Reply to Anonymous Referee #2

We thank Referee #2 for his positive feedback on the Introduction and on the carbon and greenhouse gas budgets. His main criticism applies to the statistical method that we used to evaluate the relationship between fluxes and environmental variables. We intensively discussed his concerns in our group, and we consulted Prof. Hans-Peter Piepho who holds the chair of Biostatistics at our university. Please find below our reply to this and all other discussion points the referee gave.

The authors use the eddy covariance and open path LICOR 7700. I would like to hear more about flux detection limits, calibration and errors.

We will extend our material and method with this additional information.

As for gap filling while the Falge methods are standard for CO_2 fluxes other methods should be applied for methane flux gap filling. Look to the work of Sigrid Dengel et al. and others for methane gap filling. I suggest the use of artificial neural networks to gap fill methane fluxes.

The gap filling method of Falge *et al*. (2001) is a well-established and widely used method in the eddy covariance community. We found that the correlations between environmental variables and the CH_4 fluxes were at least equally strong as those with the CO_2 fluxes. Therefore, we are convinced that for our CH_4 fluxes the method of Falge *et al*. is as reliable as for CO_2 fluxes, and we do not see the needfulness to change this well-established method to neural networks.

The authors compute one Q_{10} value through the whole data set and come up with a non biological and indefensible Q_{10} value greater than 2 and near 4. While this may be ok for gap filling, it is wrong fundamentally and can be misused by modelers who may look for a Q_{10} from these data. From lab enzymatic studies we know Q_{10} value is near 2. We have learned from CO_2 studies that the Q_{10} will be artificially high when and the basal rate of respiration changes with the season. So the basal rate must be adjusted with time; this is the main lesson from the Reichstein (Reichstein et al., 2005) paper and Mahecha (Mahecha et al., 2010) paper.

We are familiar with the paper of Reichstein *et al*. (2005). The reason we nevertheless decided to use the whole season for estimating the temperature dependency of R_{eco} is because of the limited amount of available night data and the low variation in soil temperature due to the water saturated soil. This makes the regression less reliable. But indeed, in the field the dependency of CO_2 on temperature is not only caused by enzymatic activities of microorganisms in the soil but also by, for instance, the available labile carbon which varies within the season. Although, the overestimation as described by Reichstein *et al*. will not be that high at our site (due to water-logged conditions, day time soil temperature is on average only 0.4 °C higher than night time soil temperature) we will nevertheless change our method and use Q_{10} values calculated for every 2 months (more or less the plant development stages and still supported by enough data points). The range of Q_{10} would then be from 1.53 (September-October) to 3.93 (May-June). This would on average still be a higher Q_{10} value than that found by Mehecha *et al*. (2010). A possible reason may be the use of soil temperature instead of air temperature. Soil temperature gave a much better fit in the regression and soil respiration is also expected to have a high contribution due to the high carbon content in the soil. But plant respiration will also contribute, and is more dependent on air temperature. The range of air night temperature is higher than that of soil night temperature.

To make sure that the Q_{10} values are not falsely interpreted by modelers, we will consider renaming it and add a paragraph on the whole issue in the Discussion section.

There may be difficulty in interpreting methane fluxes as the source distribution may be heterogenous. The authors need to supply us with information on the flux footprint climatology.

In general we can say that we have a very homogeneous footprint, with mainly *Phragmites* vegetation inside of it. Therefore, we do not expect a large variation of fluxes, for example, depending on wind direction. Yet, we agree that the information about the footprint was somewhat limited, and we will include more detailed information about it in the Material and Methods section.

The authors should look at both photosynthesis and transpiration as potential drivers of their methane fluxes, too. Many are showing that exudate from photosynthesis primes microbes that produce methane, so fluctuations in light could affect photosynthesis and methane production at certain time scales. This is an important alternative or complementary path and production mechanism.

This is a very good and interesting point. We will add a paragraph on it in the Discussion section.

The authors inappropriate use linear regression models for a complex, nonlinear and multifactorial process is my biggest criticism. They are bound to misinterpret their data with such an antiquated statistical method. The Gil Bohrer team fitted their data with neural networks and looked at partial derivatives with environmental drivers to explain methane fluxes. More recently Sara Knox in a paper in JGR Biogeosciences used this method to study the controls of the environment on methane fluxes. The method seems to have much power. She and colleagues found superior description of their data using neural networks compared to a simple stepwise multi-linear regression model. At least the authors should do this. In addition, the field has advanced by introducing such methods as Granger Causality and Transfer Entropy to do a better job at linking methane and carbon fluxes with drivers such as light, temperature, humidity and photosynthesis.

There are many different statistical methods available that could be used to analyze relationships between factors in datasets. Neural networks is an interesting tool for revealing connections between the input data (in our case the environmental variables) and the output data (the gas fluxes). Indeed, the advantage of neural networks is that it does not make detailed assumptions of the relation between input and output. It could be linear, but not necessarily. The main argument of the referee to use the neural network method is because of the non-linear correlations we may have in our data. We absolutely agree that a multiple linear regression might limit the outcome of the data analysis. There are also, however, disadvantages of neural networks compared to multiple regression models (Breiman 2001 and Warner & Misra 1996 give a good overview of the advantages and disadvantages of stochastic models/Neural Networks). Seen from a statistical point of view, neural networks have severe shortcomings. It is for instance not possible to test the significance of the coefficients (weights in the wording of neural networks). It is also not possible to compare different models and say which is statistically best. And the biggest

disadvantage is that the functional relationships are hidden in a kind of black box. Knox *et al*. (2016) found a better fit of neural networks over linear models. But there are also examples of an almost equally good fit of a linear model compared to neural networks (Anagu *et al*. 2009).

Methods like Granger Causality and Transfer Entropy are interesting for forecasting. But due to the auto-regression term included in the models, a large part of the variation that could be explained with environmental variables is removed. And we are not directly interested in forecasting, but in understanding the causalities in the system.

Every method has advantages and disadvantages. We highly value the multiple regression-ARIMA method because of the fact that the detected relationships can be statistically analyzed and tested for significance. Therefore, we prefer to stick to the ARIMA method. We will discuss the pros and cons of our method and that of Knox *et al*. (2016) and related ones in the revised paper.

We see as well, that the non-linearity in our data is a problem that could lead to false interpretation. To address these concerns, we will extend our ARIMA approach by testing nonlinear (quadratic and interaction) terms.

Added references:

Anagu, I., Ingwersen, J., Utermann, J., Streck, T. (2009) Estimation of heavy metal sorption in German soils using artificial neural networks. Geoderma **152** 104-112

Breiman, L. (2001) Statistical Modeling: The Two Cultures. Statistical Science **16(3)** 199-231

Warner, B., Misra, M. (1996) Understanding Neural Networks as Statistical Tools. The American Statistician **50(4)** 284-293