DEPARTMENT OF ECONOMICS

WORKING PAPER SERIES

2016-07

McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

<http://www.mcmaster.ca/economics/>

NONPARAMETRIC ANALYSIS OF COMPLEX NONLINEAR SYSTEMS

SONALI DAS

JEFFREY S. RACINE

Jeffrey S. Racine (corresponding author), Department of Economics and Graduate Program in Statistics, McMaster University, racinej@mcmaster.ca; Department of Economics and Finance, La Trobe University; Info-Metrics Institute, American University; Rimini Center for Economic Analysis; DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa

Sonali Das, Advanced Mathematical Modelling, CSIR Modelling and Digital Science, P.O. Box 395, Pretoria 0001, South Africa; School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, Private Bag 3, Wits 2050, South Africa; Department of Statistics, PO Box 77000, Nelson Mandela Metropolitan University, Port Elizabeth 6031, South Africa; DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa

ABSTRACT. In this paper we propose a nonparametric methodology designed to facilitate the statistical analysis of complex systems. The proposed approach exploits an ensemble of nonparametric techniques including conditional density function estimation, conditional distribution function estimation, conditional mean estimation (regression) and conditional quantile estimation (quantile regression). By exploiting recent developments in nonparametric methodology and also in open source interactive platforms for data visualization and statistical analysis, we are able to provide an approach that facilitates enhanced understanding of complex empirical phenomenon. We illustrate this approach by exploring the inherent complexity of the Southern Ocean system as a carbon sink, measured in terms of fugacity of carbon dioxide at sea surface temperature ($f\text{CO}_2$), in relation to a number of oceanic state variables, all measured *in situ* during the annual South African National Antarctic Expedition (SANAE) austral summer trips from Cape Town to the Antarctic, and back, between 2010 and 2015.

Date: July 14, 2016.

1991 *Mathematics Subject Classification.* C14 Semiparametric and Nonparametric Methods.

Key words and phrases. Kernel Smoothing; Conditional Density, Distribution, Mean and Quantile Estimation; Exploratory Data Analysis.

The authors would like to thank the following from the CSIR and the Southern Ocean Carbon Climate Observatory: Dr Pedro Monteiro, Dr Sebastiaan Swart, Dr Warren Joubert and Luke Gregor. Authors would also like to acknowledge the co-funding received from the DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) to host Professor Racine at the CSIR for this research. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily attributed to the CoE-MaSS.

1. INTRODUCTION

Understanding empirical relationships present in a set of variables whose behaviour is determined by a complex system is a non-trivial exercise. In order to understand such systems, practitioners sometimes resort to classical parametric approaches such as, for example, linear regression. One drawback with such approaches is that they require the practitioner to make guesses about unknown relationships. Another drawback is that they encourage practitioners to think in terms of scalars (model parameters) that are typically assumed to hold globally (i.e., presumed to be constant over the support of the variables in question). Such methods are often referred to as ‘global approximations’. However, it is known that such approaches can compromise insights about the underlying data generating process (DGP), especially when attempting to model complex non-linear systems. Furthermore, scalar summary measures can mask local features that vary over the support of such systems. In the absence of statistical methods that can assist practitioners who wish to undertake complex counter-factual analyses, important features present in complex systems may remain undetected. In this paper we propose a nonparametric approach that can be used to facilitate the statistical analysis of complex systems. The proposed approach exploits an ensemble of nonparametric techniques including conditional density function estimation, conditional distribution function estimation, conditional mean estimation (regression) and conditional quantile estimation (quantile regression). By exploiting recent developments in nonparametric methodology and in open source interactive platforms for data visualization and statistical analysis, we are able to provide an approach that is singularly well-suited to the enhanced understanding of empirical phenomenon. We are able to investigate the local nature of the DGP by locally smoothing the process using nonparametric kernel methods. One can think of smoothing as a way of regularizing the data which opens up the regularized DGP for local exploration in a way that is not possible using the data alone, all the while avoiding the pitfalls associated with global approaches.

The motivation for this work emanates from the urgent need to understand the behaviour of oceanic carbon dioxide in the Southern Ocean (SO), and consequently the role the SO plays as a carbon sink (Sarmiento & Gruber (2002), Takahashi et al. (2002), Landschützer et al. (2015)). The ability of the SO to be a carbon sink is driven by its ability to absorb atmospheric carbon dioxide (Takahashi et al. (2012)). Phytoplankton biomass in the ocean, under certain oceanic state

conditions of Salinity and Temperature, and the derived Mixed Layer Depth (MLD) measurement, will photosynthesise in the presence of oceanic carbon dioxide, and as such, an increase in photosynthesis goes in tandem with increased usage of oceanic carbon dioxide. As a consequence, the ocean waters are then depleted of carbon dioxide, thereby enabling it to absorb more atmospheric carbon dioxide. Hence, with increasing phytoplankton biomass, measured as Chlorophyll-*a* here, one would expect a lower measure of oceanic carbon dioxide, measured as the fugacity of carbon dioxide at sea surface temperature ($f\text{CO}_2$). However, this expected inverse relationship between Chlorophyll-*a* and $f\text{CO}_2$ is influenced by oceanic state variables, which have different effects on the $f\text{CO}_2$ measurement in different latitudes of the ocean, as well as in different seasons, such as early spring versus early summer (Swart et al. 2015). Furthermore, while $f\text{CO}_2$, Chlorophyll-*a*, Salinity and Temperature are measured at the sea surface, MLD is a vertical profile of the ocean water which measures the depth of uniform physical properties, linking atmospheric conditions (such as storms) with the deep waters (Dong et al. (2008), Banks et al. (2000)). MLD can be treated as a proxy for the volume of the phytoplankton biomass and hence the ability of the ocean to transport the atmospheric carbon dioxide into the deep ocean waters.

In a recent study, Rödenbeck et al. (2015) used an empirical approach to estimate the variability of surface-ocean partial pressure of carbon dioxide ($p\text{CO}_2$) in 17 biomes in the ocean based on data from the Surface Ocean $p\text{CO}_2$ Mapping intercomparison (SOCOM) initiative. Rödenbeck et al. (2015) interpolated data where there were spatial gaps before modelling $p\text{CO}_2$ using linear regression methods as well as neural-networks and self-organizing maps to investigate differences in relationships between $p\text{CO}_2$ and its drivers. The limited empirical understanding of the behaviour of the SO as a carbon sink is largely a consequence of limited data that exists on the SO as compared to, say, the oceans in the northern hemisphere (Le Quéré et al. (2007), Pretorius et al. (2014), Fransson et al. (2009), Böning et al. (2008)). Data analyzed in this paper are taken from the annual South African National Antarctic Expedition (SANAE) to and fro trips between Cape Town and the Antarctic, spanning the five austral summers of 2010 through 2015, measured *in situ*, and relate to the measurements of $f\text{CO}_2$, sea surface temperature (Temperature), salinity (Salinity), chlorophyll concentration which is a measure of productivity (Chlorophyll-*a*), and mixed layer depth (MLD). Note, Temperature, Salinity, Chlorophyll and MLD are considered to be state variables of

the ocean system, while measurements of $f\text{CO}_2$ are considered to be driven by these state variables. We implemented the proposed nonparametric approach in order to obtain enhanced understanding of the role played by the state variables as drivers of $f\text{CO}_2$. Ultimately, by putting this approach in the hands of those with domain-specific wisdom, it is our hope that they will be able to better understand such systems by the rapid investigation of local features and by performing complex exploratory and counter-factual data analysis.

As such, this paper makes two important contributions. First, it simplifies the visualization of data taken from a complex multivariate system using powerful nonparametric methods, with a focus on streamlining counter-factual analysis through manipulation of state variables. Second, we illustrate the proposed method in order to gain insights into the dynamics present in the complex SO system by focusing on the role of the state variables on $f\text{CO}_2$. This is also the first attempt to explore the SANAE 50, 51, 52, 53 and 54 trip data *in situ* data in this fashion.

2. DATA BACKGROUND

The data used in this paper relates to Leg1 and Leg6 journeys of the SANAE trips, with Leg1 referring to the journey from Cape Town to the Antarctic base, and Leg6 referring to the return leg of the journey from the Antarctic base to Cape Town. SANAE trips to the Antarctic occur between early December to late February of consecutive calendar years, which also corresponds to the austral summer when the SO waters are conducive for the ship’s journey with respect to oceanic ice along the ship’s transect. A number of biogeophysical measurements are routinely collected on-board during the SANAE trips which are georeferenced and date-time stamped. For our analysis, besides date, time, latitude and longitude, we focus our attention on 5 variables namely Temperature measured in degrees Celsius; Salinity measured in Practical Salinity Units (PSU); MLD, which is calculated from temperature profiles which are measured with expendable temperature probes, is measured in meters; Chlorophyll-*a* measured in terms of the ability of chlorophyll in the water to achieve florescence, measured in millivolts; and $f\text{CO}_2$ measured in μatm . Further, in the SANAE trips, MLD, which is calculated from temperature profiles which are measured with expendable temperature probes and calculated as a function of temperature change in the water column profiles, is measured only at certain time points, while the rest of

the variables considered here are continuously measured underway at a relatively higher spatio-temporal resolution than that for the MLD. As mentioned earlier, data investigated in this study are from 5 consecutive SANAE trips during the austral summer of 2010-11, 2011-12, 2012-13, 2013-14 and 2014-15. For an initial exploratory data analysis exercise, the relevant data obtained from the Southern Ocean Carbon Climate Observatory (SOCCO) team for the two legs of the five trips were plotted for each of the variables individually against latitude to look for any obvious suspect measurements, and thereafter based on a number of iterative discussions with the SOCCO team, such suspect measurements were replaced with ‘NA’. Details of the data cleaning exercise for each of the original files for the 5 trips are available on request. Further, MLD values corresponding to the latitude where the other variables were measured, were interpolated linearly based on the recommendation of the SOCCO team. It is important to mention at this point that due to various reasons, such as instrument failure, there were gaps in the data series for the 5 trips of the study period, including one trip where data from a whole leg was missing, such as Leg6 of SANAE 50 (2010-11).

It is also important to remind the reader that the available *in situ* data is only sampled along the ship’s transect path as it travels between Cape Town and the Antarctic and back, as such, the data is limited with regards to its spatial representation. Furthermore, there are spatial locations along the transect that exhibit large gaps due to equipment failure. To overcome these limitations, we can ‘borrow’ information using various statistical techniques that are akin to Bayesian shrinkage approaches. These issues notwithstanding, the value added from exploring local structure and conducting counter-factual experiments is potentially quite high.

3. METHODOLOGY

The SANAE data discussed above relates to a complex dynamic system where the outcome of interest, $f\text{CO}_2$, is driven by Salinity, Temperature, Mixed Layer Depth, and Chlorophyll-*a*. Let us denote by Y $f\text{CO}_2$, and by X Salinity, Temperature, Mixed Layer Depth, and Chlorophyll-*a*. Denote by Z the joint set of variables (Y, X) . We adopt the convention that uppercase Z denotes the random vector, lowercase z denotes particular realizations of Z .

The ‘probability density function’ (PDF, typically denoted by lowercase $f(\cdot)$) and ‘cumulative distribution function’ (CDF, typically denoted by uppercase $F(\cdot)$) are fundamental objects in statistics. They embody all probabilistic information contained in a set of random variables Z . They can be unconditional (e.g., $f(z)$) or conditional (e.g., $f(y|z)$). Other statistical objects are built upon the PDF and CDF. For instance, the unconditional mean of the random variable Y is defined as $E(Y) = \int_{-\infty}^{\infty} yf(y) dy$, while the conditional mean of Y given X is defined as $E(Y|X = x) = \int_{-\infty}^{\infty} yf(y|x) dy$. Note that the conditional mean is also called a ‘regression function’. Furthermore, the unconditional quantile function is the inverse CDF function, i.e., $q_{\tau} = F^{-1}(\tau)$, $\tau \in [0, 1]$, while the conditional quantile function is $q_{\tau}(x) = F^{-1}(\tau|x)$.

One can appreciate that statistical objects are related to the underlying DGP which, for better or for worse, is typically not known to the researcher. One can make parametric guesses about the DGP (i.e., write down a ‘parametric’ model), or one can use more flexible ‘nonparametric’ approaches that place minimal structure on the problem. The strength of parametric approaches flows from their ease of interpretation and, often, computational simplicity. The drawback of parametric approaches, however, is that if the presumed model does not faithfully mimic the DGP, then any inference based on the presumed model will be unsound. Nonparametric approaches, on the other hand, are far more flexible than parametric approaches and possess certain desired statistical properties (such as ‘consistency’), though they tend to be more difficult to interpret than their parametric counterparts.

One powerful application of nonparametric methods is for exploratory data analysis. That is, placing minimal structure on the data (existence and smoothness of $f(\cdot)$ and $F(\cdot)$), we can explore potentially complex relationships present in a set of data. Given that we are dealing with data taken from a complex and dynamic system, nonparametric exploratory data analysis could be beneficial to the researcher. In particular, it allows one to control for certain conditioning variables while manipulating other conditioning variables and thereby assess the impact on the conditional PDF, CDF, or conditional mean. Our approach is designed to simplify interpretation and to streamline counter-factual analysis with minimal input required from the user.

In what follows we will consider local nonparametric kernel estimates of the PDF, CDF, conditional mean and conditional quantile. These estimates are intuitively appealing as they place

minimal structure on the DGP yet can reveal complex relationships present in a set of data. Essentially, the nonparametric estimates of the conditional PDF and CDF, which we denote by $\hat{f}(\cdot)$ and $\hat{F}(\cdot)$ are constructed by taking local averages of kernel (i.e., weight) functions that themselves are suitable for PDF and CDF estimation. These kernel estimates depend on ‘bandwidths’ which define the neighbourhood in which the averaging takes place, and there is a separate bandwidth that can be set for each variable in Z . See Hall et al. (2004), Li & Racine (2008), and Li et al. (2013) for details.

Why consider both the PDF and CDF rather than simply one or the other? The reason is that the PDF is useful for conveying features such as shifts in the conditional mean and the presence of multiple modes, while the CDF is useful for determining whether the ‘probability mass’ has shifted in a particular direction, whether such shifts ‘stochastically dominate’ one another, and assessing shifts in conditional quantiles.

Below we present an interactive graphical procedure that we hope is of value to the practitioner. In particular, we consider $\hat{f}(y|z)$ and $\hat{F}(y|z)$ (Y is a scalar-valued random variable, X is vector-valued) and plot $\hat{f}(y|z)$ ($\hat{F}(y|z)$) versus y for a range of y (controllable by the user if one wishes to ‘zoom in’ on various features) for a given set of values for z (e.g. their median values). We then allow the practitioner to vary the particular values for z to assess their impact on the conditional PDF and CDF, while also computing the conditional mean. That is, we allow the practitioner to conduct counter-factual analysis and answer questions such as “what would happen to the density/distribution/mean/quantile of $f\text{CO}_2$ if MLD increased holding other state variables unchanged?” As will be seen, this approach allows the practitioner to explore the complex relationships present in a set of data with minimal structure placed on the DGP. We implement this approach using the ‘manipulate’ package in RStudio, RStudio being a front-end for R (R Core Team 2016), a language for statistical computing and graphics. The nonparametric estimates are constructed using routines in the R package `np` (Hayfield & Racine 2008).

Figure 1 presents the interactive graphical implementation and the resulting estimated conditional PDF, CDF, mean and median for Leg 1 of the trip SANAE50 holding the driving variables at their respective median values. The menu allows the user to zoom in on particular features by modifying the limits of the horizontal axis using data quantiles. It also allows you to set which

conditional quantile you wish to assess (defaults to the conditional median). Most importantly, one can manipulate any driving variable and thereby perform counter-factual analysis while controlling for the levels of other driving variables. Finally, one can assess the impact of varying the amount of local averaging by modifying the ‘scale factor’ for each variable (defaults to 0.25).

FIGURE 1. Screen shot of RStudio using the manipulate package to create the pull down menu allowing the user to modify the estimated conditional PDF and CDFs and conduct counter-factual analysis. The top figure presents the conditional PDF $\hat{f}(y|z)$ along with the conditional mean (vertical dashed line), the bottom the conditional CDF $\hat{F}(y|z)$ along with the conditional quantile (vertical dashed line). The value of the state variables are presented in the legend for each figure along with the value of the conditional mean and quantile. The number of observations are also provided in the subtext of each figure. Counter-factual analysis is achieved simply by clicking on any of the sliders for the state variables and thereby altering their value.

3.1. An Illustrative Example: Chlorophyll-*a* Counter-Factual Analysis. Suppose the researcher was interested in isolating the impact of the state variable Chlorophyll-*a* on $f\text{CO}_2$ while controlling for the levels of the other state variables. Figure 2 illustrates the impact of a change in Chlorophyll-*a* concentration from its upper quartile (0.19) to its lower quartile (-0.06) holding

the other drivers constant at their respective medians (median Salinity 34.21 PSU, median Temperature 7.10 degrees Celsius, median MLD 43.25 meters). There is a clear rightwards shift in the conditional distribution of $f\text{CO}_2$ associated with the reduction in Chlorophyll- a - a 0.25 millivolt reduction in Chlorophyll- a concentration increases average $f\text{CO}_2$ by $10.7 \mu\text{atm}$ (i.e., from 358.2 to 369.9). Let us denote the relative change in $f\text{CO}_2$ arising from this change in Chlorophyll- a concentration as $\Delta = (369.9 - 358.2)/(-0.06 - 0.19) = -46.8$.¹

FIGURE 2. Leg 1, SANA50. In the figures on the left, chlorophyll- a concentration is set to 0.19 (its 75th percentile i.e. upper quartile), and mean $f\text{CO}_2$ is 358.2. In the figures on the right, chlorophyll- a concentration is set to -0.06 (its 25th percentile i.e. lower quartile), and mean $f\text{CO}_2$ is 369.9.

Table 1 presents coefficient (i.e., derivative) estimates and average derivative estimates from a range of parametric, semiparametric, and nonparametric regression models, including linear least squares regression (lm), linear least trimmed squares regression (lts), local constant nonparametric

¹Note that negative values of Chlorophyll- a concentration reflect calibration of the measurement device, and since we shall focus on relative changes the absolute scale to which the measurement device is calibrated is irrelevant, though comparison across legs based on different calibrations could be problematic.

regression (lc), local linear nonparametric regression (ll), and a varying coefficient specification where the model’s coefficient vary with respect to longitude and latitude (scoef). The rightmost column in Table 1 summarizes the impact of a one unit increase in chlorophyll-*a* on $f\text{CO}_2$. Table 1 reveals that these (scalar) estimates range from -4.45 to -60.4. The value obtained from our exploratory method computed by comparing mean $f\text{CO}_2$ when Chlorophyll-*a* concentration is changed from its upper quartile (0.19) to its lower quartile (-0.06), $\Delta = -46.8$, is of a similar order of magnitude.

	Salinity	Temp.	MLD	Chl- <i>a</i>
lm	-23.25	-0.60	-0.15	-60.04
lts	-33.09	0.57	-0.04	-52.17
lc	2.97	2.82	-0.08	-4.45
ll	86.83	1.11	0.18	-41.31
scoef	11.30	9.69	-0.69	-43.61

TABLE 1. Leg 1, SANA50

However, by reducing this change in Chlorophyll-*a* on mean $f\text{CO}_2$ to a scalar in this manner (i.e., $\Delta = -46.8$), we may be masking features present in the data. Suppose the practitioner considers smaller changes instead and assesses the impact on mean $f\text{CO}_2$ keeping the other drivers constant at their respective medians. Table 2 and Figure 3 summarize this exercise, and there are nonlinearities present that are clearly masked by reducing the effect to a scalar. Furthermore, the impact of increasing Chlorophyll-*a* beyond its 55th percentile (0.09) on mean $f\text{CO}_2$ is negligible, at least when the other drivers are held constant at their respective medians. All of the action occurs for Chlorophyll-*a* levels between 0.00 and 0.09 when the other drivers are held constant at their respective medians it would appear.

Of course, there is much that can be learned about the underlying mechanism. One might consider this Chlorophyll-*a* counter-factual analysis when the other state variables are held constant at different levels than those used above. By modifying the other drivers one can then assess the impact on the response of mean $f\text{CO}_2$ to changes in Chlorophyll-*a* as a function of the levels of the remaining state variables while placing minimal structure on the underlying DGP. And of course one could repeat this exercise using one of the other state variables instead of Chlorophyll-*a*. A researcher with specific questions could thereby explore the structure of this complex nonlinear system using powerful nonparametric methods while focusing on the science at hand.

TABLE 2. Response of mean $f\text{CO}_2$ to changes in Chlorophyll- a at a range of quantiles of Chlorophyll- a holding other drivers constant at their respective medians (Leg 1, SANAE50, median Salinity 34.21 PSU, median Temperature 7.10 degrees Celsius, median MLD 43.25 meters).

τ	Chl- a	$f\text{CO}_2$	Δ
0.20	-0.07	370.0	NA
0.25	-0.06	369.9	-10.0
0.30	-0.04	369.8	-5.0
0.35	-0.02	369.5	-15.0
0.40	0.00	365.3	-210.0
0.45	0.03	360.5	-160.0
0.50	0.07	358.4	-52.5
0.55	0.09	358.1	-15.0
0.60	0.11	358.1	0.0
0.65	0.13	358.1	0.0
0.70	0.17	358.0	-2.5
0.75	0.19	358.0	0.0
0.80	0.21	358.0	0.0

FIGURE 3. Plot of the response of mean $f\text{CO}_2$ to changes in Chlorophyll- a at a range of quantiles of Chlorophyll- a holding other drivers constant at their respective medians (Leg 1, SANAE50).

4. CONCLUSION

Our intent in this paper is to develop an interactive visualisation method for data arising from a complex system while placing minimal prior assumptions on the underlying data generating process. Our intent is to also develop a tool which can be used to conduct complex counterfactual experiments for such systems. To this end, this paper presents an interactive ensemble of nonparametric techniques that enable a practitioner to visualize and explore local features of their data while side-stepping the limitations associated with global estimation methods that often reduce relationships to scalars thereby masking complex nonlinear functional relationships. We highlight the proposed method's potential by considering the behaviour of the Southern Ocean as a carbon sink based on *in situ* data collected from five consecutive SANAE trips between Cape Town and the Antarctic. Although this approach was inspired by the nonlinear complexity present in this unique data set, it has widespread potential application beyond the SANAE data. Code and data underlying Figure 1 and the subsequent analysis can be obtained from <http://socserv.mcmaster.ca/racinej/Gallery/Fun.html> (see item 13).

REFERENCES

- Banks, H. T., Wood, R. A., Gregory, J. M., Johns, T. C. & Jones, G. S. (2000), ‘Are observed decadal changes in intermediate water masses a signature of anthropogenic climate change?’, *Geophysical Research Letters* **27**(18), 2961–2964.
URL: <http://dx.doi.org/10.1029/2000GL011601>
- Böning, C., Dispert, A., Visbeck, M., Rintoul, S. R. & Schwarzkopf, F. U. (2008), ‘The response of the antarctic circumpolar current to recent climate change’, *Nature Geoscience* **1**, 864–869.
- Dong, S., Sprintall, J., Gille, S. T. & Talley, L. (2008), ‘Southern ocean mixed-layer depth from argo float profiles’, *Journal of Geophysical Research: Oceans* **113**(C6), n/a–n/a. C06013.
URL: <http://dx.doi.org/10.1029/2006JC004051>
- Fransson, A., Chierici, M. & Nojiri, Y. (2009), ‘New insights into the spatial variability of the surface water carbon dioxide in varying sea ice conditions in the arctic ocean’, *Continental Shelf Research* **29**(10), 1317 – 1328.
URL: <http://www.sciencedirect.com/science/article/pii/S0278434309000983>
- Hall, P., Racine, J. S. & Li, Q. (2004), ‘Cross-validation and the estimation of conditional probability densities’, *Journal of the American Statistical Association* **99**(468), 1015–1026.
- Hayfield, T. & Racine, J. S. (2008), ‘Nonparametric econometrics: The np package’, *Journal of Statistical Software* **27**(5).
URL: <http://www.jstatsoft.org/v27/i05/>
- Landschützer, P., Gruber, N., Haumann, F. A., Rödenbeck, C., Bakker, D. C. E., van Heuven, S., Hoppema, M., Metzl, N., Sweeney, C., Takahashi, T., Tilbrook, B. & Wanninkhof, R. (2015), ‘The reinvigoration of the southern ocean carbon sink’, *Science* **349**(6253), 1221–1224.
URL: <http://science.sciencemag.org/content/349/6253/1221>
- Le Quéré, C., Rödenbeck, C., Buitenhuis, E. T., Conway, T. J., Langenfelds, R., Gomez, A., Labuschagne, C., Ramonet, M., Nakazawa, T., Metzl, N., Gillett, N. & Heimann, M. (2007), ‘Saturation of the southern ocean co2 sink due to recent climate change’, *Science* .
URL: <http://science.sciencemag.org/content/early/2007/05/17/science.1136188>
- Li, Q., Lin, J. & Racine, J. S. (2013), ‘Optimal bandwidth selection for nonparametric conditional distribution and quantile functions’, *Journal of Business and Economic Statistics* **31**(1), 57–65.
- Li, Q. & Racine, J. S. (2008), ‘Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data’, *Journal of Business and Economic Statistics* **26**(4), 423–434.
- Pretorius, W. B., Das, S. & Monteiro, P. (2014), ‘Investigating the complex relationship between in situ southern ocean pco2 and its ocean physics and biogeochemical drivers using a nonparametric regression approach’, *Environmental and Ecological Statistics* **21**(4), 697–714.
URL: <http://dx.doi.org/10.1007/s10651-014-0276-5>
- R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: <https://www.R-project.org/>
- Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., Landschützer, P., Metzl, N., Nakaoka, S., Olsen, A., Park, G.-H., Peylin, P., Rodgers, K. B., Sasse, T. P., Schuster, U., Shutler, J. D., Valsala, V., Wanninkhof, R. & Zeng, J. (2015), ‘Data-based estimates of the ocean carbon sink variability— first results of the surface ocean pco2 mapping intercomparison (socom)’, *Biogeosciences* **12**(23), 7251–7278.
URL: <http://www.biogeosciences.net/12/7251/2015/>
- Sarmiento, J. L. & Gruber, N. (2002), ‘Sinks for anthropogenic carbon’, *Physics Today* **55**(8), 30–36.
URL: <http://scitation.aip.org/content/aip/magazine/physicstoday/article/55/8/10.1063/1.1510279>
- Swart, S., Thomalla, S. & Monteiro, P. (2015), ‘The seasonal cycle of mixed layer dynamics and phytoplankton biomass in the sub-antarctic zone: A high-resolution glider experiment’, *Journal of Marine Systems* **147**, 103 – 115. The variability of primary production in the ocean: From the synoptic to the global scale. The 45th International Liege Colloquium on Ocean Dynamics, Liege, Belgium, May 13-17, 2013.
URL: <http://www.sciencedirect.com/science/article/pii/S0924796314001535>
- Takahashi, T., Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R., Feely, R. A., Sabine, C., Olafsson, J. & Nojiri, Y. (2002), ‘Global seaair {CO2} flux based on climatological surface ocean pco2, and seasonal biological and temperature effects’, *Deep Sea Research Part II: Topical Studies in Oceanography* **49**(910), 1601 – 1622. The Southern Ocean I: Climatic Changes in the Cycle of Carbon in the Southern Ocean.
URL: <http://www.sciencedirect.com/science/article/pii/S0967064502000036>

Takahashi, T., Sweeney, C., Hales, B., Chipman, D. W., Newberger, T., Goddard, J. G., Iannuzzi, R. A. & Sutherland, S. C. (2012), 'The changing carbon cycle in the southern ocean'.
URL: <http://ir.library.oregonstate.edu/xmlui/handle/1957/34836>

APPENDIX A. LEG 1 SUMMARY FOR ALL TRIPS (SANA50-54)

Chlorophyll-a Leg 1

fCO₂ Leg 1

FIGURE 4. CO₂

Salinity Leg 1

FIGURE 5. Salinity

Temperature Leg 1

FIGURE 6. Temperature

MLD Leg 1

FIGURE 7. MLD