This paper presents a nice analytical trick to use fluxes optimized through atmospheric inversion to validate fluxes from bottom-up process models in a novel way (or at least novel to me). One of the weaknesses of doing this directly is that there can be considerable influence of the prior still in the posterior fluxes, especially when the measurement constraint is lacking over certain times or regions. By applying the averaging kernel of the model to the process model fluxes (somewhat like calculating the
smoothing error in the observation operator when comparing satellite column to modeled profiles), the authors have corrected for the influence of the prior. This is indeed a clever trick, with several potential future applications, but I am concerned that the authors understate some of the shortcomings of such an analysis, and overstate the information gained and the potential applications in the context of ESM projections.

Firstly, the shortcomings: This clever trick does nothing to alleviate potential problems with model transport. If, for example, the interhemispheric mixing of the model is too fast/slow, this will directly over-/underestimate northern hemisphere emissions. Similar problems occur for errors in vertical mixing, tropopause height, and stratospheric gradient (the latter two of which are of particular concern when interpreting XCH4). This does not mean that the approach is invalid, but some discussion of these caveats is sorely needed.

Furthermore, GOSAT (unfortunately?) does not only measure methane emitted from wetlands. While that would simplify matters considerably, the interhemispheric distribution of methane fluxes and concentrations is also largely dependent on anthropogenic emissions, which have a distinct interhemispheric gradient with higher values in the northern hemisphere, and not inconsiderable uncertainties. This aspect is barely touched upon, and needs to be clearly explained and discussed, also in the context of the suggested application to carbon dioxide.

Finally, I think the authors may be taking the potential applications in terms of ESM projections a bit far. This is fundamentally a data-driven constraint on process models, or perhaps better, on datasets upon which process models are dependent. Because that is one of the main findings, isn’t it? That the same model can belong to either the high or low performing ensemble depending on which wetland extent map is used. (This is not really a new finding, but rather a new way to get at this information.) It is not clear how this sort of benchmark based on atmospheric measurements should be used to analyze ESM projections. There is no reason to think that the correlations that have been deduced for this 3- or 8-year period will hold for projections for the year 2100. Specifically, in the text:

L292-292: I’m not sure I understand what you’re trying to argue with the following sentence: “Our Bayesian methodology can be used to evaluate the probability of Earth System Model (ESM) projections based on the zonal profile of wetland emissions.” This implies that only with this Bayesian methodology can you show that future wetland extent is going to be correlated with future wetland emissions... With no future data to see what the atmospheric methane distribution will be, it is hard to see what additional constraint would be brought here. Note that also the LP models have a high correlation of emissions with wetland extent - this Bayesian methodology was really more useful at sorting out which wetland extent/model combinations produce realistic fluxes, constrained by atmospheric methane measurements.
L302-303: "can" seems a bit strong here, perhaps "might" or "could"? Also, I'm not clear on what is meant by using your analysis "as emergent constraints on ESM CH4 predictions". Perhaps it would be more reasonable to suggest that your analysis could be used to test emergent constraints (in simulations from past years) that have been deduced from ESM CH4 predictions? There is a logical flaw in arguing that a data-constrained result can be used to test the plausibility of climate projections, when there are no longer data to constrain the problem. This is essentially extrapolation into a regime that you haven’t constrained with your model. Using this approach to test the reliability of process models (or perhaps more importantly, the robustness of the wetland extent simulated in ESMs?) based on current observations is reasonable, looking into the future, less so.

The novelty of the approach is however significant, and worthy of publication. Whether this is best in a general-interest journal such as Advances, or rather in a more specialist journal is fundamentally an editorial decision. The methodology is really what is interesting and unique here, the scientific findings are actually not all that groundbreaking, and the estimated latitudinal distribution and magnitude of wetland emissions is generally consistent with the findings of e.g. top-down estimates from the most recent GCP methane report, among others, as the authors point out in the discussion. The claim in the title, that this approach has allowed them to determine the climate sensitivity of wetland emissions also seems a bit overblown: Figure 4 shows rather that the climate sensitivity of the highest-performance and lowest-performance models are nearly indistinguishable. What is the knowledge gain here, really? Much of the difference between the upper third and the lower third is related to a better wetland extent map rather than an improved understanding of biogeochemical processes or climate sensitivity.

Based on this, I would recommend this study for publication, but would suggest first walking back some potentially unrealistic scientific claims and including some caveats related to the uncertainties in the inverted fluxes.

Regarding data availability: Sometimes it is not entirely clear to me which data should be made available. The authors provide access to the GCP wetland simulations, WetCHARTs and CARDOMOM, but not e.g. the posterior wetland fluxes or the averaging kernel of the transport model. The AGU data policy states that “Data used to generate, or be displayed in, figures, graphs, plots, videos, animations, or tables in a paper” need to be made available. That has not been done, but many papers are published without making such data available.

Upon reading the other reviewers comments: I agree fully with Reviewer #2, who managed to explain his or her concerns much more succinctly than I did. I’m not sure Reviewer #3 fully grasped the trick of projecting the fluxes through the averaging kernel of the inversion: this is a tricky idea for people who aren’t very familiar with inverse modelling and/or satellite retrievals. However also this reviewer shared the concern that
the novelty of the results are overstated, as one of the main results is that having a better wetland extent map leads to better wetland fluxes - which is hardly novel.

I also had many, many minor typographic comments. I have a lot of patience with language errors when none of the co-authors are native English speakers, but in cases like this I wonder if all of the co-authors have even read the whole paper before submission. This seems to be a more common occurrence with journals that do not have an open review process - the authors know it'll get cleaned up before publication anyhow. This places an undue burden on reviewers, and makes me rethink reviewing for AGU journals going forward.

Minor typographical concerns:

L3: affiliation of Zhang should be 2,3 I suppose (instead of 23)

L28: account -> accounts

L30-31: This is a bit awkward (especially for the plain-language summary): "hydrological controls on tropical wetland emissions are the dominant regulator of global wetland emission contributions to the atmospheric CH4 budget" Perhaps just: "hydrological controls are the dominant regulator of tropical wetland emissions" would be enough?

L40: Would recommend removing "corresponding variable" here.

L47: a global wetland CH4 emission -> global wetland CH4 emissions

L48: place -> places

L60-61: The very first sentence of the abstract of Kirschke et al. (2013) is: "Methane is an important greenhouse gas, responsible for about 20% of the warming induced by long-lived greenhouse gases since pre-industrial times." Also the IPCC AR5 gives different values: "Emissions of CH4 alone have caused an RF of 0.97 W/m^2", but with a "concentration-based estimate of 0.48 W/m^2" (which I guess would be direct?), compared to 2.83 W/m^2 due to changes in the concentration of all well-mixed greenhouses gases. Not sure where the one-quarter number comes from, but it's not the references given.

L63: budget -> budgets

L72: emissions -> emissions; also maybe change "the key" to "key"? It's not the only uncertainty...

L77 (and elsewhere): "to inverse" is not a verb. One could use "inverted fluxes", or, better, "fluxes resulting from atmospheric inversion".
L89: estimate -> estimates

L96: ingredients -> the ingredients; steps -> the steps (I would also suggest a different word than "ingredients", perhaps "inputs", or "datasets", or similar? "ingredients" sounds like cooking.

L117: of a priori -> of the a priori

L122: emission -> emissions

L123: as L117

L125 (and throughout): Perhaps it would be more appropriate to refer to 42 different wetland model setups or 42 different simulations of wetland methane fluxes, rather than "42 global wetland biogeochemical models". Changing the wetland extent map alone does not really make a new model... Also, the models themselves are not compared to the inversion results, but rather their simulated emissions.

L136: I’m not sure "year-difference" is completely clear. Perhaps: "the effect of using different time periods"?

L162-163: I guess just agricultural wetland emissions (rice) or agricultural emissions resulting from intentionally inundated fields are meant? This should be clarified.

L171: factor -> fact

L171: I think it would be clearer with: this cross-correlated information -> information cross-correlated

L176: is -> is the (twice)

L177: is -> is the

L178: emission -> emissions

L183: Perhaps should "therefore" be "then"? "Therefore" suggests causality related to what came before, as in, the projected fluxes are compared to the posterior fluxes to rank the models because the inversion-model mismatch is enlarged during the information-rich period. But that's not really it I think...

L188: use posterior -> uses the posterior

L192: is -> is the (twice)
I don't think you need to include this definition again in parentheses...

models -> model

emission -> emissions

from -> from the

emission -> emissions; agrees -> agree

from -> from the

emission -> emissions; is -> are

Is it really "much less than", when the ranges overlap so much?

emission -> emissions; is -> are; more than -> more than those of

"two-fold from the LP model" -> "a factor of two by the LP ensemble"

emission amplitude -> emissions

emission -> emissions

as a -> to be the; emission -> emissions

emission -> emissions

Perhaps "steps" should be “factors”?

indicate -> indicate that; emission -> emissions

of -> for

It's a bit circular to mention good agreement with the inversion results you used to determine which models performed well, isn't it?

are -> have been

"we can’t assert or refute" Perhaps better: "we can neither assert nor refute"

would just use "is" (as you can't take a stand on either side)
This manuscript presents an interesting method for comparing independent flux estimates from e.g. land surface models to fluxes retrieved from atmospheric inversions. This method rigorously takes into account the dependence of the fluxes from the
inversion on the prior information used in the inversion through the use of the averaging kernel matrix. This method could be broadly applicable to other types of fluxes and to other species. One limitation, however, is that the averaging kernel is not calculable in a lot of inversions (e.g. in those using 4D-var algorithms), which is more often the case for inversions using satellite data and at the global scale.

Although the method is interesting, I think the authors overstate its potential. For example, it is not clear how this method could be used for validating the predicted fluxes from Earth System Models, which is one claim made. Also, there is no mention of systematic errors in the inversion and how these may influence the results.

Another concern is that the authors are sometimes not very precise in what they state, both in the methodology and in the results (see specific comments).

Lastly, I am concerned about the particular inversion that was chosen for this study as it uses a linear method to solve a non-linear problem (see specific comments). I think the impact of this on the fluxes from the inversion needs to be quantified, and thus the impact on the results of this study.

Given these concerns, I recommend major revisions.

Specific comments

P3L37: Please change “inversed from atmospheric CH4 concentration” to “from atmospheric inversions of CH4 concentrations” since it is not the concentrations that are “inversed” but rather the relationship between the fluxes and atmospheric concentrations.

P3L39-40: I do not agree that “the challenge of using top-down estimates to refine bottom-up models” is “mainly because of variable a priori emissions and corresponding variable spatial resolution” as the authors say. I think the challenges of the uncertainties in the satellite retrievals and the model representation errors, which translate into flux errors from the inversions are equally relevant.

P3L44-44: By accounting for cross-correlations and uncertainties in the inversion estimates does not mean that only the information from the satellite observations is kept as a constraint. The fact that an atmospheric inversion is used means that the information provided is about the fluxes (that is having solved for atmospheric transport) so it is not “only” information from the satellite observations.

P4L68: Please specify what is meant by “future climate feedback”, isn’t the climate feedback simply a result of the sensitivity of the model to meteorological forcing etc. or the “biogeochemical processes” as the authors write?
P4L69: Please specify what is meant by "sampling" in this context.

P4L70: Please also add "atmospheric chemistry" to this list of top-down sources of error as the OH sink is a large source of uncertainty in the atmospheric chemistry transport models.

P4L74: Please change to: "Atmospheric inversions using space-borne observations provide a constraint on the global scale spatial-temporal variability of wetland CH4 emissions"

P5L79: Remove "explicitly" since the inversions do not estimate this at all.

P5L81-81: Please change "limitations in the observing system" to "limited observational constraints"

P5L82: This sentence is ambiguous, on one hand when the observational constraint is relatively weak, the inversion solution depends strongly on the prior estimates - I think this is what is meant here. When this is the case, it follows that “the interpretation of emissions processes” (based on fluxes from inversions) will also depend on the prior estimates.

P6L102: Remove "emission" before "fluxes" since the inversion estimates "fluxes" and "emission" is redundant here. And again on L105 change "emissions" to "fluxes".

L107: I think it’s important to state that the inversion of Zhang et al. 2020 estimates CH4 emissions from wetland and non-wetland emissions separately. I also read Zhang et al. and I’m a bit concerned about this inversion since they attempt to optimize OH as well, which makes the problem non-linear (since the amount of CH4 lost due to the OH reaction depends on the CH4 concentration) but their optimization method assumes linearity. I’m surprised the reviewers of Zhang et al. didn’t comment on this. I see that many of the same authors on Zhang et al. are also authors on this manuscript. The question is, what is the impact of using a method for the inversion that assumes linearity when the problem you are trying to solve is clearly non-linear? I think the error in this assumption needs to be quantified, and if it is large then another inversion result should be used in this study.

P7L117: The dependence of the inverse solution on the a priori information can be expressed as a weighting between the "true" fluxes and the prior fluxes: x_pos = Ax_true + (I - A)x_pri where A is the averaging kernel and x_true are the "true" fluxes if these would be known. Thus the impact of the prior is not just "spatial uncertainties and cross-correlations" but rather all components that determine A, i.e. the observation operator, the prior and observation error covariance matrices, as well as the number of observations.
P8L149: I am not sure if it is generally true that all GCP models are land components of ESMs. It would be more correct to say that these are all land ecosystem models.

P9L180: The approach here does not remove the impact of the prior information but allows the modelled estimate (xmodel) to be "viewed" in a consistent way with the inversion accounting for the prior information. This would be the more correct way to state this.

Fig. S1: It would be helpful to state what model is used here

L194-196: I think the list of biogeochemical models with the wRMSE should be given somewhere as it's relevant to see if there are similarities between the models that ranked as HP/LP.

L201: I'm not exactly sure what is meant by this sentence. Do the authors mean that for a regional estimate they calculated a weighted average of each variable where they weighted by the amount of CH4 emission? This should be stated more clearly.

L284: Please add "emission" after CH4 in this sentence to be clear.


L294: Actually both HP and LP have equal sensitivity to precipitation and wetland extent. Also, surely the result that wetland extent strongly determines wetland emissions is trivial.

P16L302: Please specify that this is for tropical wetlands. In contrast to the tropics, the results show that non-tropical (and global) wetland emissions are strongly dependent on temperature.

Fig. 4. It is a bit difficult to distinguish which bars belong to which variable and it would help to either draw horizontal lines between the bars for the different variables or to increase the gap between them.

P16L303: I think the claim that this analysis can be used as a constraint on ESM CH4 predictions needs to be better justified. Using inversion flux estimates to validate biogeochemical results is sound, but in order to validate ESM predicted CH4 emissions one first needs to validate the climate predictions. Also, in a future climate, it is unclear how changes in vegetation cover will affect the emissions, and in high latitudes predictions of hydrology are highly uncertain with some models predicting an increase in wetland extent and others a decrease (e.g. Andresen et al., The Cryosphere, 2019).

P17L332-P18L347: It is not clear to me how this section "Future Directions" links to the analyses and method described in this paper.
Reviewer #3

This article addresses the emissions of CH4 from wetlands. The study uses satellite-derived emissions of wetland CH4 to analyze a selection of process-based models and isolate those that provide the best fit to these satellite-derived emissions. This selection of refined models are analyzed to examine the correlation of wetland CH4 emissions to the main underlying drivers of wetland extent, temperature and carbon availability. The manuscript is generally well-written and to the point. However, I have some questions regarding what the implications and novelty of this work are, since the satellite constraint on latitudinal distribution is already provided in the work of Zhang et al (2021), and the
analysis of the sensitivity of CH4 to the underlying physical drivers adds little new in its current format.

Major comments:
This study is largely based on the work of Zhang et al (2021, hereafter Z21) which estimated global wetland CH4 emissions to be 145 Tg/yr between 2010 and 2018. The selection of high performing process-based models (HP) are then calibrated to fit the results of Z21. The finding that the median emissions from the HP models are 148 Tg/yr is therefore hardly a surprise, demonstrating a calibration to the results of Z21. The regional emissions distribution of HP models shown in Fig. S2 is also seen to be very similar to that of Z21 (Figure 11). Again, given the HP models are selected based on their fit to Z21, this is perhaps to be expected. Therefore, it seems an overstatement to say in the abstract (also the plain language summary and key points) that the study places new constraint on the latitudinal distribution of wetland emissions. Surely this constraint is provided in Z21, and the HP models selected here are merely those that demonstrate the greatest consistency with that constraint?

The novelty of the work seems to boil down to an analysis of the underlying controls on what determines whether a model is in the HP or LP group. Currently this is limited to an analysis of correlation coefficients between CH4 emissions and the underlying controls of wetland extent/precipitation, temperature and carbon availability. This analysis shows no significant difference between HP and LP models. The major factor in what determines an HP vs LP model would seem to be the total global CH4 flux magnitude, which largely seems to be a result of the wetland extent map used. This is a useful result, but beyond this any further discussion is lacking. For example, to the uninitiated, what is the difference between the GLWD and non-GLWD extent maps? The underlying fluxes per unit area are presumably the same between the same models in GCPv1 and GCPv2. So what has really been learnt, other than the major control on the fit to the results of Z21 being the wetland extent map used?

The title and abstract indicate that the work provides constraint on the climate sensitivity of wetlands. I found this term a little confusing given its normal usage in climate sciences of the change in temperature given a change in radiative forcing. Perhaps, "sensitivity to climate forcings" might be a more appropriate term. Regardless, I did not really see where in the article this quantification was provided, e.g. the calculated change in CH4 emissions given a change in temperature or precipitation. The correlation analysis shows there is a relationship between all models (not just HP) and these climate drivers, but little on the form of that relationship. It seems to be a leap to claim that the work provides new constraint on this response of CH4 emissions to the underlying drivers. Narrowing down models into HP and LP seems to add little in this regard, precisely because the main difference appears to be the wetland extent map used, rather than the underlying process model dynamics.
Specific comments:
Figure 1: This figure is not very self-explanatory i.e in what way is WetCHARTs an ingredient for evaluating biogeochemical models? How is this different from the 42 biogeochemical models themselves? Either some more info in the workflow is needed or the caption needs expanding.

P6, L111: To what extent are the wetland emission estimates of Z21 negatively correlated with the separate anthropogenic estimates? Figure 6 of Z21 suggests this might be an issue in certain regions of the world.
P7, L131: "Both GCP version are chosen because..." The meaning of this line was little unclear. Do you mean you use two different GCP versions because they use different wetland extent maps?
P9, L175 Eq1: What is $\hat{x}_{model}$?
L194: Are the models evaluated globally with respect to the w-RMSE metric or independently on a region by region basis?
L203 Results: To be clear when analyzing the models are you analyzing the raw model output, or the observation operator transformed model output? I assume it is the former given the generally low averaging kernel values presented in the supplement. Unless N. Boreal America is a particularly poorly constrained region? The point being that if A is close to zero then the results will just be largely representative of the prior. It might be worth creating a table of mean averaging kernels for each region to show the constraint that the satellite data actually have on each region.
L. 206: 5th-95th percentiles? Do these ranges refer to the percentiles of monthly emission estimates or the 5th and 95th percentiles of the 14 model medians (if such a thing exists)?
L218: Saunois et al (2020) suggest that tropical wetlands are one of the more highly uncertain aspects of the wetland CH4 budget. Yet these results suggest there is very little variation in tropical emissions between HP and LP models. Is that right, or is the model range large and it is just coincidence that the medians are similar between HP and LP?
L220-221: Why are the Amazon emissions smaller in the HP models compared to LP?
L230: What do you mean by "match the annual CH4 emissions in the Alaska region"? Do you mean consistent within the uncertainties?
L230: Are the CARVE emissions representative of exclusively wetland sources or do they include influences from non-wetland sources?
Table 1: Wouldn’t it be worth including the Z21 posterior fluxes as well since this is the quantity you are calibrating your model selection on, rather than the prior?
P14. L255: Why are LPX-Bern and JULES the best fitting models? What is it about the model parameterizations of these models compared to other models? In addition, I assume this is a global comparison, but do these models perform best for the majority of regions or is the best-performing model very region dependent?
Figure 4: Can you swap the order of HP and LP in the key to be the same as in the rest of the figure? It is just a bit confusing otherwise.
P. 16, L299: "Our analysis reveals how wetlands across the globe respond to variations in temperature, precipitation, inundation, and carbon availability." Does it? I cannot see any new quantification of how wetland CH4 responds to changes in these variations. Would it not be more accurate to say the analysis narrows the range of a given set of wetland models that fit with a satellite-derived estimate of CH4 emissions. But the response to variations in climate variables are qualitatively the same regardless of whether the model is in the HP or LP group. The main difference seems to be the use of a GLWD vs non-GLWD wetland extent map. So it is unclear whether narrowing the models down has really added anything here.

Peer Review Comments on 2021AV000408RR

Reviewer #2

General comments

The authors have gone to some length to address the concerns raised in the first review and this version of the manuscript is much improved. However, I have a few remaining, mostly minor, concerns.

Text discussing the caveats of the method, which was said to have been added at the end of the introduction appears to be missing.

About the general applicability of the method to inversions using e.g. 4D-Var methods - the authors write that the method can be used for any inversion for which the prior and posterior error covariance matrices are available. However, the method requires the Averaging Kernel, which depends on the posterior and prior error covariance matrices (or an alternative formulation using the prior and observation error covariance matrices) but also the transport matrix. Could the authors please explain if they have derived another formulation of the Averaging Kernel which does not require the transport matrix? If not, then please state that the availability of the transport matrix is a requirement of this method, as is the prior and posterior error covariance matrices.

Specific comments

Keypoint 3: Please state what "25% less" is in reference to.

P2L27 (and P3L39): Suggest changing "CH4 budget" to "CH4 source" since the budget incorporates also atmospheric loss.

P2L28 (and P3L47): I suggest removing "refine" since you are not refining the bottom-up models in this study, just comparing them with "satellite-informed" flux estimates.

P5L82: Please change this sentence since they way it is written it implies that all
inversions only use space-borne observations, which is of course not the case.

P6L102: The method described in this study only accounts for random uncertainties in atmospheric inversions, and not systematic uncertainties, due for example to atmospheric transport errors. Therefore, please add "random" before "uncertainties".

P9L157: Could the authors please add a bit more of an explanation for "C respired as CH4". Do the authors mean simply the amount of CH4 produced by methanogens, and if so, how is temperature sensitivity of methanotrophy (microbial methane oxidation) accounted for or is this not included in this parameter? Also, for CO2 respiration, do the authors mean autotrophic or heterotrophic respiration (or both)? I think in general a bit more explanation the parameter Q10 CH4:C is needed.

P12L233: The authors write that larger delta_wRMSE values indicate a better model agreement with atmospheric estimates, w.r.t the model subset. However, I would think the opposite, i.e., smaller delta_wRMSE values would indicate better agreement, as values of wRMSE_MM smaller than wRMSE mean that the specific model process agrees better than the mean of the model subset. Could the authors please explain.

P15L278: Could the authors please explain what the "emission scaling factor" is.

P15L281 (and P16L315): The authors state: "atmospheric constraints indicate lower temperature sensitivity and [...] generally leads to better agreement with atmospheric CH4 measurements. However, the parameter compared is the Q10 CH4:C, i.e. the rate of C respired as CH4 versus CO2. So in fact the result is lower temperature sensitivity of CH4 relative to that of CO2. Please also see my comment concerning P9L157 about what Q10 CH4:C represents.

P18L364: The method requires the Averaging Kernel, which depends on the posterior and prior error covariance matrices (or an alternative formulation using the prior and observation error covariance matrices) but also the transport matrix (please the general comment about this). Therefore, I think this should also be added as a constraint for this method.

P18L369: I suggest changing this to "relative uncertainties larger than that of wetland fluxes" since overall the freshwater body emissions are significantly smaller than that of wetlands, and thus also the absolute uncertainty.

Fig. 4: Is there a correlation between wetland extent (GLOBCOVER versus GLWD) and global wetland emission, for instance, does GLOBCOVER have a smaller extent and thus models using this have a smaller global wetland emission?