2 to 2^0C) and low pCO2sw (50 to 300 uatm) than others. Fig. 2 also reveals that certain histograms appear similar in shape, for example, January – April exhibit an S-shaped curve and no tilt, while June – September exhibit a diagonal tilt that reflects a tendency for higher temperatures to co-locate with higher pCO_2sw values. This being a small dataset of only 12 2D histograms, one could easily sort them into groups of similar shape just by visual inspection only. The methodology presented in this paper seeks to more mechanistically identify these groups, so that it can be confidently applied to larger and more complex datasets. We will call those organized groups, clusters or regimes or "ocean carbon states".

(Figure 2)

It is noted here that despite the broad range of values for both variables, the 2D histograms are very similar regardless of the number of bins chosen for each of the variables.

k-Means clustering

The k-means clustering algorithm (Anderberg, 1973; Jakob and Tselioudis, 2003) partitions the 2D histograms of pCO_2-SST shown in Fig. 2 into a predefined number k of groups, called clusters. In the first step of the algorithm, k histograms are randomly selected and are considered the centroid of each of the k clusters. Each other histogram in the input dataset is then assigned to its nearest centroid by computing the Euclidean distance of each bin of the 2D histogram from the same bin of the centroid. The procedure is repeated N number of iterations, each time the centroid of the resulting group is recalculated, if doing so reduces the sum of distances of each histogram to the centroid. This iterative procedure stops when the squared distance between the mean of each cluster and all the 2D histograms assigned per cluster is minimized (Jain, 2010). More than one iteration (N) is necessary to have convergent clustering results because each analysis initializes at a random cluster centroid. In this paper, convergence is reached after 10 iterations, if not less (Fig. S1 shows how this is determined for the example of the North Atlantic basin).

Sensitivity to predefined number of clusters

To ensure that the chosen number of clusters, k, is representative of the system, typically, one needs to repeat the technique for various values of k, and, using visual inspection, select the optimal value for k when the resulting clusters become repetitive or contain no additional information. Objective methods have been proposed (e.g. Bankert and Solbrig, 2015), where the average radius of each cluster (the distance from the centroid to the most distant member within a cluster) is computed for decreasing k. Bankert and Solbrig (2015) found that when the number of clusters falls below the optimal k, the average radius

grows rapidly. We employ a similar methodology here. First, we use a scoring algorithm that computes the distance of each 2D histogram from the centroid of its cluster. The higher the score, the closer the 2D histograms are to the centroid. The maximum score is 1 which indicates a perfect match. Negative or low values indicate poorly matched histograms that are the farthest from the centroid in the cluster. We run the scoring algorithm for k = 2 through k = 12, since we have 12 2D histograms and therefore there can be up to 12 clusters. Fig. 3a shows the scores only for k = 2, k = 3, and k = 4 as examples of the output of the scoring algorithm. We find that for k=2 and k=3 all 2D histograms are well matched within a cluster (i.e. all scores are high) whereas for k=4 there is one month with a negative score. To further summarize the scoring results, we average all scores for each k and normalize by k. The results are shown in Fig. 3b where we note that the averaged normalized score for k=2 is 0.4 and it then quickly drops to 0.2 for k=3 and to 0.1 for k=4 and it plateaus after that. We choose as the optimal number of clusters the k with the highest score and no significant change of the normalized averaged score thereafter. This choice implies that any reorganization of the 2D histograms within more than 3 clusters will not produce any "tighter" clusters, i.e. clusters where the members are closer together. We must note here, that the method is not entirely objective as one always need to visually inspect the clusters themselves and ensure that the choice of k is indeed the best one.

Figure 3

Data normalization

As noted earlier, pCO_2 and SST have a broad range of values. Specifically, pCO_2 values vary by about 2 orders of magnitude between 50-450 uatm while SST by one order of magnitude between -2 and 30ºC. It is customary in applications of statistical techniques such as clustering, to normalize the data (subtract the mean and divide by the standard deviation) in order to force both datasets to be in the same range of values. However, this is not always necessary (Anderberg, 1973, p. 13; Kaufman et al, 2005, p. 11). In our case, we are not clustering each variable separately in order to determine regression coefficients (as in Lefevre et al., 2005). Rather, we are clustering the 2D histograms and comparing them, in order to obtain groups of similar patterns. In addition, clusters represented in normalized data are not as easily understood physically and as well-represented on geographical maps. Therefore we choose not to normalize the data for the purposes of this study.

4. Results

4.1 The North Atlantic Ocean Carbon States

Figure 4 depicts the regimes for k = 2, k = 3, and k = 4. When only two clusters are predefined, i.e. for k = 2, the first cluster (Regime 2A, Fig 4a) is dominated (30% of the time) by pCO_2-SST pairs in the ranges of 350 – 400 uatm and 25-30ºC. The second cluster (Regime 2B) is dominated (20%) by pCO_{2SW} values within 300 – 350 uatm and SST values in the ranges -2-20ºC. When we choose more clusters initially, i.e. for k=3, Regime 3B is very similar to Regime 2A and Regime 3C is analogous to Regime 2B, in the sense that the regimes have analogous bins of highest frequencies. Regime 3A is a new state

that was unresolved in k = 2, but isn't similar to 3B or 3C. For k = 4, Regime 4B and 4C appear to be almost equivalent, and both derived from Regime 3B, which probably indicates that there is no new information gained by requiring 4 clusters. Similar visual inspection of the results for k > 4 confirms our more objective analysis result that k = 3 is the optimal number of clusters in the pCO_2-SST space in the North Atlantic basin. It should be noted here that because we implement k-clustering to the 2D histograms and not the raw data, there is really no change in the results if we use different number of bins in the histograms or if we use normalized data.

(Figure 4)

Temporal attribution for the North Atlantic carbon states

In order to characterize the ocean carbon states obtained in the previous section, we perform a temporal attribution analysis by determining when each cluster occurs. This is possible because the k-means analysis provides the distance of each member 2D histogram to the centroid of the cluster it belongs, and we are thus able to associate each cluster with certain months of the climatology. Fig. 5b shows that in the North Atlantic basin, regime 1 is represented by months January, February, March, and April and we call this the "winter regime"; regime 2 occurs during June, July, August, September and October and we will thus call it the "summer regime"; regime 3 occurs in May, November, and December and we will call this the "transition regime" because it reflects a mixed season in between the winter and summer regimes. Not surprisingly, these resulting regimes align themselves fairly well with the boreal winter and summer seasons in the North Atlantic. The cold season (winter regime) includes March and April but not November-December which are included, rather, in the transition regime. The warm season (summer regime) includes the months between June and October, again broader than the typical boreal summer. It is not surprising that we recover the seasonal cycle from the 12-month climatology, and probably because our domain is the entire North Atlantic, from the equator to the subpolar regions, these seasons are broader including months from the spring and fall, since the length of each season is different at different latitudes.

Spatial attribution of the North Atlantic carbon states

Next, we describe the geographical distribution of each regime (Fig. 5c). To do so, the frequencies of occurrence associated with each pCO_2-SST bin in Fig. 5a are averaged over the months in each regime and mapped on the North Atlantic basin. We find that in the winter regime, the dominant value pairs (300-350 uatm and 10-20ºC) are found in the subtropical North Atlantic. In contrast, the dominant range of pCO_2-SST pairs in the summer regime occurs in the tropics (values 350 – 400 uatm and 25 – 30ºC). The transition regime shows a mix of the winter and summer regimes.

(Figure 5)

We conclude that the ocean carbon states determined by a 12-month climatology of surface ocean pCO_2 and SST are characterized by a cold season where most persistent value-pairs occur in the subtropical North Atlantic and the subpolar region. In the warm season however, the most persistent value-pairs occur in the tropical Atlantic. For more complex datasets, e.g. when interannual variability is included, we expect to be able to detect regimes that correspond to processes controlled by the El Nino-Southern Oscillation (ENSO) or the North Atlantic Oscillation, for example.

The NASA-GISS climate model North Atlantic carbon states

Next we obtain the ocean carbon states from the GISS model simulations. As described in Section 2.2, we construct the ensemble mean climatologies for pCO_{2SW} and SST for the period 1995-2005 from 5 simulations of Earth's historical climate of the 20th Century performed with the NASA-GISS climate model. We then obtain the 12-monthly 2D histograms from the model climatology and using the same binning groups as in the observations, we obtain the model clusters. It should be noted here again that because we are actually clustering the 2D histograms and not the raw data, our clusters are not sensitive to the number of bins nor to normalization of the datasets prior to cluster analysis.

The sensitivity test for the model clusters is not as clear as in the case of the observations (Fig. 6). Note that there is a plateau after k = 5 and thus it appears that 5 would be a more suitable choice for k. However, as seen in Fig. 6a, some of the additional regimes include only one monthly 2D histogram. We therefore chose here k=3, recognizing that the model clusters have larger uncertainty. In a larger dataset that includes interannual variability, more than a single 2D histograms would be potentially assigned to a regime, reducing that uncertainty.

Fig. 7a shows the ocean carbon states for k=3 while Fig. 7b characterizes their temporal occurrence: model winter regime corresponds to months December, January, February, March, April, and May; the summer regime corresponds to months July, August, September, October, and November; the transition regime corresponds only to June. There is therefore good agreement with the regimes from observations (Fig. 5b). The model winter regime is somewhat broader than the observed by two months (December and May) while the model summer regime is lagging by one month (starts in July, while in the observations starts in June).

The regimes (clusters) themselves are similar to the observations (comparing Fig. 5a and Fig. 7a) in that the same bins of most likely values are identified but with somewhat different frequencies of occurrence. As an example of comparison between the temporal regimes, for the winter regime both the model and the observations show that the dominant pairs are in the range of 300–350 uatm and 20–25^{0}C, at 30% and 25% relative frequency. However, other weaker pairs are not well represented in the model, e.g. the range 50–200 uatm and -2–10ºC. During the winter regimes, the highest frequency of occurrence (25%) is for the pair of values 300-350 uatm and 10-20ºC, whereas in the model the same pair of values is found 30% of the time. Similarly,

the summer and transition regime highest frequency pairs are well simulated. In contrast, the value pairs 200-350 uatm for very cold temperatures are not well represented in the model.

We also find (Fig. 7c) that in the winter regime, the dominant value pairs (300-350 uatm and 10-20^{0}C, identified in Fig. S2) are found in the subtropical North Atlantic with higher frequency in the model (darker shading) than in the observations. The GISS model however underestimates the frequency of occurrence of the value pairs in the subpolar region (values 50–350 uatm and -2–10ºC, identified in Fig. S2). In contrast, the dominant range of pCO_2-SST pairs in the summer regime occurs in the tropics (values 350–400 uatm and 25–30ºC; Fig. 7c) and is of higher frequency in the observations than the model. In other words, the model underestimates the extent of the tropical summer regime but reproduces well the other parts of the summer regime. The transition regime shows a mix of the winter and summer regimes for both observations and model. The model results in this regime indicate higher frequency in the subpolar region than in the observations.

(Figure 6)

(Figure 7)

Model North Atlantic air-sea flux of CO_2 error analysis and bias attribution

The ocean carbon states identified in the previous section can provide a framework for model assessment against the observations. In this section we seek to identify biases in the simulated flux of CO_2 and attribute them to leading biases in physical and biogeochemical processes. For this analysis, model data will be composited on the observational regimes, i.e. the model climatology will be averaged over those months that define each observational cluster in Fig. 5b.

Figure 8 depicts the air-sea flux of CO_2 composited over the observed temporal regimes in both observations (Fig. 8a) and model output (Fig. 8b). In the winter regime, model outgassing (in shades of blue) is confined only to the tropics, whereas in the observations there is also a tongue of outgassing at about 60ºN. Similarly, in the transition regime, the model has a more extended uptake region in the subpolar North Atlantic than in the observations. While the summer regime is better represented in the model than the other two regimes, all three regimes show that model uptake is stronger than in observations at mid and high latitudes and outgassing is also stronger in the model in mid to low latitudes.

(Figure 8)

To trace the source of model biases in the air-sea flux of CO_2 we need a better understanding of the physical or biogeochemical processes that control the air-sea flux of CO_2 in the model. Using the clustering analysis results from the previous section, we can investigate the underlying processes that might be responsible for the bias.

The process attribution is performed using a Taylor expansion of the model bias as shown in Eq. (5). The model flux bias, ΔF, depends on the biases of pCO_{2SW}, SST, salinity (SSS) and wind speed (wspd) such that:

$$\Delta F \sim \frac{\partial F}{\partial pCO2_{SW}} \Delta pCO2_{SW} + \frac{\partial F}{\partial SST} \Delta SST + \frac{\partial F}{\partial SSS} \Delta SSS + \frac{\partial F}{\partial wspd} \Delta wspd \quad (5)$$

where Δq is the bias of the variable q, defined as the root mean squared error (RMSE) between the observations and the model and q is any of the variables {pCO_{2SW}, SST, SSS, wspd}. $\frac{\partial F}{\partial q}$ is a weight term that represents dependence of the flux on that variable and is determined by the slope of a linear fit in the scatter plot of the flux F with each variable q for each carbon state. Since the North Atlantic basin is a very broad basin, both zonally and meridionally, and because the carbon states regional distribution (Fig. 5c) is quite complex, we identify areas where the linear fits will be more appropriate approximations of the {F,q} relationships. The subpolar region where the value pairs are -2 to 10ºC and 50 to 350 uatm, a subtropical region (10 to 20ºC, 300 to 350 uatm), and a tropical region (20 to 30ºC, 300 to 400 uatm) are demarcated in Fig. S2. Results of the regional scatter plots and the linear fit for each regime are shown in Fig. S3 and are synthesized in Figure 9. Each contribution term (each term in the right hand side of Eq. 5) is calculated from the multiplication of the weights and the RMSEs. Figure 9 then shows that over most of the North Atlantic, the flux biases are attributed mainly to errors in the pCO_{2SW}, although in subpolar regions other terms such as salinity biases and wind speed biases become important. It therefore makes sense to further investigate biases in pCO_{2SW} and the processes these are attributed to, as presented in Eq. (2).

(Figure 9)

Similarly to Eq. (5):

$$\Delta pCO2_{SW} \sim \frac{\partial pCO2_{SW}}{\partial SST} \Delta SST + \frac{\partial pCO2_{SW}}{\partial SSS} \Delta SSS + \frac{\partial pCO2_{SW}}{\partial WSPD} \Delta WSPD + \frac{\partial pCO2_{SW}}{\partial NITRATE} \Delta NIT \quad (6)$$

We perform the Taylor expansion of the bias for each of the regimes that we computed, calculating the weights and RMSEs in the same way as described for CO_2 flux biases. The estimates of the linear fit slopes of the scatter plots are shown in Fig. S4 and the composites of the contributions in Fig. 10.

Overall Fig. 10 shows that the biases in the subpolar region are larger than anywhere else in the North Atlantic basin, as the contributions to the bias in pCO_2sw are an order of magnitude larger there. Specifically, in the subpolar region, wind speed biases emerge as responsible for the winter and transition regime biases in pCO_2sw while salinity biases dominate the summer bias in pCO_{2SW}. In the winter and transitional months, the quasi-cyclonic subpolar gyre, driven by energetic winds and wind-outbreaks, leads to Ekman divergence in the surface layer that controls the pCO_2sw biases near the coast. At the same time,

winter-time convective mixing is responsible for biases in the strength of the Meridional Overturning Circulation that are known to influence open ocean pCO_2sw (Romanou et al,. 2017). In the summer regime, GISS model sea-ice concentration is higher than observed hence melting will lead to significant surface salinity biases. Inaccurate model representation of the magnitude and fluctuations of the cyclonic wind stress curl as well as the sea-ice retreat and associated salinity changes are probably responsible for deficient physical characterization of the model ocean circulation, which would result in misrepresentations of the pCO_{2SW} and thus the CO_2 flux in the model.

In the subtropics, nitrate is found to be the largest contributor for the winter regime biases, wind speed is the main contributor in the summer, and salinity is the main contributor for the transition regime. The subtropics are characterized at the surface by anticyclonic circulation and a strong western boundary current, the Gulf Stream. Gyre subduction supports downwelling which brings nutrients and pCO_2 to depth. Nitrate utilization by ocean biology during the winter regime is probably inaccurate in the model while wind speed biases are known to be larger in the summer than the winter regime in the model.

In the tropics, biases in wind speed, nitrate and salinity are again found to be important. Here, nitrate biases, which are relatively higher in oligotrophic regions (Arteaga et al,. 2015), are probably due to misrepresentation of nitrogen fixation in the GISS climate model. Wind speed and salinity biases are associated with well-known biases in the intensity and position of the Inter Tropical Convergence Zone (ITCZ) that controls cloudiness, temperature gradient, and rainfall. The ITCZ moves north in the summer and south in the winter, therefore a wind speed bias in the transition regime in the model could be explained by an inaccurate model reproduction of how the ITCZ affects the wind during its transitional movement. The ITCZ increases precipitation, thus decreasing salinity, therefore how salinity changes by season as a result of the shifting ITCZ could explain the winter regime bias.

(Figure 10)

4.2 The Southern Ocean Carbon States

The application of cluster analysis in the Southern Ocean is presented here similarly as in the North Atlantic, with the purpose of examining whether the technique will also be able to identify some known aspects of the Southern Ocean carbon cycle. For brevity, though, the observation-based and the model regimes will be presented alongside one another and will be followed by the model-error attribution analysis.

As done for the North Atlantic basin, we look at probability density distributions of observations in the Southern Ocean which show that both SST and pCO_{2SW} exhibit a broad range of values. Temperatures range between -3 and 20ºC whereas pCO_{2SW} values span 20 and 400 uatm. The 12 monthly 2D histograms of pCO_{2SW} and SST in the Takahashi dataset are shown in Fig. S5 (Supplementary Material Figure 5).

Fig. 11a shows the scores of each regime for k=2, 3, and 4 and Fig. 11b the results of the sensitivity test to the choice of k. Following the criterion established in section 3, the largest score change is for k=3 and k=4. However, Fig. 11a shows that the overall score is better for k=3, because more months have larger individual scores, with the exception of one month (November) which has a slightly negative score. Additionally, visual inspection of apparent patterns in the 12 monthly 2D histograms (Fig. S5) also corroborates the choice of k=3 as the optimal value for the number of clusters. Again, the small climatological dataset leads to some uncertainty in determining k.

A value of k = 3 is also chosen for the model analysis based on Fig. 11c and 11d which show that this is also the optimal number of clusters. There is added ambiguity in the choice of k here, in addition to that due to the small dataset, which arises from the fact that the Southern Ocean is a very broad basin zonally and different processes become important in different regions more so than in the narrow North Atlantic basin, which make the choice of k not as clear as it was in the North Atlantic.

(Figure 11)

The observed and the model ocean carbon states are shown in Fig. 12. In the summer regime, which includes January, February and March (see Fig. 13 and explanation below), the highest frequency pair-values (i.e. the most persistent pairs) are found around 20% of the time in the observations and 25% of the time in the model for the ranges of 250-350 uatm and 0-5ºC. Another range of pair-values which also shows high frequency of occurrence (25%) is found for warmer temperatures (10-20ºC) and the same range of pCO_2, but the GISS model misplaces it towards higher values of pCO_2 (350-400 uatm). The winter regime comparison shows that the model captures the low pCO_2, low temperature state (20-150 uatm, -3-0ºC) well (30% of the time in observations and in the model). The mid-range pCO_2, high temperature state (250-350 uatm, 10-20ºC) is not as well represented in the model. The observations there show the highest frequencies of occurrence for higher pCO_2 (350-400 uatm) whereas the model for lower pCO_2 (250-350 uatm). Comparison between the transition regimes reveals much less correspondence between observations and the GISS model, considering the high frequency states in observations are quite different than in the model (e.g. observations high frequency states: 20-150 uatm, -3-0ºC; 250-350 uatm, 10-20ºC; 350-400 uatm, 0-5ºC; model high frequency state: 250-350 uatm; 0-10ºC).

Figure 12

Temporal attribution for the Southern Ocean carbon states

Temporal attribution, which is estimated using the method described in section 4.1, is shown for both the model and the observations in Fig. 13. Note that all subsequent analysis considers the austral seasons when referring to “winter” and “summer”. The temporal attribution shows that the observations and model data are clustered in regimes that correspond to

almost the same months. The only difference is that November is accounted for in the transition regime for the observations as opposed to the winter regime for the model. It is noted, however, that November is technically a "poorly matched" 2D histogram in the observations cluster routine, as discussed earlier.

(Figure 13)

Spatial attribution of the Southern Ocean carbon states

In order to further explain the model and observed Southern Ocean regimes shown in Fig. 12, the frequencies of occurrence of each bin in the cluster are mapped onto the Southern Ocean (Fig. 14). In the coastal Antarctic divergence zone, SST varies within -3–3ºC and pCO_2 20–250 uatm, in the Antarctic convergence zone SST ranges 3–10ºC and pCO_2 250–400 uatm, and in the subtropical convergence zone SST lies within 10–20ºC and pCO_2 250–400 uatm. Despite the strong temporal attribution agreement between the model and the observations, the regional attributions show much less correspondence. For example, in the summer regime, the observations show the highest frequency of occurrence between 250–350 uatm and 10–20ºC along the subtropical convergence zone (roughly along 40ºS) while the highest frequency of occurrence for the model is for the pair 250–350 uatm and 0–5ºC, occurring in the coastal region (poleward of the divergence zone along 60ºS). In the winter regime, both observations and the GISS model show highest frequency of occurrence for the value-pairs nearest to the coast where the pCO_2 is low and the temperatures coldest (20–150 uatm and -3–10ºC). Further offshore, the model highest frequency values occur for lower pCO_2 (250-350 uatm) than in the observations (350-400 uatm). Lastly, in the transition regime, the model shows a much higher persistency of values in the range (250-350 uatm and 0-10ºC) than in the observations.

(Figure 14)

It is therefore evident that the GISS model does not reproduce well the observed ocean carbon states in the Southern Ocean. There is a tendency to persistently underestimate pCO_2-SST values closer to the coast and overestimate it near the subtropical convergence zone, during the warm season. Whereas, during the cold season, the model captures well the divergence zone regime but the errors further offshore switch: the model now overestimates pCO_2-SST value-pair frequency of occurrence in the Antarctic convergence zone but is performing better in the subtropical convergence zone. The transition regime is not well represented, indicating that not enough regimes are chosen to adequately describe ocean carbon states in this small dataset.

Model Southern Ocean air-sea flux of CO_2 error analysis and bias attribution

Comparing the CO_2 flux composites on the observed regimes (Fig. 15) shows significant discrepancies between observations (Fig. 15a) and model (Fig. 15b). To better understand where these discrepancies come from, we perform an error attribution analysis, as in section 4.1 above. In the summer regime (January through March), outgassing in the model is restricted to the subtropical convergence zone whereas in the observations it is more localized and closer to Antarctica. At the same time, model uptake is stronger, confined to the coast and over a broader area than in the observations. This is consistent with the

result from the previous section, where the model underestimates pCO_2 and SST in the Antarctic convergence zone. In the winter regime (June through October), the entire model basin is a sink for CO_2 whereas in the observations there is a zonally confined outgassing belt south of 50ºS and an uptake belt north of it. Again this result is consistent with the earlier finding that the model winter regime is closer to the observed near the Antarctic coast. The transition regime shares a mix of the same discrepancies as in the summer and winter regimes.

(Figure 15)

Bias attribution is computed for the three zonally-defined regions indicated in Fig. S6. Based on the bias computations in Eq. (5) for pCO_{2SW}, SST, salinity, and wind speed with respect to CO_2 flux, pCO_{2SW} is again shown to be the driving variable in most of the flux biases in the Southern Ocean (Fig. S7; Fig. S8 for scatter plots). We therefore seek to understand the processes that control the pCO_2 biases in the model, using the Taylor expansion in Eq. (6) (Fig.16; Fig. S9 for scatter plots).

For almost all regimes and regions, biases in nitrate are large partly because of lack of a closed, state-of-the art nitrogen cycle representation in the climate model. On the other hand, observations are too scarce in the region, due to inclement weather and biases to specific seasons, so there is large observational uncertainty associated with the Takahashi climatology in the Southern Ocean. The model skill would be more adequately assessed as more in situ measurements are made (e.g. from the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) experiment; Johnson et al., 2017). Nevertheless, the model underestimates surface nitrates in the Southern Ocean in particular because of a large nitrate deficit in the subsurface ocean which upwells in the subantarctic zone and flows into the Antarctic Circumpolar Current region. This is related to processes such as denitrification and accurate remineralization in the deep ocean. SST is the second-most dominating variable for biases in the coastal Antarctic. Inspection of the model biases shows that south of 70ºS the model water column is colder than in observations, hence upwelling there will bring colder waters near the surface. Interestingly, surface salinity biases are relatively very important in the region south of the subtropical convergence zone which suggests that a study of water mass formation in that region in the model and the observations would better explain the biases.

(Figure 16)

Data availability

Data and analysis scripts can be accessed at www.data.giss.nasa.gov/oceans/carbonstates.

5. Conclusions

This proof-of-concept study presents the k-means cluster analysis and the determination of the regimes called "ocean carbon states" in observation-based data of the ocean carbon cycle. A method is described here to more objectively determine the optimal number of clusters for the cluster analysis. The study also explores how to characterize the ocean carbon states temporally and spatially in order to determine the physical-biogeochemical processes related to each carbon state. Composites of the CO_2 flux and a quantitative exploration of the effect of each field on pCO_{2SW} bias is also demonstrated.

In this study, pCO_2 and SST were chosen as the two variables that co-determine the carbon states, based on the fact that they both play critical roles on the biogeochemistry and the physics of the ocean system and control the flux of CO_2. One may choose different variables and it would be interesting to see whether and how the regimes depend on the choice of variables.

We have also tested the importance of the choice of k. A main caveat of k-means cluster analysis is that k must be predetermined through reasoning that is subject to personal bias. However, we show that by assessing the clusters from multiple angles (i.e. the score plots, the sensitivity test, visual inspection, analyzing cluster outputs with increasing k, temporal attribution), it is possible to determine an optimal k that is semi-objective. Even so, in this proof of concept study, we acknowledge the uncertainty in our choice of k as a result of the small dataset of 12 monthly 2D histograms, which occasionally results in there being only a small number of histograms per cluster.

The ocean carbon states we obtain from this climatological year data set are interesting. We found that the subtropical North Atlantic is the dominant feature in the cold season regime (the months January through April). In the same regime, the subpolar North Atlantic also features prominently which is associated with the high variability in this area due to sea ice retreat, the spring bloom and the winter-spring convection. In contrast, the tropical Atlantic dominates the warm months June through October, while the subtropical and subpolar regions play a smaller role. The transition regime, which is comprised by months that do not entirely fall in the winter or the summer regimes, shows again the lower tropics and the subtropical gyre to be more active. We would expect that a longer dataset that includes natural variability as well as the effects of longer-term climate and anthropogenic trends would result in more carbon states and hence that would be an interesting extension of the present study.

The NASA-GISS model carbon states in the North Atlantic are similar, both in temporal as well as spatial characterization, to the observed ones, with better model skill in the summer than in the winter. Specifically, the model overestimates the importance of the subtropical gyre and underestimates the subpolar gyre during the cold months. During summer, the model underestimates the tropics but not significantly. The transition months are found to behave differently in the model, although that might be a result of the small size of our input datasets.

In the Southern Ocean, during the warmer months (January-March), the observational states are more persistent along 40ºS, the subtropical convergence zone, while the colder season has prominent states (higher persistency) mostly along the Antarctic coast. The transition regime shows similar degree of variability across the entire Southern Ocean.

While the GISS model agrees in the temporal characterization of the ocean carbon states, it diverges from the observational spatial attribution particularly in the summer and the transition regimes. It is of note here, however, that the Takahashi climatology is far more uncertain in the Southern Ocean (Takahashi et al., 2009) than it is in the North Atlantic, and therefore the model lack of skill might be not as alarming. New observations in the area (e.g. from SOCCOM) will greatly benefit studies such as this.

Error analysis of the model response helps explain the GISS model biases. Applying k-means clustering analysis in two main regions of the world that are known to be critical for the global ocean carbon cycle, namely the North Atlantic region and the Southern Ocean, defines the priorities for model improvement: in the North Atlantic biases in surface salinity, wind speed and surface temperature whereas in the Southern Ocean priorities are nitrate and surface salinity. Clearly the GISS climate model would benefit from more realistic representation of the nitrogen cycle in the ocean as a whole.

The goal of this study is to enable us to apply this k-means clustering to "big" data, in order to find the interannual and regional patterns in larger, higher frequency climate datasets. This extended application will allow researchers to gain a much more comprehensive insight and intuition for physical systems by mechanistically and impartially grouping multiple variables that form the prominent features of these networks. Other variable pairs besides pCO_2 and SST will also be explored, such as CO_2 flux and chlorophyll, in order to assess other drivers in Eq. (1). Finally, higher order clustering and classification techniques will be analyzed in order to determine the most efficient and successful method for understanding the ocean carbon cycle.

All routines and datasets used in this study are freely available in the Ocean Carbon States page of the NASA-Goddard Institute for Space Studies web portal (data.giss.nasa.gov/oceans/carbonstates).

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center.
Funding was provided by NASA-ROSES Modeling, Analysis and Prediction 2013 NNX14AB99A-MAP for GISS Model-E development and NNX15AJ05A NASA Cooperative Agreement 2015-2018.
Data used to generate figures, graphs, plots, as well as analysis were archived at NCCS dirac repository, numerical codes are maintained and archived at GISS and all data and codes are available upon request from A. Romanou. Clustering analysis was performed using the MATLAB ver 2015 computing environment. The authors wish to thank Robert Schmunk for his help in setting up the Zenodo page and the GISS portal.

References

Anderberg, M. R, 1973. Cluster analysis for applications. New York: Academic Press.

Arteaga, L., M. Pahlow, and A. Oschlies, 2015. Global monthly sea surface nitrate fields estimated from remotely sensed sea surface temperature, chlorophyll, and modeled mixed layer depth. Geophys. Res. Lett, 42, 1130–1138. doi: 10.1002/2014GL062937

Bankert, R. L., J. E. Solbrig, 2015. Cluster Analysis of A-Train Data: Approximating the Vertical Cloud Structure of Oceanic Cloud Regimes. J. Appl. Meteor. Climatol. 54, 996–1008. https://doi.org/10.1175/JAMC-D-14-0227.1

Bodas-Salcedo, A., K. D.Williams, M.A. Ringer, I. Beau, J.N. Cole, J. Dufresne, T. Koshiro, B. Stevens, Z. Wang, and T. Yokohata, 2014. Origins of the Solar Radiation Biases over the Southern Ocean in CFMIP2 Models. J. Climate, 27, 41–56, https://doi.org/10.1175/JCLI-D-13-00169.1

Boyer, T.P., J.I. Antonov, O.K. Baranova, C. Coleman, H.E. Garcia, A. Grodsky, D.R. Johnson, R.A. Locarnini, A.V. Mishonov, T.D. O'Brien, C.R. Paver, J.R. Reagan, D. Seidov, I.V. Smolyar, and M.M. Zweng, 2013. World Ocean Database 2013, NOAA Atlas NESDIS 72, S. Levitus, Ed., A. Mishonov, Technical Ed.; Silver Spring, MD, 209 pp., http://doi.org/10.7289/V5NZ85MT

Dlugokencky, E. and P. Tans, Trends in atmospheric carbon dioxide. National Oceanic & Atmospheric Administration, Earth System Research Laboratory (NOAA/ESRL), [last accessed 2014 8 August]. http:// www.esrl.noaa.gov/gmd/ccgg/trends.

Fay, A. R. and G.A. McKinley, 2014. Global open-ocean biomes: mean and temporal variability, Earth Syst. Sci. Data, 6, 273-284, https://doi.org/10.5194/essd-6-273-2014

Halverson, M. J. and S.W. Fleming, 2015. Complex network theory, streamflow, and hydrometric monitoring system design, Hydrol. Earth Syst. Sci., 19, 3301-3318, https://doi.org/10.5194/hess-19-3301-2015

Hoffman, F.M., W.W. Hargrove, R.T. Mills, S. Mahajan, D.J. Erickson, R.J. Oglesby, 2008. Multivariate Spatio-Temporal Clustering (MSTC) as a data mining tool for environmental applications, in: M. S'anchez-Marr'e, J. Ḃejar, J. Comas, A.E. Rizzoli, G. Guariso (Eds.), Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software Society (iEMSs 2008), Barcelona, Catalonia, Spain.

Hoffman, F.M., J.W. Larson, R.T. Mills, B.J. Brooks, A.R. Ganguly, W.W. Hargrove, J. Huang, J. Kumar, R.R. Vatsavai, 2011. Data Mining in Earth System Science (DMESS 2011). Procedia Computer Science, 4, 1450-1455, https://doi.org/10.1016/j.procs.2011.04.157

Hugue, F., M. Lapointe, B.C. Eaton, A. Lepoutre, 2016. Satellite-based remote sensing of running water habitats at large riverscape scales: Tools to analyze habitat heterogeneity for river ecosystem management. Geomorphology, 253, 353-369, https://doi.org/10.1016/j.geomorph.2015.10.025

Jain, Anil K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666, http://dx.doi.org/10.1016/j.patrec.2009.09.011

Jakob, C., G. Tselioudis, 2003. Objective identification of cloud regimes in the Tropical Western Pacific. Geophys. Res. Lett., 30, doi:10.1029/2003GL018367

Johnson, K. S., J.N. Plant, L.J. Coletti, H.W. Jannasch, C.M. Sakamoto, S.C. Riser, D.D. Swift, N.L. Williams, E. Boss, N. Haentjens, L.D. Talley, J.L. Sarmiento, 2017. Biogeochemical sensor performance in the SOCCOM profiling float array. J. Geophys. Res. Oceans, 122, 6416-6436, doi:10.1002/2017JC012838

Kaufman, L. and P. Rousseauw, 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey: John Wiley & Sons, Inc.

Landschützer, P., N. Gruber, D. C. E. Bakker, U. Schuster, S. Nakaoka, M. R. Payne, T. Sasse, and J. Zeng, 2013. A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. Biogeosciences, 10, 7793–7815, doi:10.5194/bg-10- 7793-2013

Landschützer, P., N. Gruber, D. C. E. Bakker, and U. Schuster, 2014. Recent variability of the global ocean carbon sink. Global Biogeochem. Cycles, 28, 927–949, doi:10.1002/2014GB004853

Lefèvre, N., A.J. Watson, and A.R. Watson, 2005. A comparison of multiple regression and neural network techniques for mapping in situ pCO2 data. Tellus B, 57, 375–384, doi:10.1111/j.1600-0889.2005.00164.x

Nakaoka, S., M. Telszewski, Y. Nojiri, S. Yasunaka, C. Miyazaki, H. Mukai, and Usui, N., 2013. Estimating temporal and spatial variation of ocean surface pCO2 in the North Pacific using a selforganizing map neural network technique. Biogeosciences, 10, 6093–6106, doi:10.5194/bg-10-6093-2013

Oreopoulos L., C. Nayeong, D. Lee, S. Kato, 2016. Radiative effects of global MODIS cloud regimes. J. Geophys. Res., 121, 2299–2317, doi:10.1002/ 2015JD024502

Peron, T.K.D., C.H. Comin, D.R. Amancio, L. da F. Costa, F.A. Rodrigues, and J. Kurths, 2014. Correlations between climate network and relief data, Nonlin. Processes Geophys. 21, 1127-1132, doi:10.5194/npg-21-1127-2014

Phillips, J.D., W. Schwanghart, T. Heckmann, 2015. Graph theory in the geosciences. Earth-Science Reviews, 143, 147-160, https://doi.org/10.1016/j.earscirev.2015.02.002

Radebach, A., R.V. Donner, J. Runge, J.F. Donges, J. Kurths, 2013. Disentangling different types of El Niño episodes by evolving climate network analysis. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 88, doi: 10.1103/PhysRevE.88.052807

Romanou, A., W.W. Gregg, J. Romanski, M. Kelley, R. Bleck, R. Healy, L. Nazarenko, G. Russell, G.A. Schmidt, S. Sun, and N. Tausnev, 2013. Natural air-sea flux of CO2 in simulations of the NASA-GISS climate model: Sensitivity to the physical ocean model formulation. Ocean Model., 66, 26-44, doi:10.1016/j.ocemod.2013.01.008

Romanou, A., J. Marshall, M. Kelley, and J. Scott, 2017. Role of the ocean's AMOC in setting the uptake efficiency of transient tracers. Geophys. Res. Lett., 44, 11, 5590-5598, doi:10.1002/2017gl072972

Rossow, W.B., Y.-C. Zhang, and J. Wang, 2005. A statistical model of cloud vertical structure based on reconciling cloud layer amounts inferred from satellites and radiosonde humidity profiles. J. Climate, 18, 3587-3605, doi:10.1175/JCLI3479.1

Sarmiento, J.L., and N. Gruber, 2006. Ocean Biogeochemical Dynamics. Princeton University Press.

Sasse, T.P., B.I. McNeil, and G. Abramowitz, 2013. A new constraint on global air-sea CO2 fluxes using bottle carbon data, Geophys. Res. Lett., 40, 1594–1599, doi:10.1002/grl.50342

Takahashi, T., S. C. Sutherland, R. Wanninkhof, C. Sweeney, R. A. Feely, D. W. Chipman, B. Hales, G. Friederich, F. Chavez, A. Watson, D. C. E. Bakker, U. Schuster, N. Metzl, H. Yoshikawa-Inoue, M. Ishii, T. Midorikawa, Y. Nojiri, C. Sabine, J. Olafsson, Th. S. Arnarson, B. Tilbrook, T. Johannessen, A. Olsen, R. Bellerby, A. Körtzinger, T. Steinhoff, M. Hoppema, H.J.W. de Baar, C.S. Wong, B. Delille and N. R. Bates, 2009. Climatological mean and decadal changes in surface ocean pCO2, and net sea-air CO2 flux over the global oceans. Deep-Sea Res. II, 56, 554-577. doi: 10.1016/j.dsr2.2008.12.009

Telszewski1, M., A. Chazottes, U. Schuster, A.J. Watson, C. Moulin, D.C.E. Bakker, M. González-Dávila, T. Johannessen, A. Körtzinger, H. Lüger, A. Olsen, A. Omar, X.A. Padin, A.F. Ríos, T. Steinhoff, M. Santana-Casiano, D.W.R. Wallace, and R. Wanninkhof, 2009. Estimating the monthly pCO2 distribution in the North Atlantic using a self-organizing neural network. Biogeosciences, 6, 1405–1421, doi:10.5194/bg-6-1405- 2009

Trans Mills, R., F.M. Hoffman, J. Kumar, W.W. Hargrove, 2011. Cluster Analysis-Based Approaches for Geospatiotemporal Data Mining of Massive Data Sets for Identification of Forest Threats. Procedia Computer Science, 4, 1612-1621, https://doi.org/10.1016/j.procs.2011.04.174

Trochta, J.T., C.B. Mouw, T.S. Moore, 2015. Remote sensing of physical cycles in Lake Superior using a spatio-temporal analysis of optical water typologies. Remote Sensing of Environment, 171, 149-161, https://doi.org/10.1016/j.rse.2015.10.008

Tselioudis, G., W. Rossow, Y. Zhang, and D. Konsta, 2013. Global Weather States and Their Properties from Passive and Active Satellite Cloud Retrievals. J. Climate, 26, 7734–7746, https://doi.org/10.1175/JCLI-D-13-00024.1

Wanninkhof, R., 1992. Relationship between wind speed and gas exchange over the ocean. J. Geophys. Res., 97(C5), 7373–7382, doi:10.1029/92JC00188

Wanninkhof, T.,G.-H. Park, T. Takahashi, C. Sweeney, R. Feely, Y. Nojiri, N. Gruber, S.C. Doney, G.A. McKinley, A. Lenton, C. Le Quéré, C. Heinze, J. Schwinger, H. Graven, and S. Khatiwala, 2013. Global ocean carbon uptake: magnitude, variability and trends. Biogeosciences, 10, 1983-2000, http://dx.doi.org/10.5194/bg-10-1983-2013

Williams, K., and M. Webb, 2009. A quantitative performance assessment of cloud regimes in climate models. Clim. Dyn., 33, 141–157, doi:10.1007/s00382-008-0443-1

Wood, R., M. Wyant, C.S. Bretherton, J. Rémillard, P. Kollias, J. Fletcher, J. Stemmler, S. de Szoeke, S. Yuter, M. Miller, D. Mechem, G. Tselioudis, J.C. Chiu, J.A. Mann, E.J. O'Connor, R.J. Hogan, X. Dong, M. Miller, V. Ghate, A. Jefferson, Q. Min, P. Minnis, R. Palikonda, B. Albrecht, E. Luke, C. Hannay, and Y. Lin, 2015. Clouds, Aerosols, and Precipitation in the Marine Boundary Layer: An Arm Mobile Facility Deployment. Bull. Amer. Meteor. Soc., 96, 419–440, https://doi.org/10.1175/BAMS-D-13-00180.1

Zagouras, A., A. Kazantzidis, E. Nikitidou, A.A. Argiriou, 2013. Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. Solar Energy, 97, 1-11, https://doi.org/10.1016/j.solener.2013.08.005

Figures:

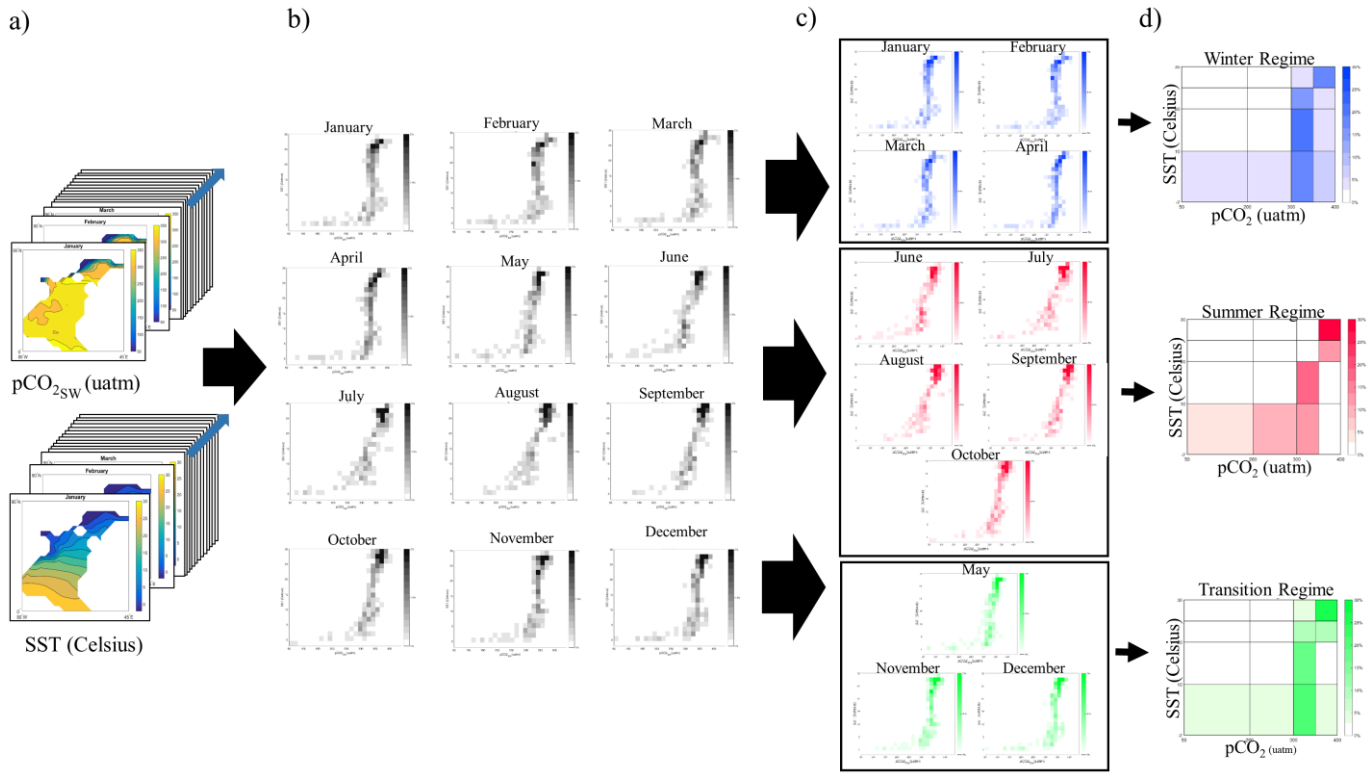

Figure 1: Schematic diagram of the clustering methodology used in this paper: a) 12 monthly mean climatological year data of two variables, pCO_2 and SST, b) monthly 2D histograms, c) clustering of the 2D histograms into groups by similarity in the bivariate distributions, d) clusters resulting when k=3 is assumed.

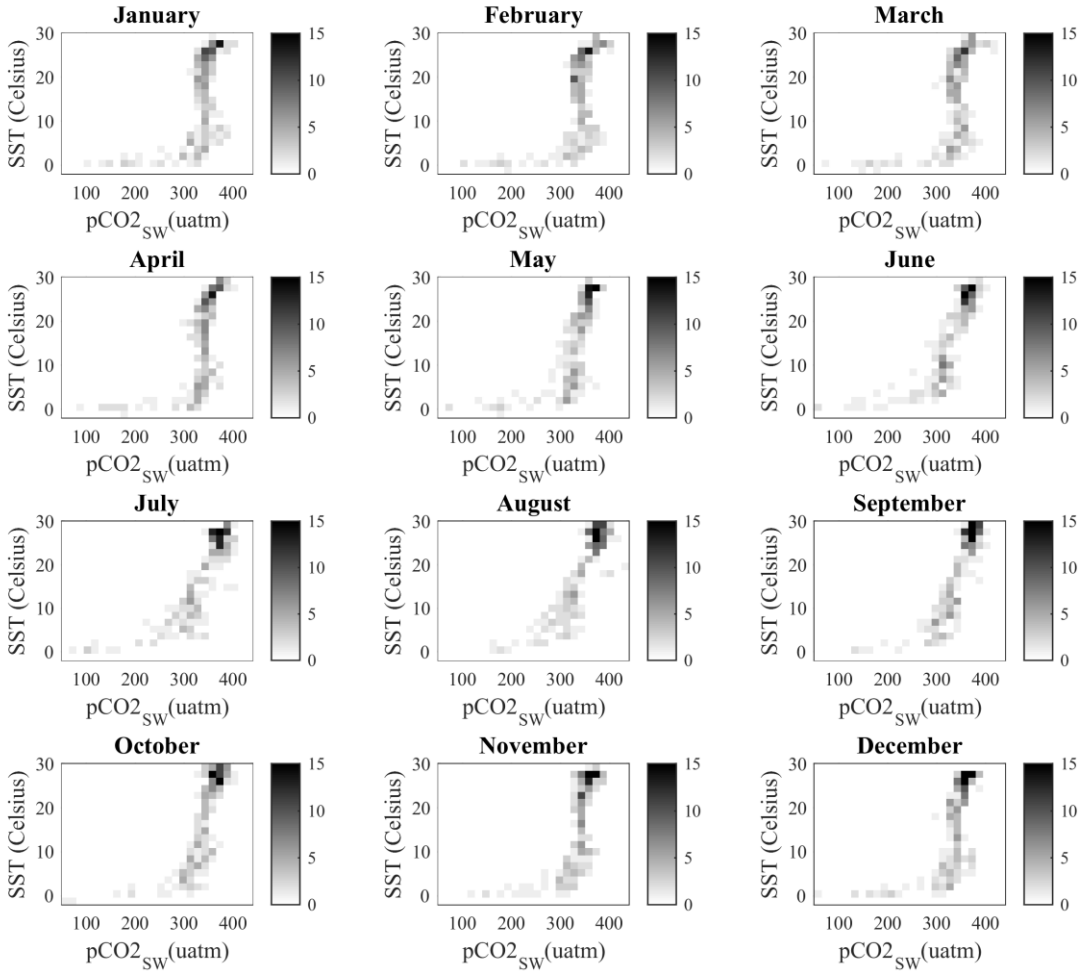

Figure 2: Monthly 2D histograms of partial pressure of CO_2 in the surface water (pCO_2sw) and sea surface temperature (SST) in the North Atlantic (defined as 80°W to 45°E, 0° to 90°N) from the Takahashi observational dataset. The horizontal axis is pCO_2sw (uatm) and the vertical axis is SST (ºC). The bin interval is 15 uatm and 1.6ºC. The colorbar describes the actual frequency of occurrence of each bin.

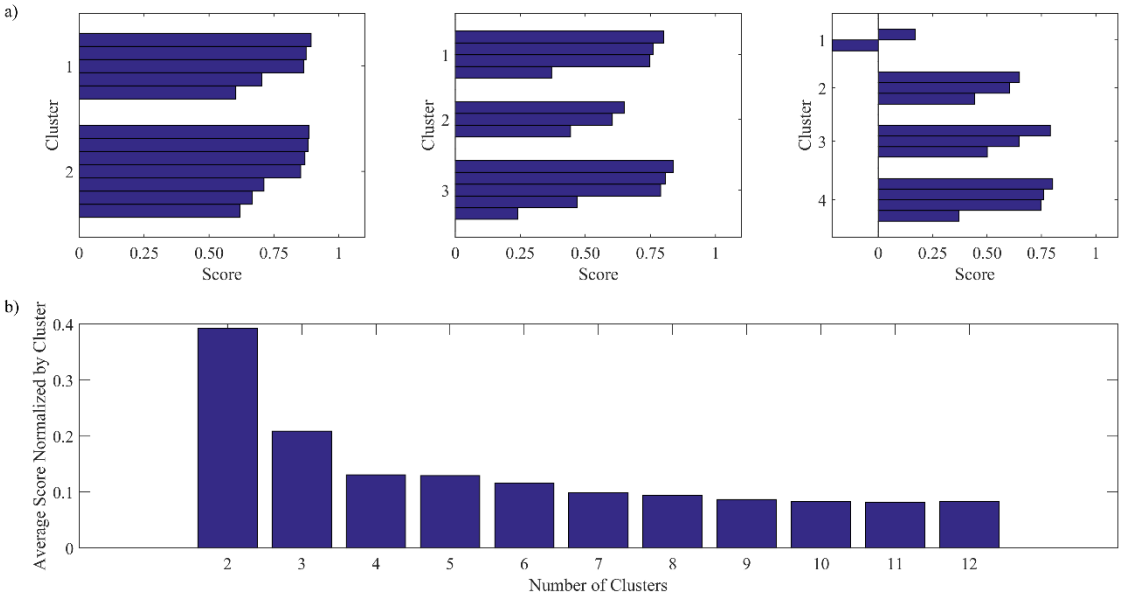

Figure 3: a) Scores for each cluster analysis of observational data in the North Atlantic for k = 2, k = 3, k = 4, where cluster k is the predetermined number of clusters and each bar represents the score per 2D histogram. b) Average scores for each cluster analysis with increasing k, from k=2,…,12, normalized by the number of clusters k.

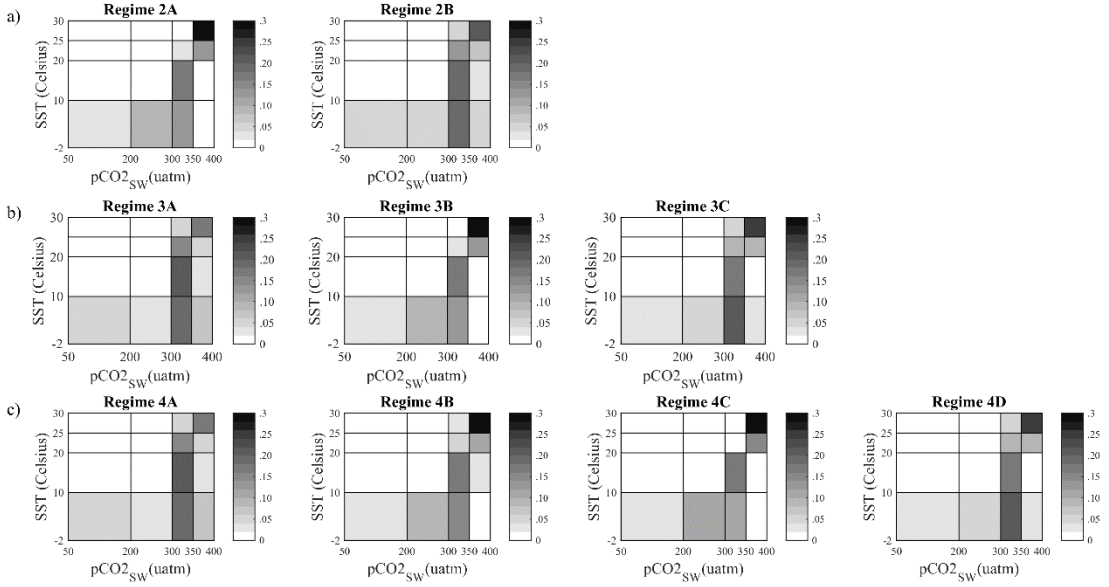

Figure 4: Cluster analysis output (regimes) for a) k = 2, b) k =3, and c) k = 4 for the North Atlantic, from the Takahashi observational dataset. The colorbar represents the relative frequencies of occurrence of each value-pair interval, i.e. the frequencies are divided by total number of frequencies per regime.

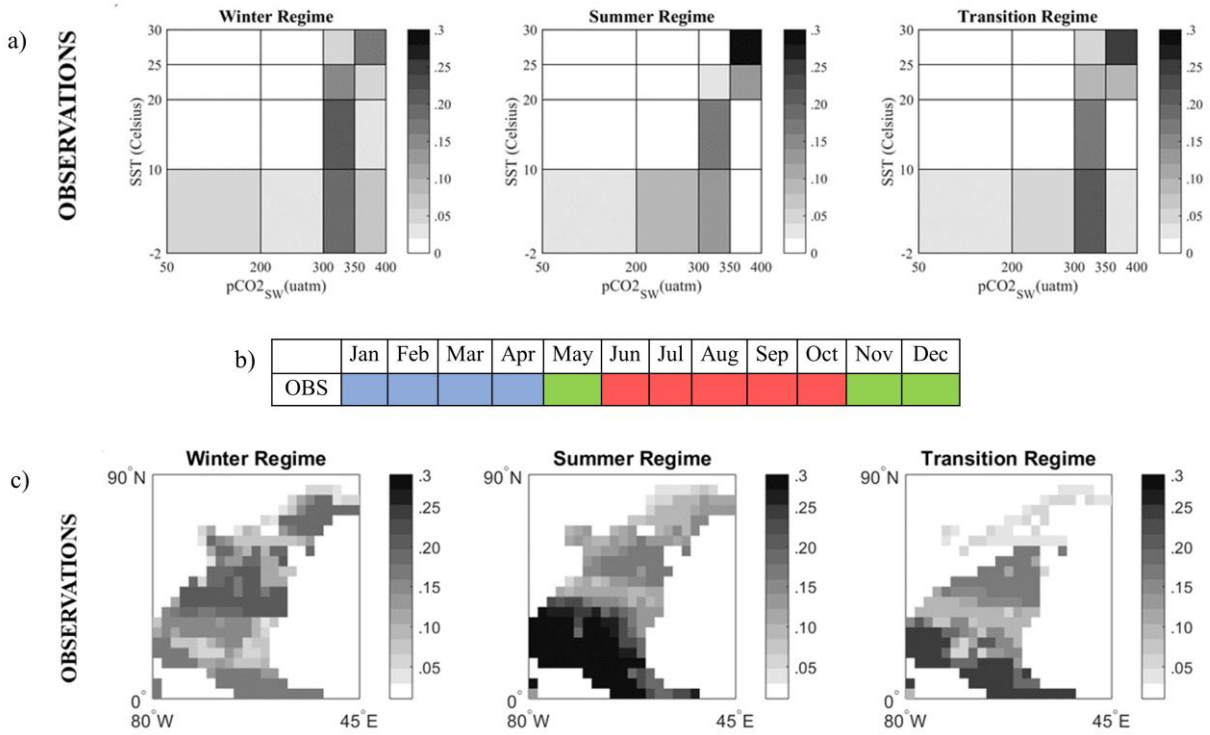

Figure 5: a) Ocean carbon states (regimes) in observation-based data of the North Atlantic. b) Monthly attribution of each ocean carbon regime in the Takahashi observational dataset. Temporal attribution is based on the distance of each monthly 2D histogram to the centroid of each cluster. Each color represents different regime: blue denotes the cold months regime (winter regime), red the warm months (summer regime) and green the May/Nov-Dec transition regime, since this is more a mix of the other two. c) Regional attribution of each regime depicted in 5a. The colors in 5(c) correspond to the ones in 5(a), i.e. to the frequencies of occurrence of each bin (value pair) in the clusters of 5(a).

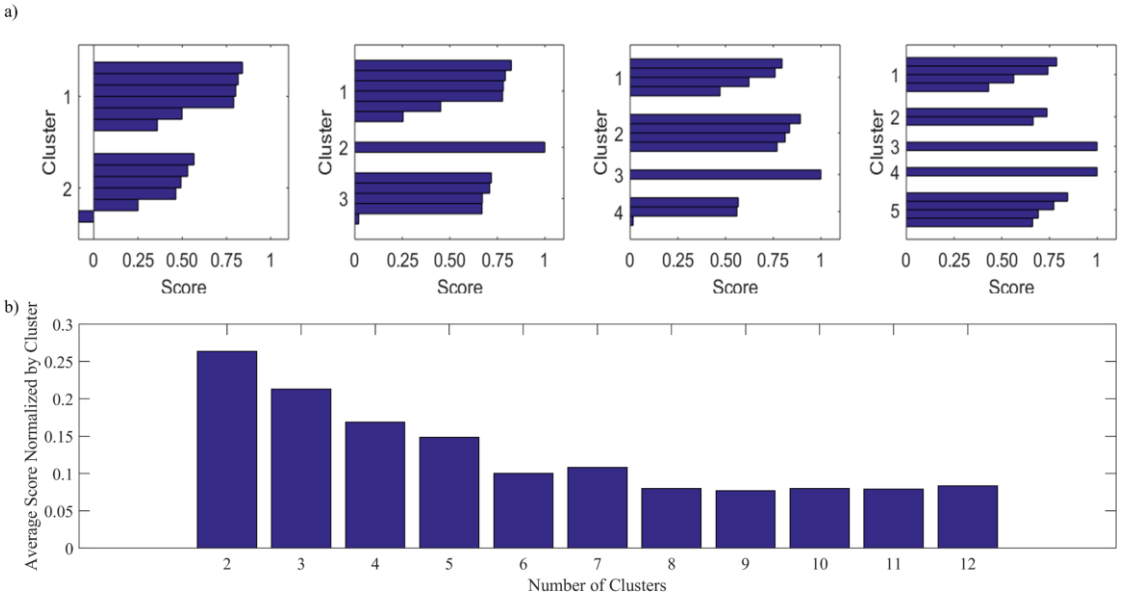

Figure 6: a) Scores for each cluster analysis of the GISS model data in the North Atlantic for k = 2, k = 3, k = 4, k = 5, where cluster k is the predetermined number of clusters and each bar represents the score per 2D histogram. b) Average scores for each cluster analysis with increasing k, normalized by the number of clusters k.

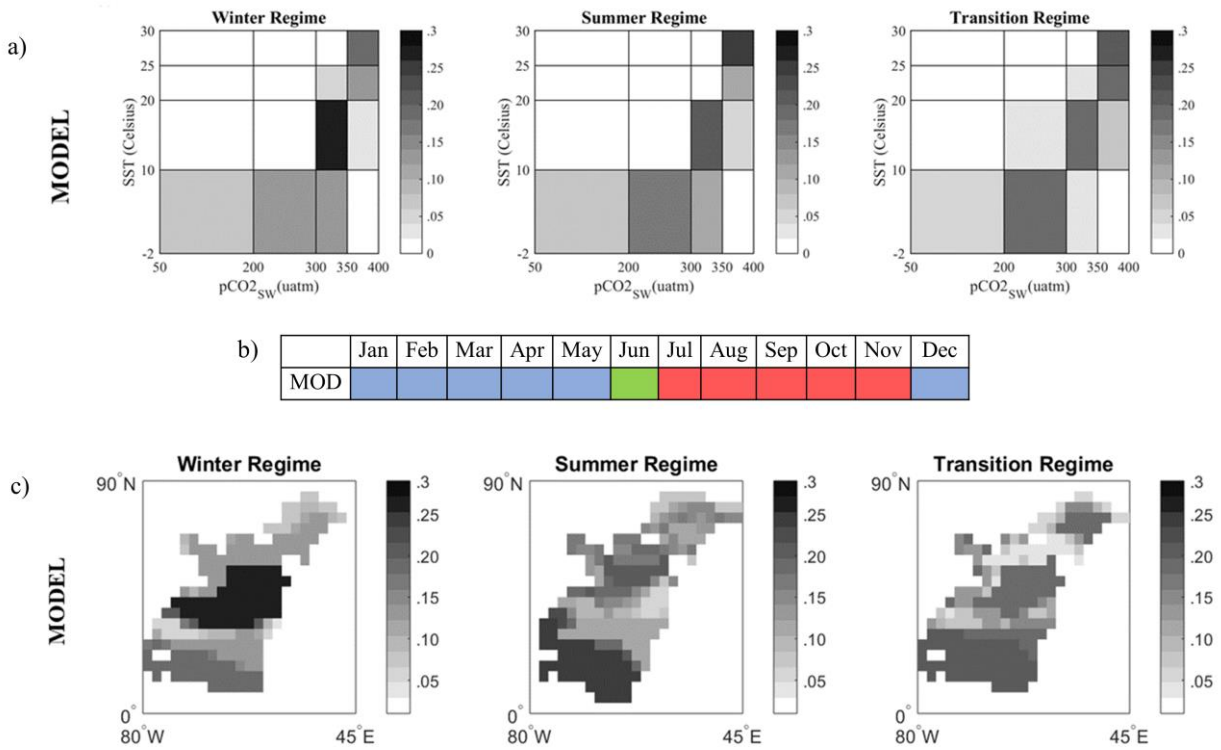

Figure 7: a) Ocean carbon states (regimes) in the North Atlantic from the GISS model output. b) Monthly attribution of each ocean carbon regime. Temporal attribution is based on the distance of each monthly 2D histogram to the centroid of each cluster. Blue denotes the cold months regime (winter regime), Red the warm months (summer regime) and Green the transition regime (only June), since this is more a mix of the other two. c) Regional attribution of each regime depicted in 7(a). The frequencies of occurrence of each bin (value pair) in the clusters of 7(a) is mapped onto the North Atlantic grid.

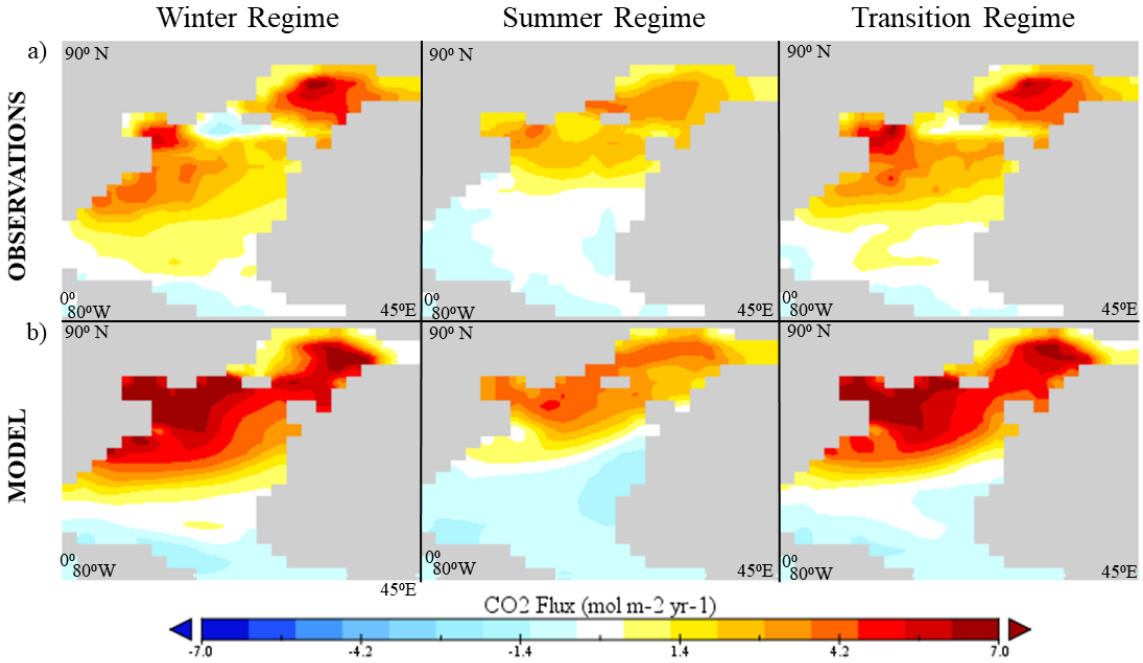

Figure 8: Composites of the CO_2 Flux field over the observed regimes for (a) the observations and (b) the model. The composite fields are computed as averages of the field over the months included in each regime. Both the observations and the model data are composited over the same months as determined by the temporal attribution of the observations dataset, shown in Fig. 5b. Blue shades indicate outgassing, red shades indicate uptake.

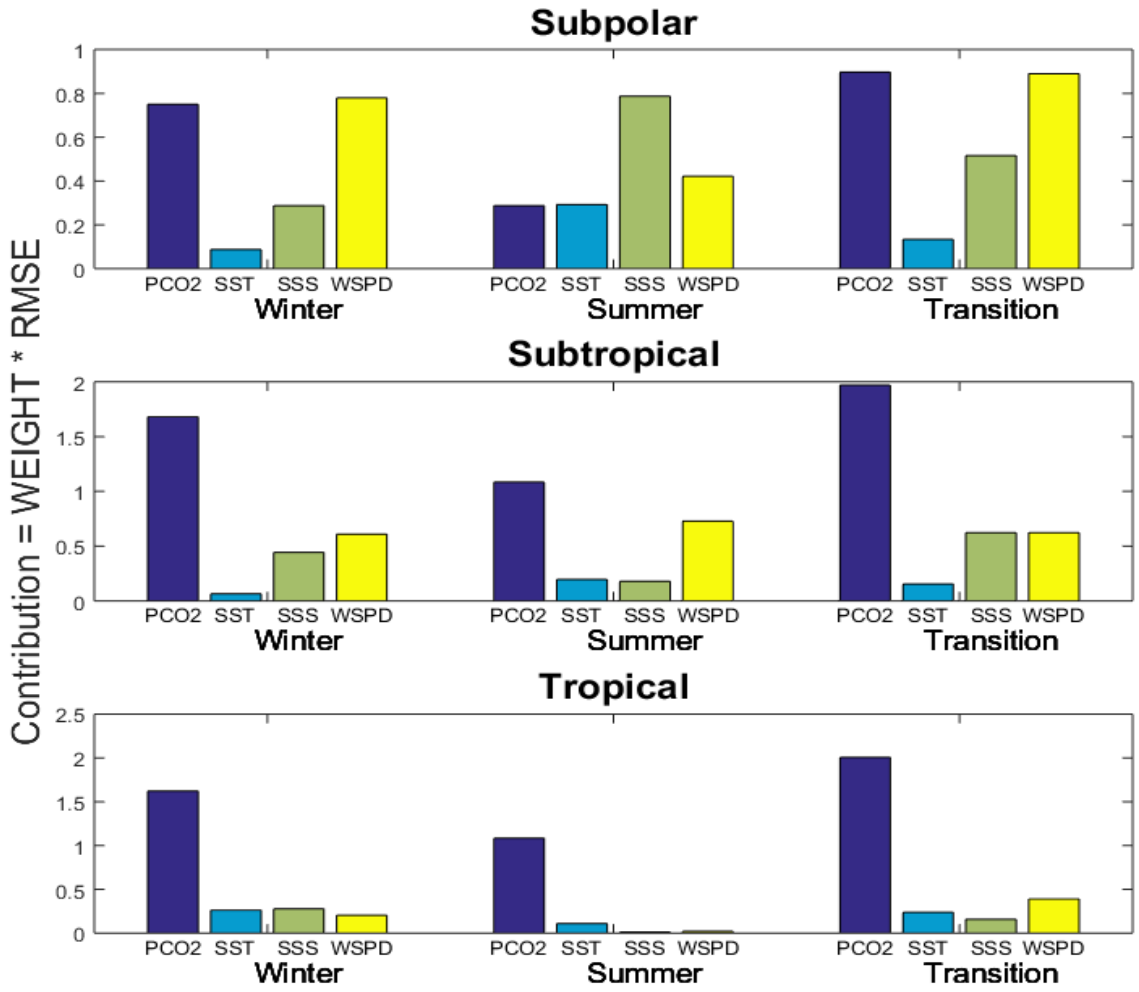

Figure 9: Contributions of each of the variables pCO_2, SST, SSS, WSPD to the overall air-sea flux ΔF bias. The contributions are computed as the products of the weights and the RMSEs of each variable q as described in Eq. (5). See text for detailed explanation.

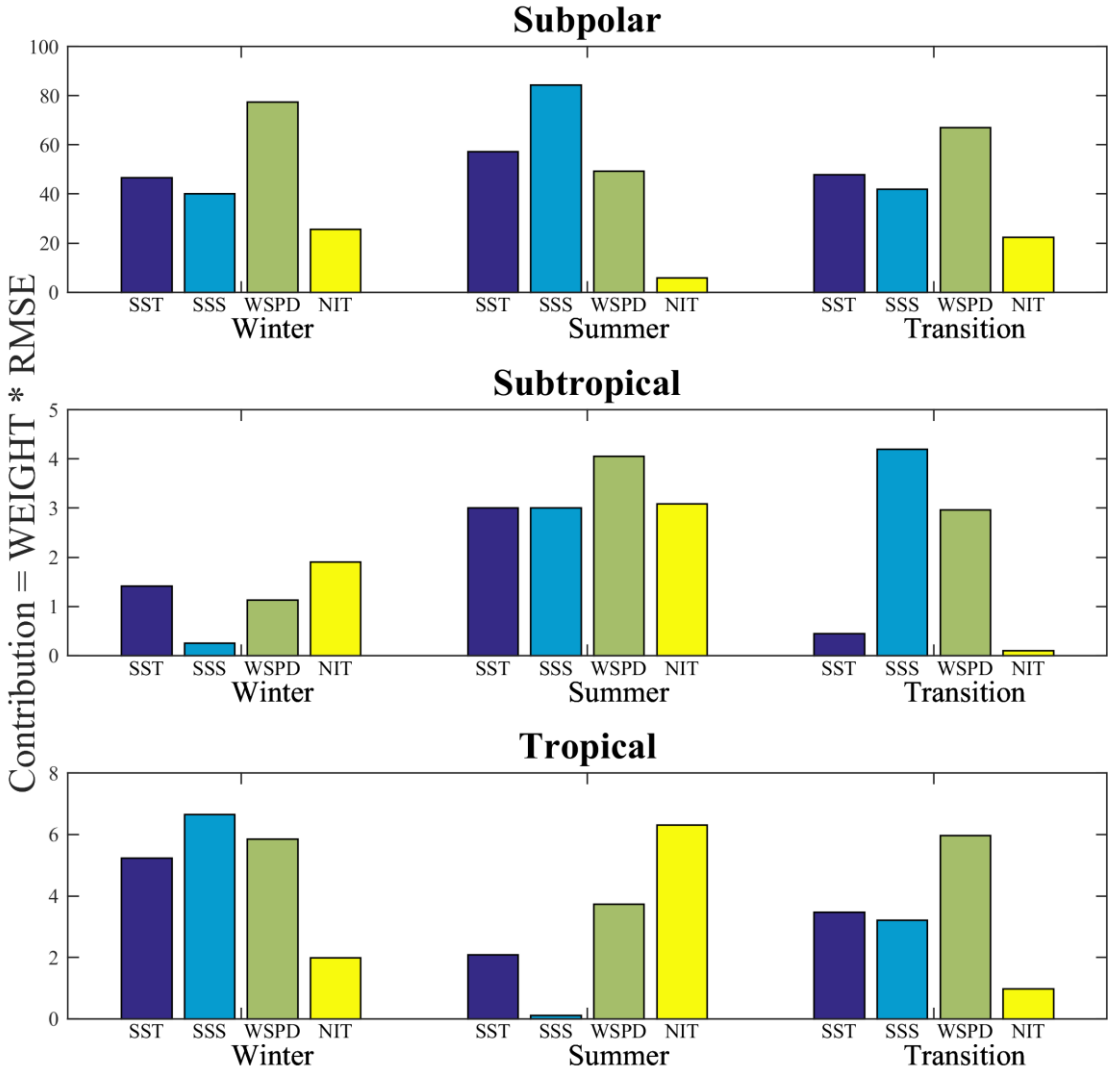

Figure 10: Contributions to the pCO_2sw bias in the model from SST, SSS, wind speed (WSPD) and nitrate (NIT) in the winter, summer, and transitional regimes. Contributions are computed as in Fig. 9 (see details in the text). The entire North Atlantic is differentiated into subpolar, subtropical, and tropical regions to better account for regional differences in the model biases and obtain better linear fit for the computation of the weights in Eq. 6.

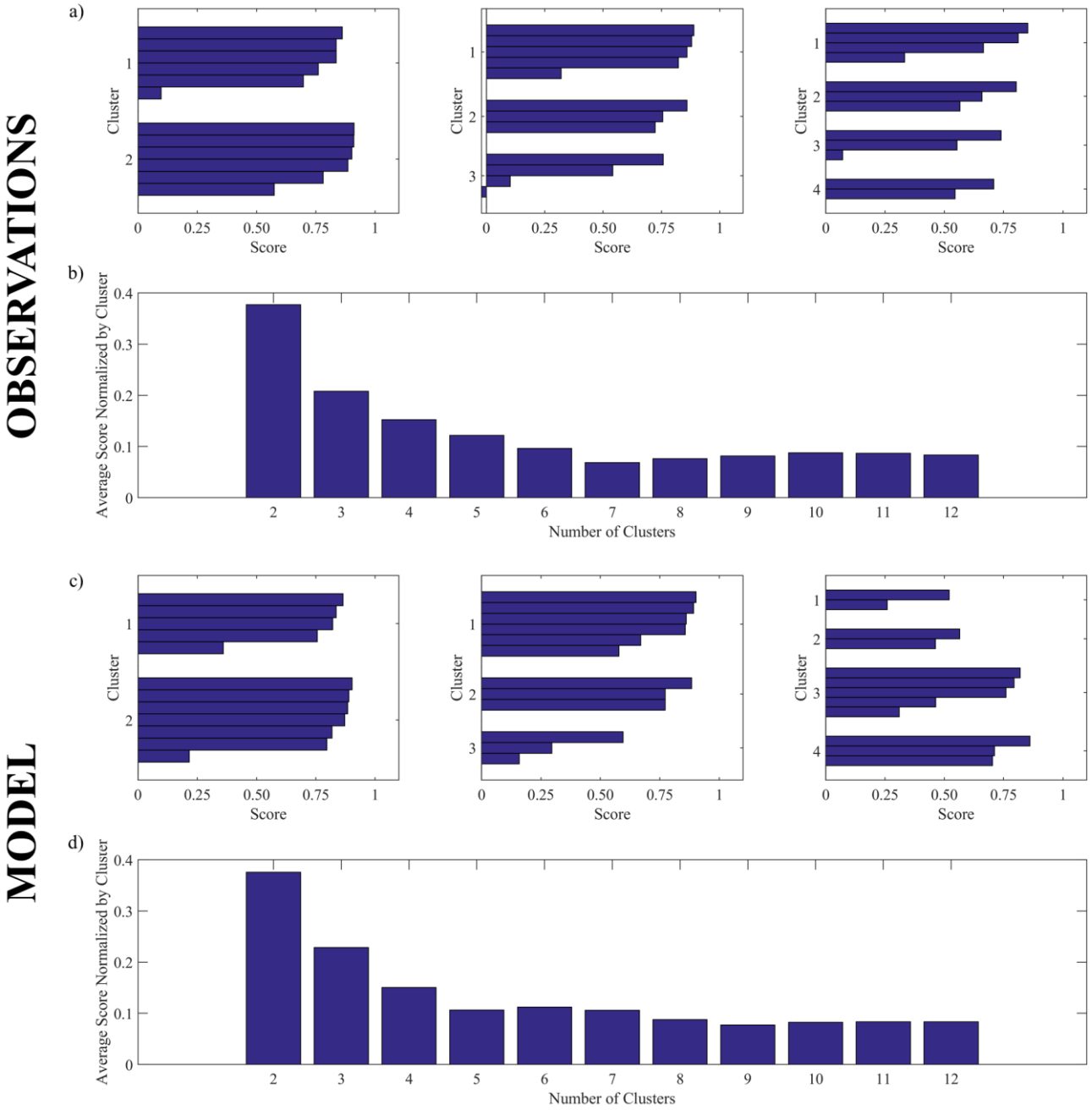

Figure 11: a) Scores for each cluster analysis for k = 2, k = 3, k = 4 for the observational data in the Southern Ocean where cluster k is the predetermined number of clusters and each bar represents the score per 2D histogram b) Average scores of each clustering analysis for increasing k, from k=1 to k=12. c) Scores for each cluster analysis for k = 2, k = 3, k = 4 for the model data in the Southern Ocean. d) Average scores of each clustering analysis for increasing k.

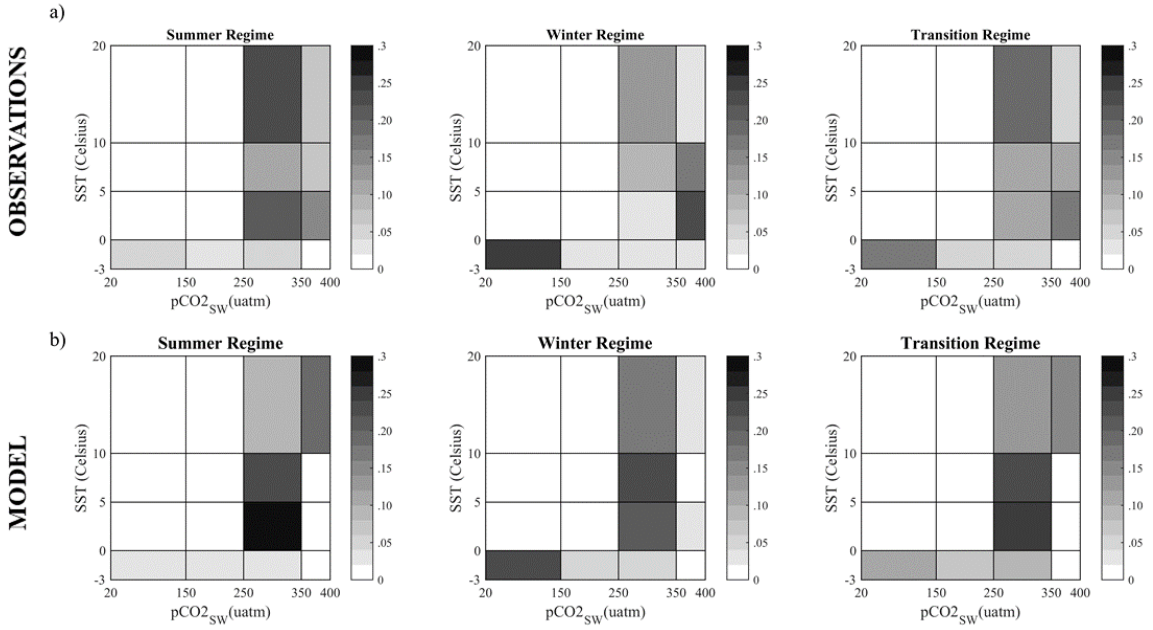

Figure 12: Comparison between the regimes in (a) the observations and (b) the model output.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
OBS												
MOD												

Figure 13: Monthly attribution of each ocean carbon regime in observations and the GISS climate model. Temporal attribution is based on the distance of each monthly 2D histogram to the centroid of each cluster. Referring to austral seasons, blue denotes the cold month regime (winter regime), Red the warm months (summer regime) and green the transition regime, since this is more a mix of the other two.

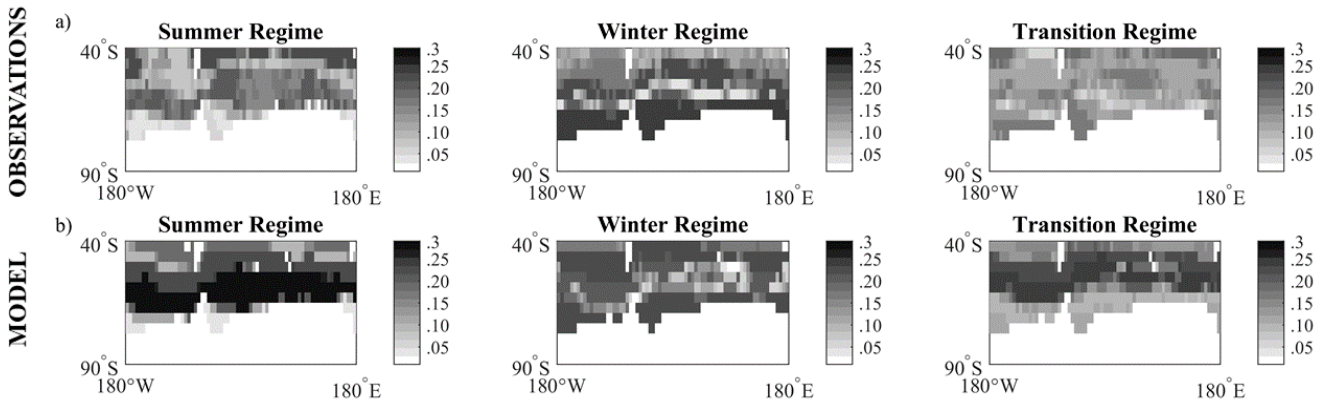

Figure 14: Regional attribution of each regime for k=3 in the Southern Ocean in (a) the observations and (b) the GISS model simulations. Each spatial grid point for every month is associated with its relative frequency of occurrence in the cluster output, and then the months are averaged per regime to output the average frequency of occurrence in each regime. Model regimes are calculated using the monthly groups identified by the observation temporal attribution.

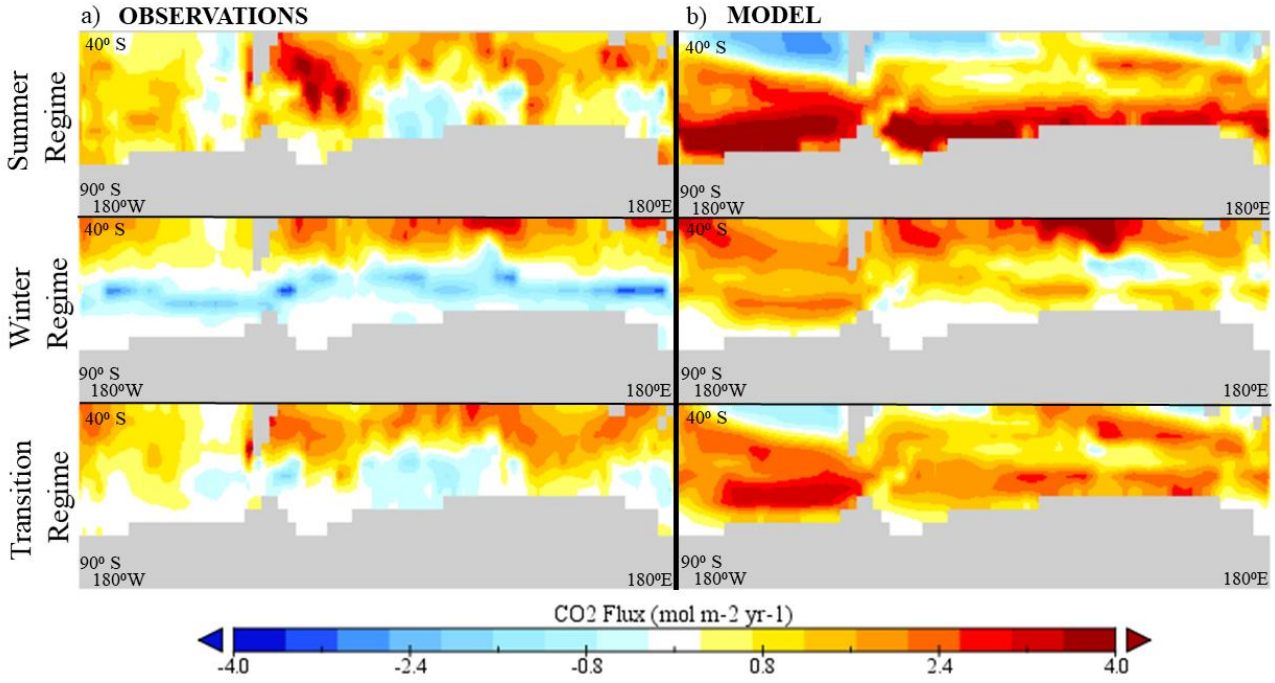

Figure 15: Composites of the CO_2 Flux field over the regimes in the Southern Ocean for (a) the observations and (b) the GISS model. Both the observations and the model data are composited over the same months as determined by the temporal attribution of the observed regimes. Blue shades indicate outgassing, red shades indicate uptake.

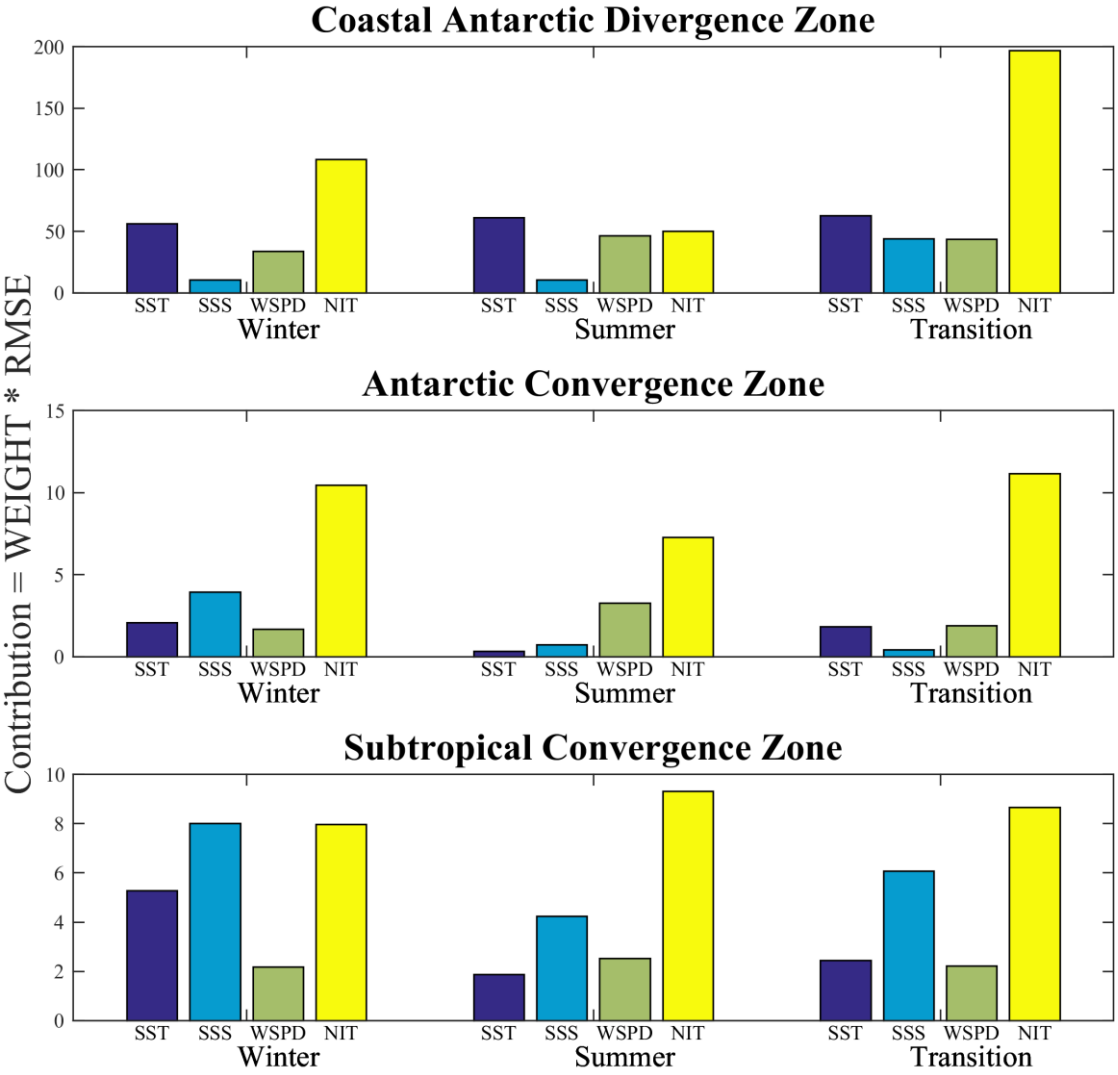

Figure 16: Contributions to the pCO_2sw bias in the model from SST, SSS, wind speed (WSPD) and nitrate (NIT) in the winter, summer, and transitional regimes. The Southern Ocean is differentiated into the coastal Antarctic divergence zone (roughly polewards of 60ºS), Antarctic convergence zone (roughly 60ºS-50ºS), and subtropical convergence zone (roughly 50ºS-40ºS) to better account for regional differences in the model biases and obtain better linear fit for the computation of the weights in Eq. 6 (see Fig. S9).