Reviewer #2: Comments

I agree with referee #1 that the paper ’Uncertainties in the national inventory of methane emissions from rice cultivation: field measurements and modeling approaches’ by Zhang et al. is an important and nice study regarding general uncertainties evolving during regional/national GHG emission inventories. I also agree with referee #1 that national estimates of CH4 emissions should be more emphasized. My main criticism relates to the presentation of the study. Material and Methods, Results and Discussion sections all need revisions in order to improve the reader’s access to the main points of this study (see specific comments).

Re: We thank the reviewer for the comments and made revision to the MS to show the national estimations of CH_4 emission, and also the writing of the MS emphasizing the formulation and nomenclatures.

Specific comments:

P1 L25: Mention that regression models are taken from literature.

Re: Revised (P1 Line18 in the ‘clean revised manuscript’).

P1 L27-28: Use clear measures and give respective values instead of using the vague term ’model performance’ only.

Re: The ‘model performance’ refers to how the model representing the variation in the observations, evaluated by the difference between the observations and the corresponding model outputs. Conceptually, the model performance here covers the ‘parameter uncertainty’ and ‘model inadequacy’ in Kennedy and O’Hagan (2001) and errors in observations, because we can’t distinguish them with model validation, which was used to evaluate the ‘model performance’ in the present study. The measures to quantify the model performance here are two statistical parameters of the modelling residuals (difference between the observations and modelling outputs): bias (means of the residuals) and variance (statistical variance of the residuals) as showed in Equation 1 and Equation 2. We revised the equation and the main text to make the meaning of the terms clearer.

P1 L30: Absolute values of simulated methane fluxes are meaningless here since context (e.g., different irrigation, straw management, ...) is not clear yet.

Re: The modelling result here is the result of CH4MOD with available information of irrigation, straw management and soil properties of paddies in rice cultivation of China. We revised the sentence as ‘As simulated by CH4MOD with data of irrigation, organic matter incorporation and soil properties of rice paddies, the modelling methane fluxes varied from 17.2 kg CH_4 ha^{-1} to 708.3 kg CH_4 ha^{-1}’

P4 L19-21: Statement is not very intuitive. Why should ’non-key’ factors lead to significant errors? Factors leading to significant errors are implicitly named key.

Re: Here we intended to say that they were not ‘non-key’ at all. To avoid misunderstanding, we revised the sentence by replacing ‘non-key’ with ‘other’.

P5 L21-22: Imprecise formulation, inaccuracies of models are manifold and should be defined more clearly based on common nomenclatures in literature, see for example nomenclature and definitions by (Kennedy and O’Hagan, 2001). Nomenclatures and definitions should be revised and standardized in many parts of the paper. Kennedy, M.C., O’Hagan, A., 2001. Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B Stat. Methodol. 63, 425–464. doi:10.1111/1467-9868.00294

Re: We thank the reviewer for the comments and recommending the literature. Here, the model inaccuracy refers to the combination of ‘model inadequacy’ and ‘parameter uncertainty’ in Kennedy and O’Hagan (2001). In other places of the MS, ‘model fallacy’ means the same. We revised the MS to use the term ‘model fallacy’ throughout the MS and explain explicitly the mean of it. We also rewrite the equations of the MS.

P5: L34-36: Why were these two regression models chosen? It would be very interesting to see how IPCC emission factors, which also account for, e.g., different amounts of straw and different irrigation schemes would behave.

Re: One of the objectives of the study was to compare the performance of models with different complexity with different levels of data availability. We chose the two regression models because: 1) they had been used to estimate regional/national/global methane missions in many previous studies, and 2) they differed from each other and from CH4MOD explicitly in levels of complexity. There are many other models that developed and used widely in modeling methane emissions from rice paddies and wetlands *etc*. But we can’t tell which one is more complicated in structure that the other. We briefly explained it in the revise MS (P6 Lines 13-15).

P6-7 Formulas 1-4: Unclear why these measures have been used. Give proper descriptions, meanings and references to ’bias’ and ’total error’ and compare both to each other.

Re: ‘bias’ is the statistical mean of the modelling residuals. We admit that ‘total error’ is not a proper term. In the MS, it means the ‘mean of squared errors’ in model validation. In the revised MS, we used the term ‘mean of squared errors’ in the main text.

P7 L12: ’ errors in the performance of the method’: unclear formulation, use consistent nomenclature for different error/uncertainty sources

Re: revised as ‘model fallacy’. Because in the MS, our emphasis was on the quantification of the uncertainty in the national inventory by modelling approaches, the rationale of the uncertainty was mainly provided in the Supporting Information (Appendix C and D).

P7 L 15: Give more information regarding your Monte Carlo simulation and PDFs since this is an important determinant of posterior uncertainty.

Re: To measure the uncertainties in model outputs due to insufficient data quality and availability, we applied Monte Carlo simulations to the CH4MOD model. Statistical characteristics were derived from the available datasets to develop probability distribution functions (PDFs) for each model input variable. The PDF of field irrigation were defined by the occurrence percentage of each irrigation pattern (Table B2). Table B1 shows the statistical parameters of the PDF (normal distribution) of organic matter incorporation in each province. The PDF of the soil sand percentage was also built as normal distribution with parametric information from the literature (Shi et al., 2004).

We performed Monte Carlo simulation in the way of randomly drawing values of the model input variables from their PDFs and then run the model (e.g. CH4MOD). This process iterated 1000 times and at the last step, the mean and 95% CI of the calculated methane fluxes were derived from the iterations (P8 Lines 3-6).

P8 L1 On what is this assumption (amount of stubble) based?
P8 L1-11: What is the difference between stubble and incorporated straw?

Re: Stubble is the part of rice stem that left after rice harvesting. Traditionally, both the rice grain and rice straw were harvested and stubble was left in field. The harvested straw may be taken away or left in field, but stubble was always left. The amount of stubble accounts for about 10% of the aboveground biomass of rice according to previously published literatures (Huang et al., 2004; Zhang et al., 2011). We noted the literature in the revised MS (P8 Line24).

P8 L15-16: Be more precise here and mention considered irrigation schemes and how the model handles them.

Re: The irrigation in rice cultivation in China was summarized into five patterns: 1) flooding-drainage-flooding-intermittent irrigation, 2) flooding-drainage-intermittent irrigation, 3) flooding-intermittent irrigation, 4) continuous flooding and 5) continuously intermittent irrigation (Gao and Li, 1992; Huang et al., 2004). Appendix B in the Supporting Information provides necessary information of the irrigation in China. Table B2 list the percentage of each water pattern in different regions of China. More information of how CH4MOD handles the irrigation may refer to the literature of the model development (Huang et al., 2004). We also add brief description of the irrigation in the revised MS (P8 Lines 32-34 and P9 Line 1).

P8 L28: Probably Appendix B is meant.

Re: Revised.

P8 L33: In section 2.4, the description of used formulas should be improved since the combination of model and model input uncertainty is a central point of this study. The derivations of formulas in the Appendix are unclear. Give consistent names and meanings to each symbol that is used. Parts in the Discussion sections refer to the meaning of formulas and measures and should be moved here.

Re: We revised the relevant part of the MS and the Appendix, emphasizing the consistence of the names and expressions.

P9 L27: Do you mean ’harvested-area-weighted’ or ’cultivated-area-weighted’? Since cropping intensity (number of crops per year) varies, the weighted mean should be derived based on harvested area. In addition to area weighted means you should also consider seasonal means. A given amount of data may refer to different seasons, e.g., winter, spring, summer and autumn with strongly varying potentials of CH4 emissions. Most likely the seasonality distribution of observations does not correspond to the actual seasonality distribution of rice cultivation in China.

Re: Yes. The ‘area-weighted’ in the MS means ‘harvested-area-weighted’ and we revised the expression.

In China, the rice cultivation is different from north to south: single rice cultivation in north-eastern China, rice-upland crop rotation in eastern China and double rice cultivation in southern China. The ‘harvested-area-weighted’ analysis in the present study distinguished the harvested area of different water irrigation, because irrigation is the most important factor for methane emission. Seasonality also affects the methane emission but not as important as irrigation, according to both observational and modelling studies (Yan et al., 2005; Zhang et al., 2011). We agree with the reviewer that mismatch between the seasonality of the observations and the actual rice cultivation in China may bias the national estimation of CH4 emission via the statistical summation of the observations, and contributes to the uncertainty of the estimation.

P10 L18-21: Be more precise how measurements are dependent from each other. The potential dependency of measurements is not discussed in the Discussion section.

Re: The dependence of measurements here means the possible spatial correlation among them because of the common environmental conditions they may share. It is not the meaning that they were dependently obtained by sampling. The spatial aggregation of the measurements obtained at different places to produce national estimations may introduce biases if the spatial correlation among the measurements were not handled properly. We didn’t make in-depth discussion about the spatial correlation because it is beyond the topic of the present study. In the revision, we revised the sentence to clear that it is about the spatial correlation. We also provide literature reference (Legendre, 1993; Dormann et al., 2007) for those interested in spatial correlation.

P10 L12-16: Standard Error (SE) and deviation are very common measures and do not need explanations/references. To my understanding, the presented SE refers to the variability of different observed mean fluxes from different field sites. How are measurement errors reflected? What do you mean with representative error?

Re: Yes, we agree with you concerning the standard error. The measurement errors were not discussed separately in the present study. The reason of doing so was explained in Appendix C. The representative error in the present and other literatures (e.g., Van Bodegom et al., 2002a; Verburg et al., 2006) stands for the representativeness of the measurement obtained at a site to the area that enclose the site.

P10 L25: Present average values of overestimations for both models.

Re: Revised.

P10 L35: Why is ’total error’ and not ’bias’ interpreted as model performance? In order to underline this statement, more measures should be used, e.g., root mean squared error, R2, model efficiency.

Re: ‘bias’ is the average of the modelling residuals, accounting part of the errors. We use ‘mean of the squared errors’ to interpret model performance. ‘total error’ is not a proper expression and we replace it with ‘mean of the squared errors’ in the revision. There are other indexes, e.g., R^2 and RMSE, we use bias and ‘mean of the square errors’ in the MS owing to they are directly comparable to the errors from data availability.

P11 L22: I miss the discussion of these values. Are such uncertainties small or large compared to other studies?

Re: The within-grid estimation error ($\sigma_{T,i}$, calculated with Equation 5) is the error in each grid cell due to both the model fallacy and data scarcity when making estimation of a grid cell (10 × 10km). They are not shown in details because we emphasized the uncertainty in the national inventory, which was the spatial aggregation of the uncertainty in each grid cells. We didn’t compare the result of the ‘within-grid estimation error’ in the present study with other studies because no study had make estimation of the uncertainty in the way of the present study.

P11 L23-25: Discussion is missing.

Re: In the revision, we discussed the difference of the national methane emissions and the uncertainties estimated with different approaches and the data availability scenarios, as showed in the revised Table 3.

P11 L30 - P12 L14: This is rather introduction and representing of results than discussion.

Re: Thanks for this comment. We revised the MS by moving it to the introduction.

P12 L10: Temporal variations are not presented.

Re: Here in the sentence, we noted that there are temporal variations, annual, seasonal and even diurnal, in the methane emissions. But in the present study, we discussed the spatial variation and the estimation uncertainty in the national inventory of a specific year. Temporal variations of the methane emission were not discussed.

P12 L15-17: Unclear argumentation.

Re: Thank you for pointing it out. We revised the sentence as ‘This was partly due to the discrepancy in the spatial representativeness of the methane fluxes in field observations and model estimations’.

P12 L18-22: Unclear argumentation. Model performance was assessed with site-specific input and not with regional averages. The representation of experimental measurements for larger regions and associated uncertainties should be independent of models. Discussion of comparison between model and measurements at site scale could be moved to a separate subsection.

Re: This is what the ‘representative error’ means, which had been discussed in previous studies (Verburg et al., 2006; Van Bodegom et al., 2002a) and described in Appendix C of the MS. Model performance was assess with site-specific input. Here the ‘site’ means a small scale (e.g., a hectare or smaller) instead of a ‘point’, when the experimental sampling was taken at several ‘points’ called ‘duplicates’ at the experimental site. When we use the model for regional estimation, we make estimations for each grid cell (10 × 10 km in the present study). The mismatch of the scale supports the meaning of ‘representative error’.

P12 L31-33: Should be moved to the Results section. Use consistent nomenclature, i.e., the term ’model fallacy’ has not been used beforehand. Do not repeat formulas from the Material and Methods section in the Discussion.

Re: Thanks for the suggestion, we revised accordingly (P12 Lines 24-26).

P12 L35 - P13 L14: Much of this information belongs to the Material and Method Section and to the Discussion. Key results (e.g., ’56.6% of total uncertainty originates from the model’), which are also presented in the abstract should be first presented in the Results section and subsequently discussed. Appropriate discussion regarding the different uncertainty sources (model versus input) is missing. Argumentation regarding ’imprecision random noise and/or unknown factors’ is unclear.

Re: Section 4.1 discussed the different error sources to the uncertainties in the inventory. This paragraph around Fig. 8 was about the aggregation of $\sigma_{v,i}^2$. Material and Method Section described how the errors were quantified and aggregated, as showed in Fig. 2. We thank the reviewer for the revision suggestion and revised the MS accordingly.

P13 L27 - P14 L16: Remove this section from the Discussion. This is partly Material and Methods and seems to be an arbitrary example of model parameter uncertainty that has been neglected and thus is not much contributing to this study.

Re: Section 4.2 discuss how model improvement (e.g., parameterizing rice cultivar more specifically) affect the uncertainty analysis. We agree with the reviewer that the model parameter uncertainty wan not separately analyzed in present study. But because the parameter uncertainty contributed significantly to the model fallacy, it should be noted briefly in the discussion.

Fig. 5: Use identical axes for all plots.

Re: We guess you meant Fig. 5. We had at first used identical axes for Fig. 5-(a), Fig. 5-(b) and Fig. 5-(c). But it looked a little awkward, we, therefore, changed the y-axe of Fig. 5-(a) and kept the other axes identical.