Reply to Reviewers for the manuscript egusphere-2022-96:" Inter–annual global carbon cycle variations linked to atmospheric circulation variability":

Na Li, Sebastian Sippel, Alexander Winkler, Miguel D. Mahecha, Markus Reichstein, Ana Bastos

Reply to Reviewer #2

The authors are investigating the ability of SLP anomaly field to predict global carbon inter-annual variability (IAV) when used in a ridge regression (RR). In particular, the IAV of de-trended global observed atmospheric CO_2 growth rates and modelled global land sink are reconstructed. This RR is compared to a another RR taking 15 teleconnection indices as predictors and to a linear regression only based on SOI. The use of SLP allows a good reconstruction of the different carbon cycle time-series IAV.

In general, the article is a bit difficult to follow. Indeed, the word 'global' is mentioned several times throughout the paper but its meaning is different whether it is CO_2 (single global value) or SLP (800 grid-point). An effort should be made to ease the reading. This paper is showing some potential. However, the paper needs some clarification/modification before publication.

Thanks for the in–depth comments and constructive suggestions. We address each point separately below.

Major comments:

A About the estimation procedure: what is the influence of the LOO consisting in using three consecutive years as test sample? What would happen if the test sample is bigger?

We would like to thank the reviewer for pointing this out. Since we have only limited samples (less than 60), we selected leave–one–out (LOO) rather than other cross–validation approaches with bigger test samples. We agree though that results may be sensitive to this choice. We select three consecutive years, but only the year in the center is used as a test sample each time. We remove the years before and after the test sample to avoid the influence of time–series autocorrelations. We note that in the dependence study (Fig. A3 of the manuscript), the time–series autocorrelation in CO_2 IAV is relatively small, so that this step might not have a strong influence on the predictability. However, the predictability also depends on the stationarity (particularly variability) of the dataset: the more stable the variance is, the less predictability difference when using leave–one–out or k–fold. Here, if the test sample is bigger, varying predictability is expected. To verify this, we conducted the k–fold cross–validation, with test sample sizes of k=1, 2, 3, 6, and 9. This time, all selected test data are predicted (our LOO approach removed the samples with year before and after the middle year), and we run 100 times of each estimation with different shuffled train/test grouping. Below we show the resulting correlations between predicted and test values:

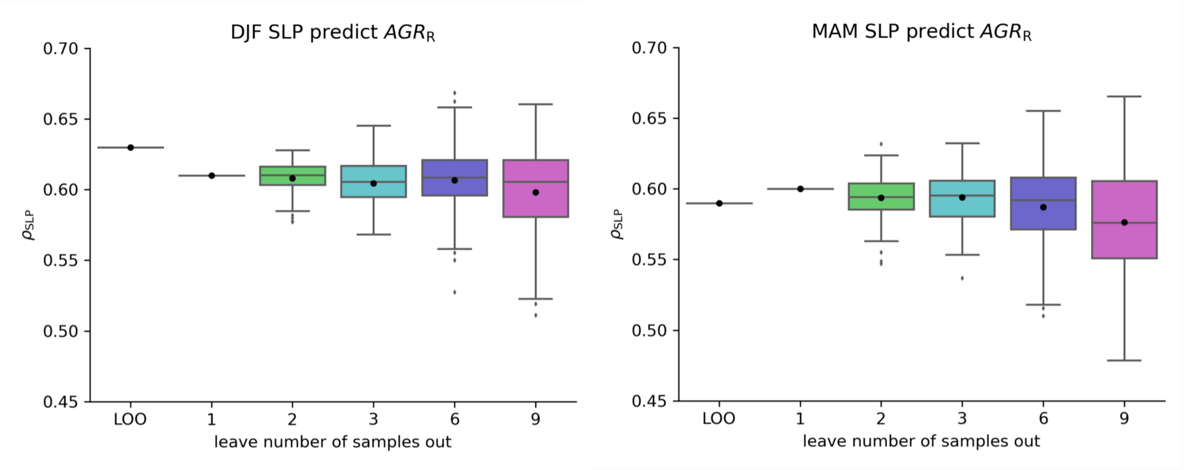

Figure R7. Comparison of global seasonal SLP predictability with different estimation approaches: LOO of our approach (leave–one–out with the samples before and after the test year removed from test and train group), leave–one–out, and leave 2, 3, 6, 9 out approaches, in the period 1959–2017.

The predictability shows the larger spread, the larger the number of test samples is, but the median remains relatively stable. Note that the k=1 leave–one–out approach has no uncertainty in this experiment, since no matter how many times we shuffle the train/test dataset, the predictability is the same. There is some difference between the LOO and k=1 cross–validation, it might be due to the removal of the years before and after the test sample. In such a case, the LOO has two less train samples (same train samples as k=3) compared to k=1 cross–validation, which might influence the predictability due to different train sample numbers. This test run shows that the smaller the test sample sizes, the more robust the predictability. We have added one note on this in the discussion text (line 168–169), which now reads:

" Given the relatively short period ($n < 60$), **and generally the smaller the number of the test samples, the more robust the predictability,** here we use leave–one–out (LOO)… "

B About the SLP anomaly fields as predictors: predictor numbers evolve from 4 to 800 depending the predictors domain. How ever it seems impractical to perform multiple RR with up to 800 predictors to estimate one global value and select the best predictor domain. If the intend of the authors is to provide an alternative to study the relationship between C-cycle and circulation variability this can be perceived as heavy. Besides, based on Figure 2, the SLP-based RR is not necessarily better than the indices-based RR or the SOI-based linear regression. A user would be tempted to use one of those.

We agree with the reviewer that “the global SLP–based RR is not necessarily better than the indices–based RR” according to Fig. 2 in the manuscript. If the user wants to evaluate a simple correlation between C–cycle and atmospheric circulation variability, teleconnection indices may yield equally good predictions. However, this approach might be too inflexible for certain applications, where identifying more general patterns of atmospheric circulation driving variability in the target variable might provide more comprehensive information. This is likely the case for local/regional studies for climate variables such as precipitation or temperature, as shown in Sippel et al. (2020), where the SLP–based RR was shown to be a more robust approach than using EOF–based circulation components for regression.

A more fundamental justification for this approach is that teleconnection indices summarize the variability of different modes of atmospheric circulation in simple time–series, where the spatial patterns of SLP, SST or associated variables for calculating the indices are usually fixed. This can lead to multiple, slightly different definitions of the same mode, for example the multiple indices that can be used to describe the El–Niño/Southern–Oscillation phenomenon. Moreover, teleconnections are known to interact, so that the resulting circulation pattern is a combination of different modes, for example, ENSO, SAM and IOD together influence the Australian precipitation and drought (Cleverly et al., 2016). Finally, teleconnection indices reflect the dominant modes of atmospheric variability, which are not necessarily the dominant atmospheric circulation patterns controlling CO_2 variability.

Our approach could identify the spatial patterns of SLP variability driving most of global CO_2 IAV (Fig. 3a in the manuscript) these patterns include the ENSO pattern but also reveal other important regions which do not correspond necessarily to a single index (e.g. the west Pacific area in MAM, which can be associated with multiple teleconnection patterns as discussed in the reply to Reviewer #1).

(a) The main problem is to compare results of regression with very different number of predictors only based on ρ_{SLP} . What is the trade-off between adding predictors and the RR improvement? Since the objective is to capture the IAV, using the principal mode of variability of SLP fields instead of the entire fields could remedy the aforementioned issue. For instance, the first EOFs of SLP fields can be used as predictors. The number of EOF can be chosen according the proportion of the variance captured by the EOFs. (b) RR is adapted for large numbers of predictors. It would be interesting to see the performances of a usual generalised linear model based on the EOFs of SLP fields.

We thank the reviewer for pointing out this critical and important issue. The Ridge Regression is especially tailored to address the problem of a large number of (correlated) predictors, since it attributes low weights to predictors that carry little additional information, while keeping those that do contribute the most to the variance of the target variable. Furthermore, we performed a preliminary sensitivity study as stated in section 2.4.1. By using different resolutions of SLP at

2 ° * 2 ° (16200 predictors), 5 ° * 5 ° (4536), and 9 ° * 9 ° (800) (Fig. A1 in the manuscript). The predicted correlations show only a slight difference with the changing number of predictors (e.g., r varied less than 0.03 with AGR_R), as expected by the use of RR.

The reason we did not use EOF of SLP fields is that EOF would capture the main variance of the SLP field, but not necessarily that of the main patterns that influence CO_2 IAV (similar to our reasoning about teleconnection indices in the reply above). Moreover, mathematically Ridge regression and EOF regression are deeply connected: EOF regression cuts off all components with small variance beyond a certain threshold, while ridge regression shrinks them, which allows that some information that is hidden in low variance components can still be used for prediction (von Wieringen et al., 2021). This approach might reveal hidden components of SLP that are driving CO_2 IAV. A comparison of these two approaches has been performed in Sippel et al., (2019), where the performance of RR and EOF when using SLP anomalies to predict temperature/precipitation variations are compared. They show that RR performs better than EOF (hence also justifying our answer to the previous comment). Finally, we followed the reviewer's advice and compared our results with those based on a generalized linear model using the first EOFs of SLP fields. The results are shown in Fig. R8.

First, we selected the first 10 components of SLP anomalies (DJF) using EOF analysis. We did this for three different SLP spatial domains (global, 18° N-18° S, and 18° N-72° S), and for the periods 1959-2017 and 1980-2017 separately.

We then reconstruct SLP fields based on an increasing number of components, from 1 (based on components 1), explaining 25% in period 1959-2017 and 23% in period 1980-2017 of the variance in SLP, to 10 components (1… 10), explaining 75% and 79% of the variance in SLP. We use these as predictors of AGR_R in a simple linear regression (also with LOO estimation).

The results are then compared with those using RR (Fig. R8)

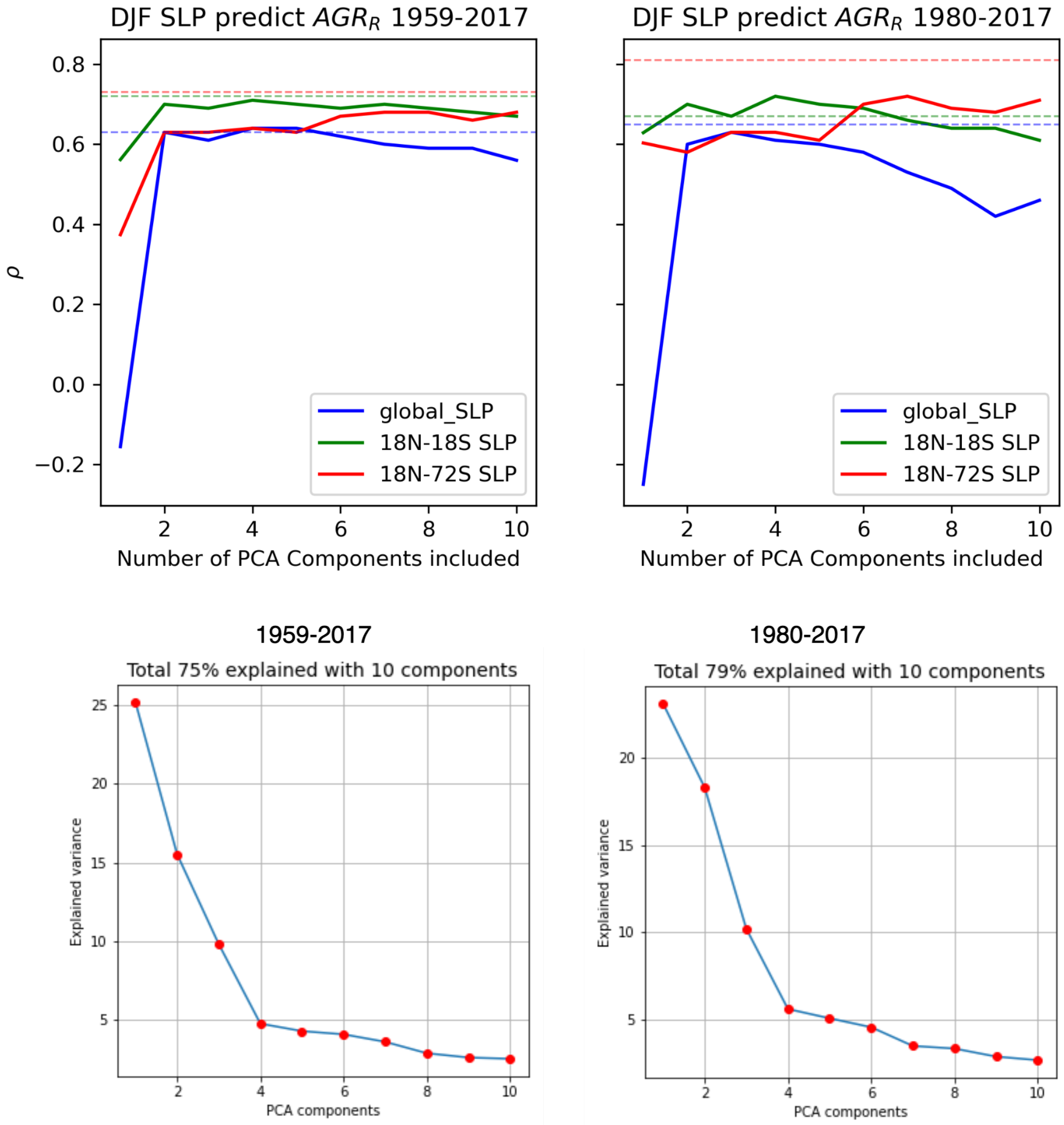

Figure R8. Comparison of EOF–linear to RR approach under different DJF SLP spatial domains. The upper plots are the comparison of predictability, The y label shows the Pearson correlation between the predicted CO_2 IAV to original CO_2 IAV. The solid line represents the results by using the EOF–linear approach under a different number of extracted SLP EOF components that are included in the linear regression. The dashed line represents the results by using RR. Different colored lines represent different SLP domains used. The lower plots are the corresponding explained SLP variance by different EOF components.

The EOF–linear approach returns in most cases lower predictability than RR, depending on how many components are included. According to Fig. 4 in the manuscript, the domain from 18° N to 72° S in DJF shows the highest predictability (r=0.81 in 1980–2017 and r=0.73 in 1959–2017) when using RR. In this SLP domain, the EOF-linear approach generally shows much lower predictability with all different numbers (1~10) of components included (r < 0.75 in 1980–2017 and r < 0.7 in 1959-2017).

This test shows that a EOF–linear approach can generally achieve lower/similar predictability than RR when selecting global/tropical SLP anomalies, which is consistent with the mathematical explanation above.

Minor comments :

— line 29: ’plagued’ may be a little harsh

We thank the reviewer for pointing this out, this sentence has been corrected to: “since some of these processes are ~~plagued~~ confounded by large uncertainties”.

— line 42 : Replace ’These dynamics’ by ’These climate variability modes’. These variability modes may be subject to irreducible noise but they can not be considered as ”noise”, please rephrase this.

Thanks, the sentence has been corrected to:” These ~~dynamics~~ **climate variability modes** are generated within the coupled atmosphere–ocean …”

— line 68 : In ”while at the same time”, at the same is redundant.

We agree with the reviewer and have removed “at the same time”.

— Section Data pre-treatment : clarify this section as follows : 1) trend removing (CO2, SLP and indices) and anomalies computing (SLP) 2) spatial and temporal aggregation.

We thank the reviewer for the good advice. We now add paragraph headers for "Trend removal" and "Spatial and temporal aggregation". We put the pre–treatment of CO_2 and teleconnection indices under “trend removal” and SLP under “spatial and temporal aggregation”. We have made some changes in the treatment of teleconnection indices. The new text in section 2.2 Data pre–treatment is:

Trend removal

The long–term trend of CO_2 time–series was removed by locally weighted scatterplot smoothing (LOWESS) of the annual time–series with fixed window size of 25 % interval longer than 30 years (1959–2017) and 45 % for shorter period (1980–2017). For monthly teleconnection indices, we first ~~remove the long–term trends by applying the LOWESS as for the SLP fields~~ calculate DJF, MAM, JJA, and SON mean values ~~accordingly~~, **we then remove the long–term trends by applying the LOWESS as for the CO_2 time–series** ~~SLP fields~~, and further include DJF and MAM combined (DJF+MAM) as treated in SLP (as described below).

Spatial and temporal aggregation

The monthly mean SLP fields are area–weighted and aggregated to 2 ° * 2 °, 5 ° * 5 °, and 9 ° * 9 ° spatial resolution, and the seasonal cycle removed by subtracting the monthly mean values for each pixel. We then aggregate SLP values in seasonal means for: December of the previous year to February of each given year (DJF), March–May (MAM), June–August (JJA), and September–November (SON) and further consider DJF and MAM combined (DJF+MAM), so the number of **pixel-based time–series (predictors)** ~~grid points~~ in DJF+MAM is double of DJF. **Note that a large fraction of the pixel–based time–series of seasonal SLP anomalies show no long–term trend, and the predicted differences between LOWESS detrended and no detrended SLP are small. Here we keep the analysis of SLP anomaly with no LOWESS detrending.** Here, we refer to DJF and MAM as boreal winter and boreal spring."

We would like to note that the SLP fields were not detrended (see reply to Reviewer #1). We have added two sentences in the above text:

"Note that a large fraction of the pixel-based time–series of seasonal SLP anomalies show no long–term trend, and the predicted differences between LOWESS detrended and no detrended SLP are small. Here we keep the analysis of SLP anomaly with no LOWESS detrending."

— from line 300 : scale is used to refer to the spatial predictor domain or temporal learning periods. Please be precise, in those case scale is not appropriate.

We thank the reviewer for pointing out, "scale of" has been removed.

— line 328 : Maybe 2001 instead 2003 ?

We thank the reviewer for correcting that, it is "2001" and has been corrected.