Author response bg-2013-533

We thank the referee for the careful review and the constructive comments. Please find the answers to all specific comments below. Author comments are given in turquoise, extracts from the paper in blue.

Both reviewers criticized that the definition of Time of Emergence (ToE) in the submitted manuscript is flawed. Unfortunately, there is an error in the definition given in the submitted manuscript. This has now been corrected. We emphasize that all calculations were done using the correct definition and results presented in this revised version remain unchanged compared to the first version.

Review #1

The study investigates the minimum length of the time period where trends in physical or biogeochemical ocean variables become detectable against the natural variability background. It uses a "Time of Emergence" concept applied to modelled fields of SST, DIC, pCO2, and pH.

This is a very relevant topic in the context of anthropogenic climate change. Nevertheless, as detectability is a statistical question, its investigation needs a clear statistical context, including tests of statistical significance of the trend with respect to a defined significance level. Such context or tests are completely absent from the manuscript. Rather, the "ToE" concept is introduced vaguely (and wrongly, see below). The threshold "S/N>2" is given without any motivating rationale, such that is remains completely open what a certain ToE value would tell in practice. I agree that the relative statements (such as "ToE of SST longer than that of DIC") are less affected by the exact definition, but still think that a clean statistical treatment is a necessity for meaningfull results.

We realize and acknowledge that the ToE method is not sufficiently explained and that the ToE formula has to be rearranged to provide the correct unit, i.e. time.
Concerning the lack of statistical tests: by contrasting measures for signal and noise, the ToE method itself is related to certain statistical tests. Yet, it is not a standard procedure with given thresholds, significance levels et cetera. Still, it is possible to test whether the trend of the variables (like DIC) is significant or not. The linear trend representing S is based on 30 years of annual data, we checked the trend signal of these 30-year time-series given the null hypothesis that there is no trend using the approach suggested by von Storch and Zwiers, 2001. Further, we checked using an F-test that sdv remains approximately constant over the simulation period. The according part of the method section was rewritten and extended:

“ToE is defined as ToE = (2xN)/S where S is the trend and N a measure for variability.

For each grid cell, S is defined as the linear trend (per year) over the period 1970–1999. A different approach is the computation of the trend by applying a smoothing spline on the time series; test calculations for 1970–1999 yield comparable results. Figure 1 exemplarily illustrates the two approaches and the good agreement between them based on the model NCAR CESM1. We note that by applying the linear trend from the 1970 to 1999 period also in the future, any changes in trends are not explicitly accounted for. Changes in trend are likely to remain relatively small in the next few decades, but trends will differ considerably between business-as-usual and stringent mitigation scenarios by the end of this century (e.g., Steinacher et al., 2009; Cocco et al., 2013; Bopp et al., 2013). For N, the standard deviation (sdv) over the entire simulation, 1870–1999, is used. Prior to this last step, the data is detrended via a spline approach (cut-off period: 40 yr; Enting, 1987).

For illustration purposes, we calculate ToE for DIC at a location in the subtropical North Pacific (see also Fig. 1). By inserting the respective values for S (0.94 mmol m-3/yr) and N (7.24 mmol m-3), we obtain (2x7.24)/0.94 = 15.4 yr, i.e., a (rounded up) ToE of 16 yrs. The ensemble mean of ToE is computed from the ToE of individual models, and not from the ensemble mean of S and N. Note that the presented ensemble mean patterns, i.e., the averages of all 17 models, are not necessarily physically consistent.

ToE is a measure for the point in time when the trend signal (S x ToE) exceeds two times the background variability N, i.e., the approximate 95% confidence interval of the background variability. The choice of the detection threshold differs between studies, other approaches are e.g. one sdv of seasonal or annual means (Hawkins and Sutton, 2012), observation-based thresholds (Ilyina et al., 2009; Ilyina and Zeebe, 2012) or the range of the pre-industrial annual cycle (Friedrich et al., 2012). Here, we use the rather conservative value of two sdv of interannual variability. For a threshold of one sdv ToE would be half, accordingly.

By calculating S over a time period of 30 years, we can to a certain degree rule out interference of low-frequency variability in the detection of the trend (see e.g., McKinley et al., 2011). A ToE of only a few decades, as we find it especially for the three carbon cycle variables (see Sec. 3.1), is thus a strong indicator for the significance of the respective trend. This is confirmed by a significance test (t test, 5% level) of the trend of the underlying 30 year time series (not shown): For all 17 models, all trends in pH are significant. The trends in pCO2 are also significant, yet with localized insignificant exceptions in the Southern Ocean (BCCR BCMC, IPSL-CM5A-MR) and the upwelling region off Peru and Chile (CanESM2). Trends in DIC are significant in large parts of the global oceans, exceptions are the high latitudes and the equatorial Pacific. Statistically significant trends in SST are less widespread and corresponding regional results are highly model-dependent.

In using these definitions, we assume that (i) the trend from 1970–1999 is linear, (ii) the sdv is constant over time and, by using annual averages, (iii) that trends and sdv patterns are comparable for annual, seasonal or monthly data. To verify (i), the global trends of surface DIC and SST for the period 1970–1999 are investigated. For all models, we find that trends in global surface DIC can be represented by a linear function. SST shows larger inter-annual variability, yet likewise with a linear underlying trend. For (ii), we investigate the detrended data (1870–1999) of DIC and SST of all 17 models (not shown). The comparison of sdv fields calculated for the first and second 65 yrs (F test, 5% level) illustrates that differences only occur very localized, consequently we suggest that this assumption is confirmed. Assumption (iii) can be confirmed for the trend patterns. The standard deviations, however, differ considerably in magnitude – we address this issue in Sect. 3.2."

I also miss any process-related explanation, or implications for biogeochemistry, from the numerical findings. Why are the different variables behaving so differently? What do we learn from the numbers? Is there any particular advice for observations (other than just collecting more data)? Why are the results different from previous studies (e.g. the cited ones by Mc Kinley et al, Fay et al)?

i.) A comprehensive investigation of underlying processes is beyond the scope of this paper. Still, we added a statement to the results section concerning the difference in ToE between DIC and pCO2:
"ToE of pCO2 and pH show a very similar pattern. However, the trends emerge much faster for pCO2 and pH than for DIC – after approx. 12 yr for the majority of the global ocean area, 14–18 yr in the Arctic Ocean and approx. 20 yr in the equatorial Pacific. A likely reason of these different timescales of DIC and pH/pCO2 are nonlinear processes in ocean chemistry described by the buffer factor (or Revelle factor; Revelle and Suess, 1957), which result in increases of pCO2 of approx. 10 times the magnitude of the corresponding relative increases in DIC."

ii.) Our results show that, depending on the research question, different variables might be of interest. We stress, however, that the correct estimation of trends relies on a sufficient length of time of the underlying data. We have expanded to conclusions:
"ToE of pH and pCO2 has rather low values (around 10 yr) in many regions of the surface ocean. It is, however, generally difficult if not impossible to reliably determine variability and long-term trends in the surface ocean from data that extend over such a short period only. Trends in surface ocean variables can vary significantly between different 10-year periods and even reverse sign (see Fig. 1 and Tab. 1, ALOHA data for an illustration). As a consequence, model data, or measurements, over a longer period are needed to reliably determine anthropogenic trends (Fay and McKinley, 2013) and the ToE. Here, trends and variability are estimated from 30 years (1970 to 1999) and 130 years of model data, respectively. The choice of a 30-year period minimizes the influence of climate modes such as NAO, ENSO or AMOC on trends as demonstrated by Fay and McKinley (2013) for surface ocean pCO2 measurements, while at the same time the 1970 to 2000 period still provides an approximate measure of the

current and near-future anthropogenic trend in the surface ocean. The ToE is indicative for the time required for the anthropogenic trend to leave the variability band, but it should not be confused with the period required to detect this trend in observational or model data. [...]
The study clearly illustrates the need of more long-term measurements with sufficient seasonal data coverage. DIC is a very important variable and crucial for our understanding of processes. For the sole detection of trends, however, pCO2 and pH seem to be a better choice. "

iii.) Results presented here are not in conflict with earlier results by Fay and McKinley and McKinley et al.

a) For example, Fay and McKinley 2013 investigated trends in pCO2 measurements between 1981 and 2010 considering periods of 4 to 30 years. They conclude in their summary and conclusion section (text in brackets included by us):
On decadal timescales [here less than ~25 yrs] *, signals of variability abound, and there are indications of influence from climatic oscillations such as PDO, ENSO, SAM, NAO, and AMV. However, these signals fade away as timescales lengthen, with the exception of PDO and AMV that continue to have influence on the longest timescales (Figures 4b and 8).*

The implication of this is that long-term trends in pCO2 can in general be reliably determined from a 30-yr period, as used in our study. A direct comparison of the trend signals computed by Fay and McKinley with our trend signal is hampered by the fact that Fay and McKinley use relatively sparse observational data to determine trends. We have revised the text in section 3.1 to read:
"This issue is addressed in a recent study by Fay and McKinley (2013). These authors investigated trends in surface ocean pCO2 measurements between 1981 and 2010 for periods of 4 years to up to 30 years. They found that, on shorter timescales, trends of surface pCO2 are sensitive to variability presumably linked to climatic oscillations and, consequently, may vary between different periods. Consequently, this caveat has to be taken into account when comparing modeled and observed trends over relatively short time periods. Fay and McKinley also find that the influence of climatic oscillations fades when analysis periods are between 25 to 30 years, e.g. as used in this study to determine trends. We note that a direct comparison of the trend signals computed by Fay and McKinley with our trend signal is hampered by the fact that Fay and McKinley use relatively sparse observational data to determine trends."

b) We also find large values of sdv in the equatorial Pacific ocean and the Southern Ocean (Figure 3) for surface ocean pCO2. Taken at face value, this is in agreement with the conclusion of McKinley et al, GRL, 2004 of a Pacific dominance to global air-sea CO2 flux variability. These authors state in the abstract: "We find that, for 1983–1998, both novel high-resolution atmospheric inversion calculations and global ocean biogeochemical models place the primary source of global CO2 air-sea flux variability in the Pacific Ocean. In the model considered here, this variability is clearly associated with the El Nino-Southern Oscillation cycle.

Both methods also indicate that the Southern Ocean is the second-largest source of air-sea CO2 flux variability." We do not comment on this in the manuscript As we investigate pCO2 and not air-sea flux. Variability in air-sea flux results both from variability in gas exchange rate and air-sea partial pressure difference.

I would like to encourage the authors to continue studying detection of trends, but feel that the clean statistical treatment is needed for publication.

[Minor comments]

Sect 2: In the decription of the ToE concept, it is unclear how time actually enters ToE, since trends are only calculated over a fixed period. Presumably S should not be trend but "trend times length of period" (otherwise "S/N" also would not be dimensionless).

see above

p 18070 line 8: "well approximated" is vague and needs explanation

Rewritten:

"For all models, we find that trends in global surface DIC can be represented by a linear function."

p 18070 lines 9-12: I expect that standard statistical tests exist for this question, that should be used.

To address this, a F test was conducted. Rewritten:

"For (ii), we investigate the detrended data (1870–1999) of DIC and SST of all 17 models (not shown). The comparison of sdv fields calculated for the first and second 65 yrs (F test, 5% level) illustrates that differences only occur very localized, consequently we suggest that this assumption is confirmed."

p 18071 line 1: Delete sentence as this has been said in Methods.

DONE

p 18072 line 10: "reasonably" is vague.

Rewritten:

"For DIC, features like a stronger trend at BATS compared to ESTOC are captured by the ensemble mean."

p 18073 lines 2,8: Putting ranges by ":" is unusual in texts and hard to read, rather use "-" or "...".

DONE: replaced with "-"

p 18074 lines 15-16: As far as I see, these two points are actually the same.

Rewritten:
"It illustrates where and when the models diverge and is thus a measure for uncertainty. "

p 18075 line 8: missing "of"

DONE: added

p 18076: I agree that these cases are revealing. Concerning the "Monthly" case, wouldn't it be a practically relevant option to deseasonalize?

With the "Monthly" case, we would like to illustrate the full range of intra-annual variability. This signal would get lost when the data is deseasonalized. Yet, there would be an alternative to the applied approach. We address this:
"For case two (monthly averages), an alternative approach would be to define N as the full range of the seasonal cycle. In doing so, it is possible to make a clear distinction between inter- and intra-annual variability. However, we focus on the combination of both since it is closer to what we find in reality. "

p 18077 line 9-17: I agree that the "aliasing" of seasonality into yearly values is the major problem here. As far as I see, similarity of July and January unfortunately does not mean that there is no such "aliasing".

The model IPSL shows similar patterns for January and July and thus no clear 'aliasing'. However, the majority of the models, with CESM as its representative, does. Rewritten:
"The "January/July" patterns are similar in parts of the global oceans. This indicates that, at these locations, statements based on irregularly sampled data are valid representatives for the whole year. In large areas, however, intra-annual variability might interfere such a generalization."

Fig 1: Hard to believe that the grey band gives the standard deviation (rather than e.g. +/- 2 sigma). Further, the agreement of spline and linear trend is seen to only apply to the specific period.

The grey band gives indeed +/- 2 sigma, which is indicated in the graph. For clarification, caption is changed to "The grey bar represents two times the standard deviation of the detrended time series (i.e., annual-spline)."
Concerning "agreement of spline and linear trend": Obviously, the trend over the whole time series is not linear. The spline shows distinctly stronger trends after approx. 1960. However, this study aims at the time period necessary to detect a trend signal with a magnitude comparable to recent observations. We suggest that the trend 1970-1999 is a more realistic estimate for the present and the near future than the lower trend values found prior to 1970 or calculated over the complete simulation. Consequently, the trend was calculated for the time period 1970-1999. And for these years, spline and linear trend agree.

Fig 3: Numbers are printed very small and not readable (probably even in full-width print). Variables should be arranged in the same order as in text and Tab 1 (also in Fig 2).

DONE: numbers are increased. Concerning the order: we prefer to keep the order of variables in Fig. 2 and Fig. 3. The reason is that the same label bars are used for DIC/SST and for pCO2/pH, respectively. Regrouping as requested would thus increase the complexity of the figures.

Fig 4: For the January and July cases, N is larger than for Annual, so why is ToE smaller rather than larger?

For DIC, the ratio S/N is very delicate – changes can have substantial impact. There are some localized deviations of the spatial patterns, both in S and N. For example, both January/July show slightly higher trend signals in the eastern eq. Pacific compared to annual. These deviations are big enough to cancel out the higher N, the result are comparably shorter ToE.

References:

Revelle, R. and Suess, H.: Carbon Dioxide Exchange Between Atmosphere and Ocean and the Question of an Increase of Atmospheric CO2 during the Past Decades, Tellus, 9, 18–27, 1957.

Von Storch, H. and Zwiers, F. W.: Statistical analysis in climate research, Cambridge University Press, 2001.