Constraining biospheric carbon dioxide fluxes by combined top-down and bottom-up approaches

Samuel Upton, Markus Reichstein, Fabian Gans, Wouter Peters, Basil Kraft, and Ana Bastos

Response to reviewers

REVIEWER 1)

R1C1: Honestly speaking, I have conducted research for inverse modeling and eddy covariance observations for more than 10 years, but still feel this MS (egusphere-2023-805) by Upton et al. (2023) is very hard to follow and understand in recent version, including both the logical organizing for different paragraphs and detailed expressions in sentences. This MS is not a bottom-up and top-down based research paper, but a mathematic based way that made use of avalable dataset, and without enough details. Therefore, I do not recommend for publication unless large and significant improvement have been made to improve its reliability and robustness and readability.

Thank you for your frank critique of the manuscript. We realized that the presentation of the methods and results was not clear, so that the novelty and value of our approach did not come through. We also recognize that we did not sufficiently emphasize the central modeling structure of our study, a bottom-up data-driven flux model, constrained by top-down data. We have thoroughly revised the text to improve clarity, and include new conceptual diagrams (see below figures R2, R3), so that both our logic and methodology are more understandable to the reader. We do, however, not agree with the reviewer's point that our approach is not bottom-up and top-down based (see R1C2), nor with the suggested lack of reliability or robustness of our work as presented, as the review did not provide critical scientific argumentation on these aspects for us to respond to. Below, we provide a point-by-point reply addressing the reviewer's concerns and indicating whenever possible the main changes that will be introduced in the manuscript upon revision.

R1C2: First, although the authors combined “bottom-up” CO2 flux and “top-down” atmospheric inversion CO2 flux data set, from my first feeling when go through the introduction and methods sections, it reads like the authors conducted at least one of above approaches. However, after I read through results for quite a few times, it seems the authors only used some mathematic method (i.e. a region-specific sparse linear model) to make use of both datasets and made some adjustment. Hence, the mathematic method is the main advantage, but it has not been descripted in details how to use this method and its advantage when compared with above “bottom-up” and “top-down” methods.

In this study, our goal is to improve NEE estimates from a well-established "bottom-up" eddy-covariance upscale model (FLUXCOM, Jung et al., 2020) by additionally constraining its calculations by "top-down" estimates of NEE from atmospheric inversions. The reason, as described in the MS Lines 40-57, is that FLUXCOM is known to perform well in representing small-scale and sub-seasonal to seasonal variability in NEE, but has important biases in global NEE and limited skill in representing interannual variability (Jung et al. 2020). On the other hand, atmospheric inversions constrain well the global NEE and its interannual variability, but increase in uncertainty with decreasing spatial scale. Hence, we adapt the underlying FLUXCOM machine-learning model structure to optimize NEE by adding the atmospheric constraint

provided by inversions. Our study provides a first step towards a merged "bottom-up" and "top-down" model by addressing some of the technical and scientific challenges when trying to combine these distinct types of datasets (as noted by the reviewer).

As the reviewer points out below, there are issues of scale and model structure that make this effort non-trivial. To link the different spatial scales at which each approach is able to constrain NEE, we add region-specific sparse linear models to bridge across scales. This allows the underlying model to be directly compared with, and to learn from top-down data from atmospheric inversions. These "bridge models" are not the core of our approach (see R1C6 for further discussion), but rather a way to implement two cost-functions (loss terms) in the optimization of NEE (L_{EC} and L_{ATM}). The final model optimizes NEE by assimilating top-down and bottom-up data at the same time and allows evaluating whether a combination of top-down and bottom-up constraints allows resolving the limitations of each individual approach.

We do not understand the statement that "...it has not been descripted in detail how to use this method and its advantage…". We acknowledge that the methodology might have been difficult to understand at points, but Sections 3.1 through 3.7 documented many details, with Appendix B expanding even on that. Can the reviewer point out which details are missing specifically? We would appreciate the chance to add the needed information. As for the advantage, we would like to point to lines 15 through 20 and line 25 of the abstract, as well as to line 79 of the introduction, line 228, 267, and 298 of the Results, and 373 of the Conclusion.

To address the concern of the reviewer, we will however add a separate paragraph on the advantages and disadvantages of our new method. The paragraph now reads:

"Here, we show that adding regional atmospheric constraints to a bottom-up data-driven flux model influences NEE estimates from monthly to inter-annual and local to global scales. This approach minimizes the limitations of both top-down and bottom-up systems. It yields a model that preserves the local, small-scale view of the bottom-up approach while bringing the regional and globally integrated results inline with other best estimates of NEE. [...] The system presented in this study is still dependent on the priors and transport models in the underlying atmospheric inversions, and is still subject to the underlying uncertainty of these data, particularly in the tropics where both bottom-up and top-down systems lack sufficient observations."

Final Draft:

"Our results show that adding regional atmospheric constraints to a bottom-up data-driven flux model improves NEE estimates from monthly to inter-annual and local to global scales. This approach minimizes the limitations of both top-down and bottom-up systems. It yields a model that preserves the local, small-scale view of the bottom-up approach while bringing the regional and globally integrated results inline with other best estimates of NEE. […] The system presented in this study is still dependent on the priors and transport models in the underlying atmospheric inversions, and is still subject to the underlying uncertainty of these data, particularly in the tropics where both bottom-up and top-down systems lack sufficient observations."

R1C3: Lines 15-21, "The inferred global terrestrial carbon flux from land, excluding fires and riverine evasion across 2001-2017 is -3.14±1.75 PgC $year^{-1}$ (±1 σ). This is a strong improvement over the -

20.28±1.75 PgC year -1 from the exact same data-driven flux model trained without the additional regional top-down constraint (i.e., single constraint) when compared to current best estimates of the global carbon flux from land.", the authors said this result is considerably different with another result, and why the readers can believe your results is robust as you said "is a strong improvement"? there are many previous studies that calculated global NEE, how does your results compared with theirs?

Multiple studies have shown that while there is still uncertainty on the absolute magnitude and trends of the land CO2 sink, its magnitude is of a few PgC/yr (Friedlingstein et al., 2022; Ruehr et al., 2023; Crisp et al., 2022). In this study, we chose the estimate of the residual land sink from the Global Carbon Project (Friedlingstein et al. 2022) for comparison in the main text, as this estimate encompasses many different methodologies (modeling, bookkeeping, measurement-based) in its determination of global NEE. It is thus a synthesis of most recent insights on what that number should be. The residual land sink also handles fire emissions, partially integrated in the land-use term, in a way that is coherent with our inversion system targets. We find that the abstract is not the place to elaborate further on this, but we will explain this more clearly in the results as follows:

"*As shown in Figure 1, the addition of an atmospheric constraint to a data-driven flux model (i.e. the EC-ATM model) leads to a global estimate of NEE (-3.14±1.75 PgC year^{-1} (±1 σ) that is closer to the current best estimates from GCB22 (-2.99±0.9 PgC year^{-1} (±1 σ) than the model only based on eddy-covariance flux data (EC model, -20.28±1.75 PgC year^{-1}). The GCB22 estimate represents an independent view of the sink magnitude as it is calculated from as the residual of other major independent terms in the global carbon budget (emissions from fossil fuels and industry (E_{FF}) + emissions from land-use change (E_{LUC}) - the ocean sink (S_{ocean}) - atmospheric growth rate (G_{ATM})), which are otherwise not used in our approach. Moreover, it agrees well with estimates of the land sink from other recent studies ({Here Table 1}). The EC-ATM annual mean NEE has an RMSE of 0.91 (compared to GCB22) for the period 2001-2017, compared to, with an RMSE of 17.32 for the EC model.*"

Final Draft:

We first compare the global NEE estimates from the EC and the EC-ATM models (Fig. 2) with the residual estimate of the land sink from the Global Carbon Budget (GCB22) (Friedlingstein et al., 2022). The addition of an atmospheric constraint to a data-driven flux model (i.e. the EC-ATM model) leads to a global estimate of NEE (-3.14±1.75 PgC year-1 (±1 σ)) that is closer to the current best estimates from GCB22 (-2.99±0.9 PgC year-1 (±1 σ)) than the model only based on eddy-covariance flux data (EC model, -0.28±1.75 PgC year-1). The GCB22 estimate represents an independent view of the sink magnitude as it is calculated from as the residual of other major independent terms in the global carbon budget ($E_{FF} + E_{LUC} - S_{ocean} - G_{ATM}$) which are not used in our approach. Moreover, it agrees well with estimates of the land sink from other recent studies (Table C1). The EC-ATM annual mean NEE has an RMSE of 0.91 PgC year-1(compared to GCB22) for the period 2001-2017, compared to an RMSE of 17.32 PgC year-1 for the EC model.

Nevertheless, we have revised the abstract, as follows:

"*Compared to reference estimates of the global land sink from the literature, e.g. Global Carbon Budgets, our double-constraint inferred global NEE shows a considerably smaller bias in global and tropical NEE compared to bottom-up data driven model estimates (i.e., single constraint). The mean seasonality of our double-constraint inferred global NEE is also more consistent with the Global Carbon*

Budget and atmospheric inversions. At the same time, our model allows for more robustly spatially resolved NEE and smaller errors at local scale, when compared to atmospheric inversions alone."

Final Draft:

Compared to reference estimates of the global land sink from the literature, e.g. Global Carbon Budgets, our double-constraint inferred global NEE shows a considerably smaller bias in global and tropical NEE compared to the underlying bottom-up data-driven model estimates (i.e., single constraint). The mean seasonality of our double-constraint inferred global NEE is also more consistent with the Global Carbon Budget and atmospheric inversions. At the same time, our model allows for more robustly spatially resolved NEE. The improved performance of the double-constraint model across spatial and temporal scales demonstrates the potential for adding top-down constraint to a bottom-up data-driven flux model.

Furthermore, we now add the GCB22 estimate of the land sink in Figure 2, which will be updated as shown in Figure R1.

By comparing the outcome of our system with/without additional constraint we demonstrate that our method can overcome known problems with the bottom up flux model. We feel that our choice for this benchmark is easily defensible, and we will make this choice more clear in the revised manuscript.

We note, however, that a thorough assessment of the performance of our approach compared to both large-scale NEE from inversions and in-situ observations from eddy-covariance towers can be found throughout the submitted manuscript, namely Table 2, Figures 3-5, 7-9, C1 and C2.

R1C4: Line 60, please give us the extent of recent studies who calculated global NEE here, including both top-down and bottom-up approaches.

To accommodate the reviewer, we will add a summary of the recent studies in the revised manuscript, reproduced below.

Table T1

Study	Land sink mean with fire (Pg C yr−1)	IAV magnitude	Period
Crisp et al., 2022	−2.2	0.6	2000-2009

Friedlingstein et al., 2022	-2.86 S_land -2.99 Residual	0.81 S_land 0.9 Residual	2001-2017
Gaubert et al., 2019	Northern ExtraTropics: −2.17 Tropics+Southern ExtraTropics: −0.06	Northern ExtraTropics: 0.36 Tropics+Southern ExtraTropics: 0.11	2004–2014
Ruehr et al., 2023	-3.1	0.6	2010-2019

Table 1: Summary of estimates of NEE from recent studies

R1C5: Line 70-85, after reading this part, which aim to describe the method the authors used in this MS, I am still confused why only a few sit-scale based regressions can be used to represent whole regional scale? As the authors said this is based on atmospheric inversion results, and the available global NEE products as Carbon Tracker provide NEE at more than 1o, how does this coarse spatial resolution be used with FLUXNET and eddy covariance tower at site scale (i.e. 100m*100m)?

The reviewer's remark points out the exact reason why the proposed "bridge models" are needed to develop a model that optimizes NEE at the same time using site-scale data from eddy covariance measurements and large-scale atmospheric constraints.

We assume that the reviewer's reference to "site-scale based regressions" relates to the bridge model that connects local NEE to regional integrals. The method to connect the coarse spatial resolution of CarbonTracker to the eddy-covariance scale is explained in Section 3.4 (not the introduction) and in R1C2: we connect the two constraints at regional scales (and not at 1 degree nor at 100m resolution). To do this in a computationally efficient manner, we need to find a small enough but representative enough number of grid cells that can be used to infer regional NEE and estimate L2. See Figures B3 and B4 for an assessment of land-cover composition of the contributing pixels. These sparse regional models represent five orders of magnitude fewer forward and backward passes through the machine-learning model compared to a naive approach calculating the per-pixel NEE over the whole land surface for each timestep executed in the training loop. This reduction provides a computationally efficient method to integrate bottom-up fluxes regionally, while the regional scale NEE is robustly represented by the atmospheric inverse models, which is the basis for our innovations.

Sections 3.3 - 3.6 of the manuscript will be thoroughly revised to improve the clarity of our methods. Additionally, we will replace figure 1 with the figure R3, shown below.

R1C6: The main assumption in this MS is "The central hypothesis of this study is that individually trained regional sparse linear models can serve this function", how to certify your hypothesis is robust from site to local and regional scales?

The description of the full central hypothesis continues on to describe how these regional sparse models "enable a complementary constraint based on (top-down) atmospheric inversions to the (bottom-up) data-driven flux model, improving the estimates of regional and global land surface CO2 fluxes". The

complementary constraint is actually the core innovation of our study, where the bridge models help improve computing efficiency (see R1C5). We acknowledge that this was not clear in the original text, and the revised manuscript will make the central motivation and hypothesis more easily understood.

Our full hypothesis is demonstrated by our Results section in 3 ways:

(1) Lower RMSE between the regional integrals for the EC-ATM NEE and both the magnitude and seasonality and variability of the inverse model results, demonstrating the transfer of information and propagation to the regional scale (Figures 2, 3, C1, Tables 2,3)

(2) is the improved correspondence to the independent GCB22 results (see R1C3), demonstrating the robust transfer to global scales. We realize that this information was missing beyond the abstract and it will be added in the revised manuscript.

(3) The continued, or even improved, correspondence to eddy-covariance observations, demonstrating the robust back-propagation of information from the regional to the local scale (Figures 8, 9, C2).

We make no further claims regarding the regional models beyond this efficiency (also, Figure 4), and the proposed statistical relationship on which they are based. These statistical models are based on a single dataset derived from the underlying "bottom-up" model used here (FLUXCOM RS+MET). The results show that the spatial distribution of NEE in the globally upscaled FLUXCOM product is sufficiently spatially stable over time to allow for a sparse solution.

The sentence regarding our hypothesis will be revised accordingly to:

"*Here, we aim to test the hypothesis, [...], that a bottom-up data-driven flux model trained with a complementary constraint based on top-down atmospheric inversions can improve the estimates of regional and global land surface CO2 fluxes.*"

R1C7: Line 92 "At tower locations the meteorological data is derived from the FLUXNET and ERA5 data,", the highest spatial resolution of ERA5 is 0.25o (~25 km), and the EC tower site is point scale, so how to evaluate the footprint mismatch between them. And for the section of 2.1, what data is produced by the authors and what is produced by the references? I am still confused here what the authors' contribution.

It is not within the scope of this study to evaluate the performance mismatch between the eddy-covariance scale and the grid-cell, as this approach is well documented in the FLUXCOM approach (Jung et al., 2020) which is consistent with the gap-filling approach in FLUXNET2015 (Tramontana et al 2016) on which this study is based. The eddy-covariance data is from the La Thuile synthesis data set and CarboAfrica network (Valentini et al., 2014). Here, we use the publicly available NEE estimates from FLUXCOM RS+MET (Jung et al., 2020), and guarantee that the EC model follows exactly the same set up (structure, predictors) as described in (Jung et al. 2020) for consistency. To improve clarity, we have reformulated the description of the modeling approach and of the use of FLUXCOM v1 in revised Section 3 (Statistical bridge-models) as follows:

"*The dependent variable for the Lasso regression is the regional integral of the monthly ensemble mean NEE derived from the global NEE data from the FLUXCOM intercomparison, specifically the RS+METEO setup (Jung et al., 2020). […] [T]he spatial logic of the regression is to find a stable set of*

weights relating the NEE at geographic locations within the region to the regional integral of NEE. This logic is then extended using Lasso regression (Eq. 3), where the reduction of weights in the model to zero creates a sparse solution. This can be interpreted as discovering a subset of geographic locations whose NEE are minimally sufficient to produce a high-quality estimate of the regional NEE from the FLUXCOM RS+METEO results. This quality is quantified as the R2 value of the NEE inferred by the sparse Lasso model regressed against the regional integrals of NEE. The FLUXCOM RS+METEO data are thereafter only used for model evaluation."

R1C8: Section 2.2, the atmospheric inversion community have found the flux results by different assimilation systems largely varied by both global total NEE, and its magnitude, especially for spatial distributions, which are caused by difference in atmospheric transport models observed concentration, how to choose which is reliable in your research?

Indeed, we agree with the reviewer that differences between atmospheric inversion systems based on their priors and atmospheric transport are important to consider. This is precisely why we used an ensemble of inversion systems in parallel during training, rather than a single one, so that the objective function considers values for all inversion members at each step. We hope that this exposes our model to the range of estimates in the atmospheric data, and allows the model to learn from the full spread of data without an explicit a priori choice of the 'correct' model. We will improve our description of the approach in section 2.2, along with the methods generally, to make this logic more clear to readers.

R1C9: Line 111, "These variables are were computed" delete were

Done

R1C10: Line 113. "and are reconstructed here by the on the percentage of the component PFTs", it's hard to follow.

The underlying driver variables from Jung 2020 and Tramontana 2016 use these same data, and they are computed and stored as a global gridded dataset according to the majority plant functional type (PFT) of the land pixel. We will improve the language to make this more understandable. The sentence now reads:

"*… and are reconstructed here by the percentage of the component PFTs in each pixel following the approach in Jung et al. (2020) and Tramontana et al. (2016)*"

R1C11: In Figure 2, what's the reason for the large difference between red, blue lines and green, yellow lines, emissions from rivers as you stated?

We apologize that this was not clear in the manuscript. The differences between the lines are the core finding of the study and we will improve our text to make this more concrete. The yellow line is the previous FLUXCOM RS+MET result published in Jung et al. (2020), which strongly over-estimates the tropical sink and shows a global integral that is not close to GCB22's best estimate of NEE. The red line is the ensemble

mean of the atmospheric inversions integrated globally over land. Since these inverse estimates feed into, and are extensively benchmarked in, the GCB22 estimate we use this as a reasonable estimate of the 'true' value of NEE. We realized that the GCB22 estimates were missing from the manuscript, which might have been the cause for confusion. We will now add these in the revised version. The green line shows the EC model, which is trained using only eddy-covariance data and corresponds, therefore, to an additional member of the FLUXCOM RS+METEO ensemble. The purpose of this comparison is to show that our model (with small revisions in the predictors) reproduces the NEE from the FLUXCOM v1 dataset from Jung et al. (2020), based on which we trained the "bridge" models. The blue line corresponds to the EC-ATM model which is trained with eddy-covariance and atmospheric data. The central result of the study is the difference between the blue and green lines, indicating the learned difference in global NEE when the eddy-covariance model is enriched with atmospheric data.

Below, you can find an updated version of the figure with GCB22 constraints based on the residual land sink (Fig. R1).

R1C12: Section 5.1, I am very confused why the authors believe atmospheric information (inversion flux) is reasonable at grid scale? Large bias exist even for large regional scales as whole continent or country.

We agree that grid level data from the atmospheric inversions is not robust, and in this manuscript we do not state anywhere that we believe this flux is reasonable, as stated by the reviewer. In fact, we explicitly avoided the use of gridded data and in training we used the data only in regional integrals and only as an ensemble (see R1C2 and R1C5). Section 5.1 refers to Figure C1, which compares regionally aggregated NEE, precisely for the reason mentioned by the reviewer. Note that we choose to compare maps of NEE based on EC and EC-ATM with inversions (Figures 5, 8), so that we can show that conversely, combining a "bottom-up" approach with the "top-down" atmospheric inversions allows to infer NEE with more finely resolved spatial variability in NEE. This was intended to demonstrate that the model is not over-learning the unreliable spatial distribution of the atmospheric inversions. We will revise the text in Section 4.3 to make these decisions more clear.

R1C13:I have not provided coments on the spatial-temporal patterns of NEE, which seems not robust before the authors add more descriptions to verify their method.

Please see replies to R1C2, R1C5 and R1C12, which we believe address the concerns about the reliability and robustness of our method. We will improve the manuscript accordingly and are happy to provide any further proof of robustness if the reviewer provides concrete suggestions.

Reviewer 2)

R2C1: The paper addresses the quantification of Net Ecosystem Exchange (NEE) from multiple datasets. The concept is a multi-scale approach to estimate NEE fluxes using an additional atmospheric "top-down" constraint to further extrapolate the bottom-up data-driven flux model, originally derived from eddy covariance sites. Overall, the paper is well written, and mostly need to educate more the readers about the author's innovative framework.

We thank the reviewer for their feedback. We appreciate the need for additional work on the manuscript and will provide a more complete description of the logic and process that better explains the study in the revised version.

R2C2: The proposed method seems to provide a substantial contribution to the field by attempting to reach reliable continental scale carbon dioxide flux estimates. However, in order to reach such conclusions, the authors should outline better and more clearly the steps regarding the EC-ATM optimization constraints. This could be done simply by presenting more illustrations of the technical implementation such as, for example the one shown in the appendix.

We agree that the description of our method was not clear enough (see replies to R1). Therefore, we will include new figures showing the diagrams describing the two models (see Figures R2, and R3) that better show the spatial logic and processing of data through the system and the optimization approach. We hope that these three figures, and corresponding description in the revised manuscript, will make the system more transparent to the reader.

R2C3: Specifically, it looks like the various weights employed in the objective functions and parameters such as the region-specific parameter are playing an important role in the optimization and extrapolation from local NEE to RECCAP regions. The considerable change in the global flux suggests that the top-down inversion's weights will drive the flux estimation at regional scale. It seems important to quantify such effects to further appreciate the multi-scale flux estimate and its uncertainties at the RECCAP-2 level. This is illustrated in Fig. C1 where the EC-ATM can diverge from both the inversion mean and the EC model.

The weighting of the terms in the objective function are a learned parameter of the model, and it is difficult to directly quantify the effect as the model evolves during training. Figure R4 tracks the progression of the weighting terms for the EC-ATM model. However, the reviewer's intuition is correct, the weight of the atmospheric objective term does drive the regional flux estimation. The lower bound of this is the EC ensemble, which can be understood as the model with the relative weight of the atmospheric term reduced to zero. Although we do not undertake the analysis here, it is likely that a model trained with a zero-weighted eddy-covariance term (the other extreme) would follow the mean of the atmospheric inversions as closely as the free parameters of the model allowed. Given the importance of this term to the overall model performance, we wanted to avoid an arbitrary weighting scheme. This is why we opted for a learned weighting scheme (Kendall et al., 2018) to make this important model parameter dependent on the optimization rather than an expert assertion.

Again as the reviewer notes, the addition of atmospheric inversion data to the training loop (i.e. the new loss term L_{ATM}) has a very strong impact on the final estimate of NEE, particularly at regional and global scales (which was our hope), but is also propagated to smaller scales. Inversions show varying agreement in NEE over space and time (see Friedlingstein et al., 2022), therefore we consider their spread as a measure of NEE uncertainty at each given point in space (region integral) and time (monthly values). The individual inversion-based NEE values for a given region and month are used to estimate individual loss terms before they are aggregated to $L_{ATM,r}$ as described in Eqs 6 and 7, so that the final loss term incorporates information

about inversion uncertainty. Indeed we were initially surprised that in some regions the EC-ATM NEE departs from both the inversion and EC model. This results from the fact that the final atmospheric loss term (L_{ATM}) is a regionally-weighted sum of loss terms, so that in the training, regional NEE is optimized based on a globally aggregated metric. We now add a more in-depth discussion of Figure C1 in Section 5.1.

To address the reviewer's comment, we propose to add supplementary material with additional model tests where a version of EC-ATM is optimized with each individual inversion, and a set of partial ensembles with two and three members, so that we can attempt to quantify the value of using the inversion ensemble, rather than individual ones (Figures R5-R7). These figures show how much the EC-ATM model is driven by the atmospheric term on regional scale. The regional responses can vary widely across inversions, and in balancing terms with the eddy-covariance objective term, the optimized models can show meaningfully different mean season cycles from the inversion systems used as their constraint. As additional inversions are added, there is a tendency for EC-ATM NEE to move closer to the inversion mean, as the target for optimization contains a larger amount of NEE values from the ensemble of inversions.

We now include a thorough discussion of this sensitivity analysis, as follows:

"We perform a "one-against-many" sensitivity analysis, where we trained several models with the atmospheric constraint coming from either one inversion (zero spread), two randomly selected inversions, or three randomly selected inversions. This analysis, shown in Supplement (Figs. C3, C4, C5) allows us to evaluate how the specific inversion NEE estimates interact with the loss mixing scheme in the model training. In extra-tropical regions where the eddy-covariance and atmospheric observations are dense, the specific inversion NEE trained on does not critically influence the model response. There is strong correspondence between all EC-ATM instances with individual or smaller groups of inversions and the full inversion ensemble mean. In tropical regions, there is definite movement towards the specific inversion NEE trained on (the dotted lines in the Figs. C3, C4, C4). This response is balanced against the model initialization and the learned weighting scheme during training. At the global level there is a closer agreement between the trained models and the target inversions, as the relative noise of the tropical regions is dampened by the more consistent extratropical regions. As additional inversions are added, there is a tendency for EC-ATM NEE to move closer to the inversion mean, as the target for optimization contains a larger amount of NEE values from the ensemble of inversions. This analysis demonstrates that the EC-ATM model inherits the uncertainty from the ensemble of atmospheric inversions, with largest uncertainty remaining in the tropical regions where the available observations for both top-down and bottom-up is lacking."

R2C4: The results seem to be overconfident in regions such as the tropics where there are much fewer observations to drive both the bottom-up and the top-down datasets. This suggests that the uncertainties in top-down inversions is underestimated. I apologized if I missed something and I did not understand fully the evaluation

We agree that the EC-ATM result is probably overconfident in the tropics. We tried to account for the uncertainties in two ways; the error between the EC-ATM result and the inversion data is normalized in the loss term by the spread of the inversion estimates by time and region, and the model learns to weight the objective term to account for the task-dependent uncertainty in the atmospheric loss term, which should tend

to reduce the weight where there is larger systematic disagreement between inversion systems. We will add a discussion on this in the revised manuscript as follows:

"Given the small number of tropical observations considered in top-down constraints, and the inherited uncertainty from priors and transport models, which are not directly managed by the system in this study (Baker et al., 2006), the EC-ATM model result might still be prone to large uncertainties in the tropics. Here, we discuss the handling of uncertainty in the EC-ATM model in more detail.

During training, the EC-ATM model tries to account for uncertainties in two ways: the error between the inference and the inversion data is normalized in the loss 2 term by the spread of the inversion estimates by time and region, which reflects uncertainty across inversion models (Eq. 10); additionally, the model learns to weight the objective term relative to the estimated uncertainty in the atmospheric loss term, which should tend to reduce the weight where there is larger systematic disagreement between inversion systems (Eq. 12)."

Final Draft:

The sparsity of observations does not only concern the FLUXNET network, but also the network of atmospheric monitoring sites used by atmospheric inversions. Given the small number of tropical observations considered in top-down constraints, and the inherited uncertainty from priors and transport models, which are not directly managed by the system in this study (Baker et al., 2006), the EC-ATM model result might still be prone to large uncertainties in the tropics. Here, we discuss the handling of uncertainty in the EC-ATM model in more detail.

The system presented in this study is still dependent on the priors and transport models in the underlying atmospheric inversions, and is still subject to the underlying uncertainty of these data, particularly in the tropics where both bottom-up and top-down systems lack sufficient observations. During training, the EC-ATM model tries to account for uncertainties in two ways: the error between the inference and the inversion data is normalized in the loss 2 term by the spread of the inversion estimates by time and region, which reflects uncertainty across inversion models (Eq. 10); additionally, the model learns to weight the objective term relative to the estimated uncertainty in the atmospheric loss term, which should tend to reduce the weight where there is larger systematic disagreement between inversion systems (Eq. 12)

R2C5: There must be a way to propagate uncertainties to estimate the combined errors associated with the EC-ATM.

While quantifying the errors associated with EC-ATM is not straightforward, we propose to include supplementary material comparing EC-ATM runs with individual inversions vs. the ensemble, so that the effect of uncertainty from the ensemble spread can be appreciated (see R2C4).

Minor comments:

Section 3.3: Can you expand about how inversions are used here, is the model is trained using 16 out of the 18 years? Does it mean only the ensemble mean is considered?

The regional integrals of the inversions are considered individually at each time step within those 16 out of 18 years. In the objective function, for each timestep, the neural network produces an estimate which is compared with the inversion ensemble values and then normalized by the spread of the inversions for that

timestep. The inversion regional and global means are only used for the figures to improve the legibility. The approach is later described in more detail in Section 3.4-3.6, which will be revised for clarity. We hope that Figures R2, and R3 allow for a clearer understanding about the use of inversions in our model.

The shorten word Fig has to be spelled with a dot as Fig.

This will be fixed.

P2L43: Gaubert et al., showed improvements at the scale of the latitudinal distribution of fluxes, not at the scale of continental-sized regions.

Thank you for this correction, the revised text will improve the language around our use of the inversions and the confidence we have in them as a constraint for our model.

"*While pixel-level NEE estimated by atmospheric inversions are known to be underconstrained (Ciais et al., 2010; Kaminski and Heimann, 2001), and are unlikely to provide robust constraints of NEE to train our model, atmospheric inversions produce reliable estimates of the magnitude and variability of latitudinal distribution of NEE, finding solutions that are in line with the global atmospheric growth rate (Gaubert et al., 2019). Therefore, we aggregate the NEE from the inversions by a set of 18 very large regions consistent with the regions in the Regional Carbon Cycle Assessment and Processes-2 (RECCAP2) project (Tian et al., 2018). This aggregation leverages the growing consensus about the magnitude of global NEE from atmospheric inversions (Gaubert et al., 2019) allowing for a 'global' constraint from a set of smaller regional constraints which cover the land surface.*"

P4L152: "from the as an independent data to test" There are two article *the* and *an*.

Done

P8L198: The accent ^ should be on the m (not on the r).

Thanks. Note that our notation has been revised and we corrected the typo.

P9L214: "are are" repeated word

Done.

P9L235: "The EC-ATM ensemble mean preserves the correlation with the scaled anomalies, producing very similar results to the FLUXCOM RS+METEO results 3."The sentence is not clear, what is result 3?Maybe it is Table 3, I cannot find mentions of Table 3 in the text.

We apologize, it should have read Figure 2B, which shows the detrended anomalies of the different annual estimates. The point of the sentence is that the annual integral from the EC-ATM results agree with the sign, if not the magnitude of the FLUXCOM estimate. The sentence will read:

"The EC-ATM annual mean is strongly correlated with the detrended annual anomalies of the atmospheric inversion ensemble mean (R^2 of 0.67), producing very similar results to the FLUXCOM RS+METEO results (Fig. [2]B) "

P19L384: xCO2, the x is usually upper case. Maybe you could spell out what are XCO2.

The text will be corrected to read:

"In the future, this logic could also be extended to other datasets, for example by pairing the archive of eddy-covariance observations with novel 'flux towers in the sky' (Schimel et al., 2019) estimates from satellite retrievals of total column CO2 (XCO2)"

FIGURES:

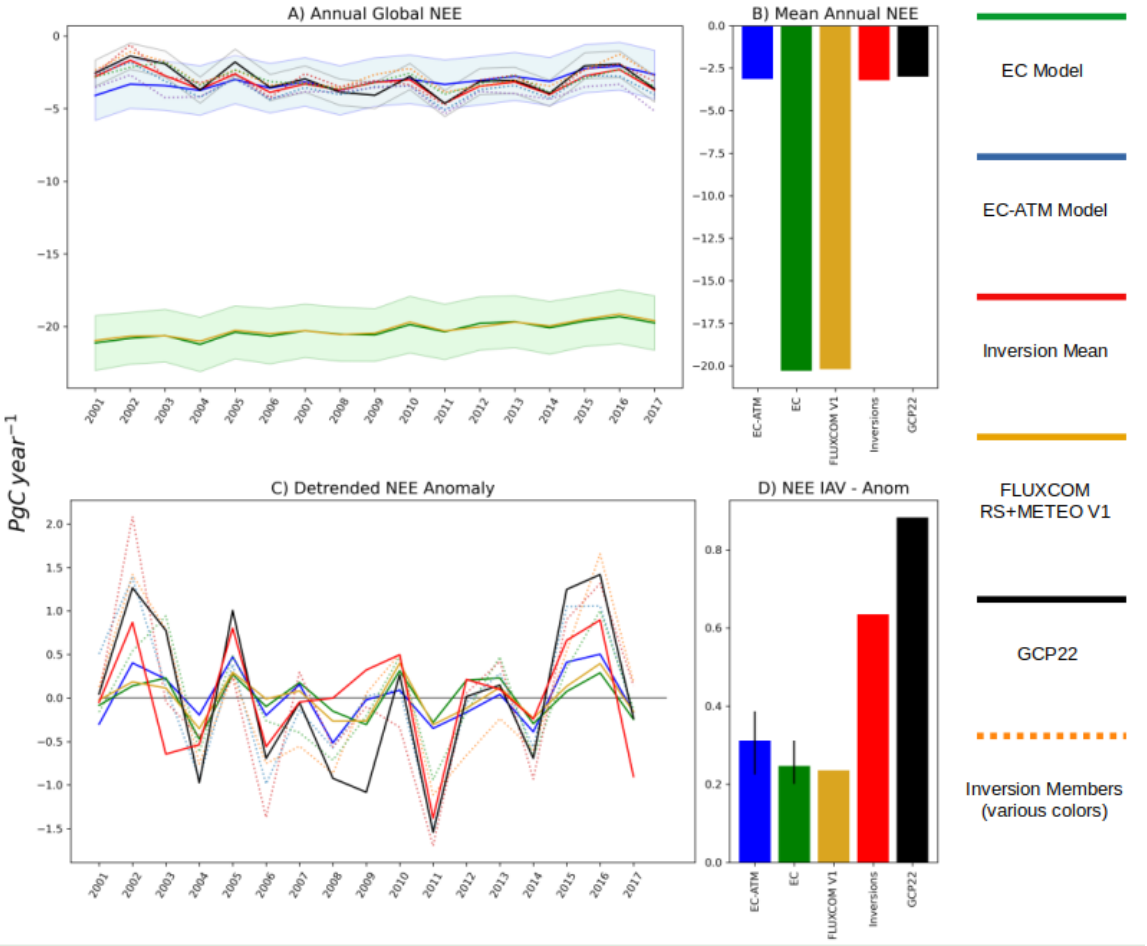

Figure R1: Panel (A) is the global annual NEE in PgC year-1. The blue line is the EC-ATM ensemble mean across the 10 folds The blue shaded area is the EC-ATM uncertainty across the ensemble (±1σ). The green line is the EC ensemble mean across the 10 folds. The green shaded area is the EC-ATM uncertainty across the ensemble (±1σ). The red line is the ensemble mean of the atmospheric inversions. The individual inversions are shown in dotted lines. The black line is the GCB22 residual land sink. The grey shaded area is the published GCB22 uncertainly. The yellow line is the ensemble mean of FLUXCOM RS+METEO V1. (B) is the annual mean across 2001-2017. (C) shows the detrended anomalies in PgC year-1. (D) shows the magnitude of IAV, estimated as the standard deviation of detrended annual anomalies. The bars on the EC and EC-ATM bars indicate the spread of IAV across the 10 folds.

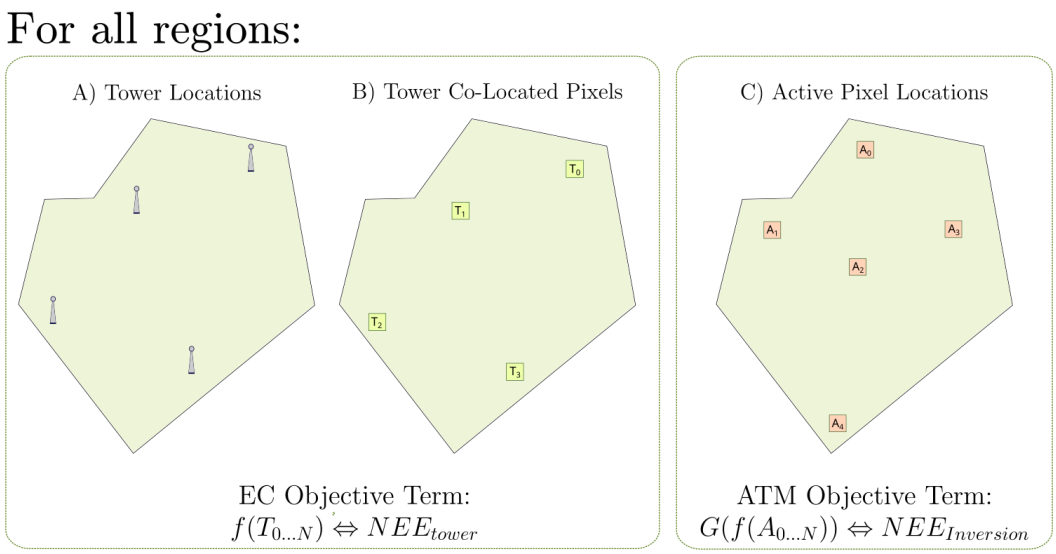

Figure R2: Spatial logic of the objective function terms: At each training step, for all regions, the neural network f(x) is run for a set of global eddy-covariance tower locations [T0,...TN] using meteorological (MET) data from the tower, and remotely sensed (RS) data from co-located pixels, which are compared with the observed NEE at the towers creating the first objective term. For each region, the neural network is also run for the region contributing pixels from the sparse linear model [A0,...AN]. Then, the specific regional linear model G() is used to generate a regional estimate of the integral, which is compared with the inversion ensemble. The results from all regions create the second objective term.

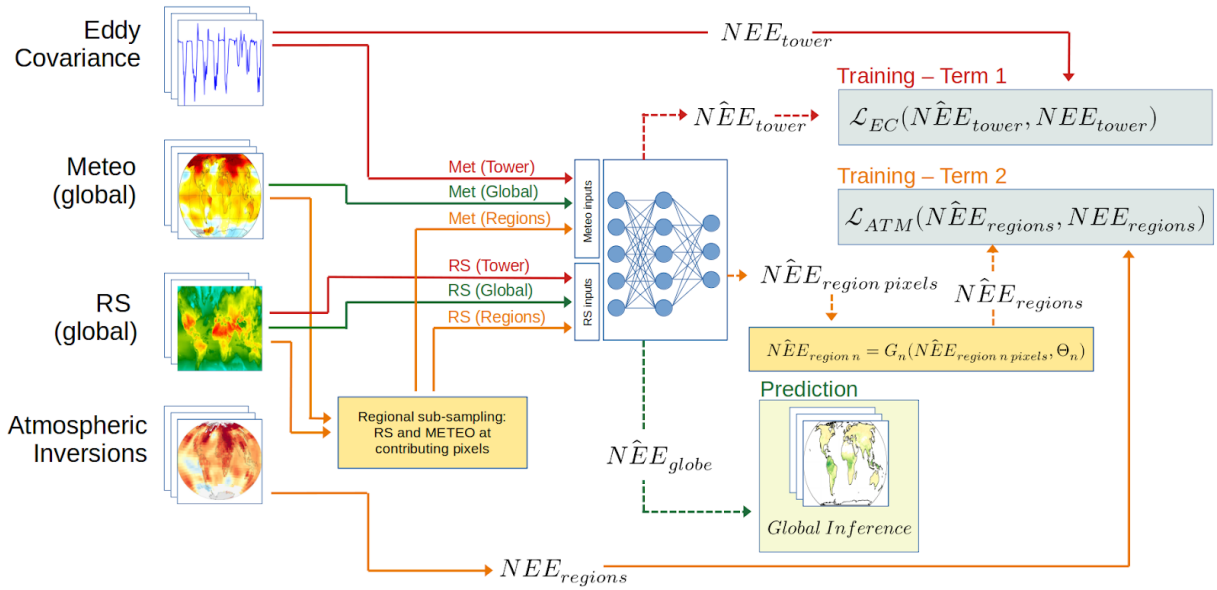

Figure R3: Flow diagram for the EC-ATM model. Red lines show the movement of data during training which creates the eddy-covariance based term in the objective function. The orange lines show the movement of data which creates the atmospheric term in the objective function. The yellow boxes show the function of the sparse regional models; selecting pixels for calculation and generating regional inferences from regional pixel NEE values. The green lines show the movement of data for a global inference.

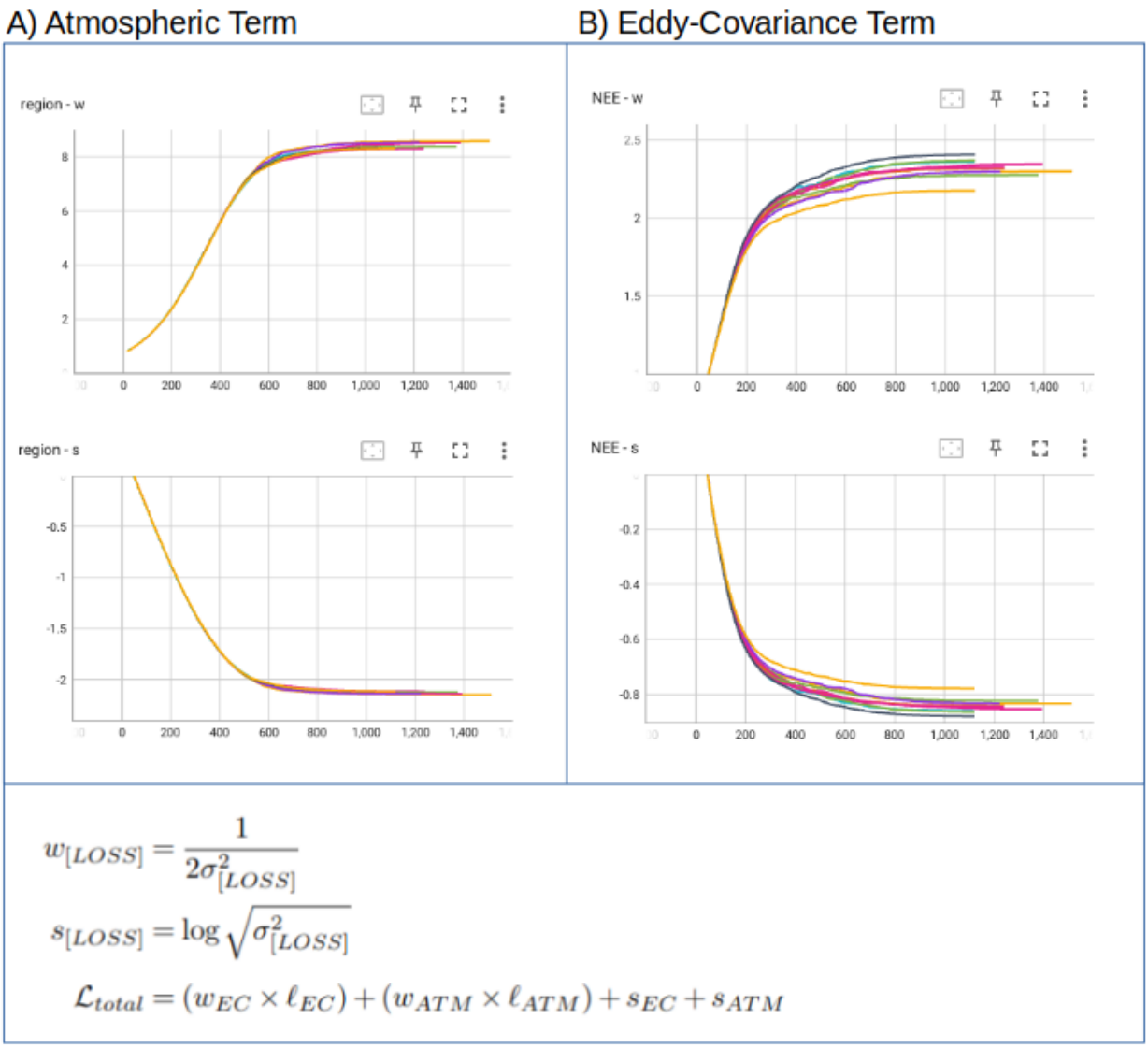

Figure R4: The evolution during training over the 10 folds of the learned weighting terms in the objective function. The weighting scheme, shown in the bottom panel (Kendall et al., 2018) creates a weighted loss according to two parameters that the model learns during training. Panel A shows the two learned parameters for the atmospheric term, and panel B shows the two learned parameters for the eddy-covariance term. The two terms ‘w’ and ‘s’ are defined in the bottom panel, but are not easy to interpret independently. Taken together the evolution of ‘w’ and ‘s’ shows a consistent learned estimate of the variance of the two objective terms over all data folds leading to a consistent weighting scheme derived solely from the data.

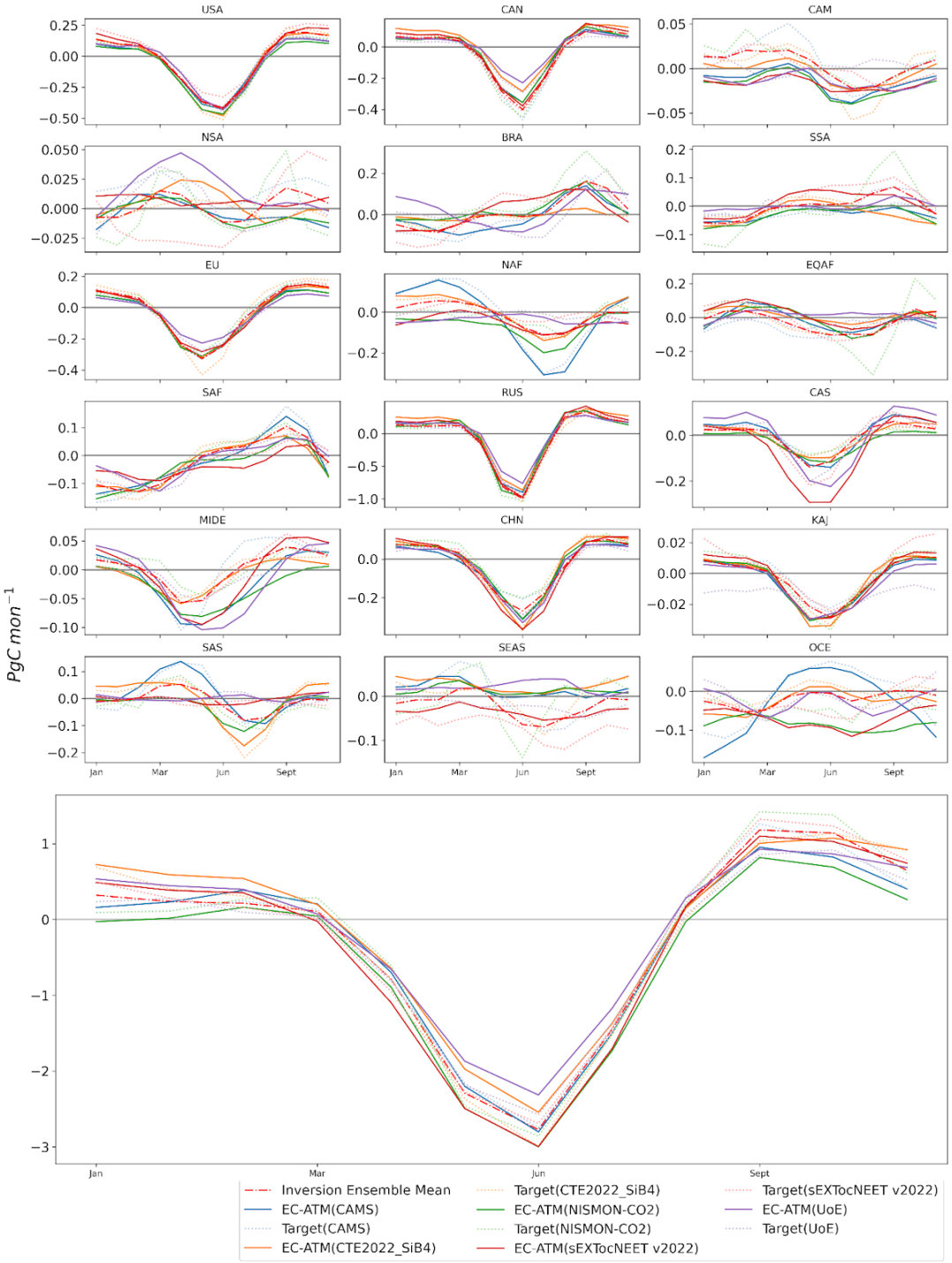

Figure R5: The MSC results for the ‘one-against-many’ training runs. A separate EC-ATM model was trained for each individual inversion system to test the impact of the full ensemble and loss normalization in the full study. The solid lines are the EC-ATM models trained using the named inversion system, and the dotted lines are MSC of the inversion system regional and global integrals. For all panels, the x-axis is the yearly cycle, and the y axis is PgC mon-1.

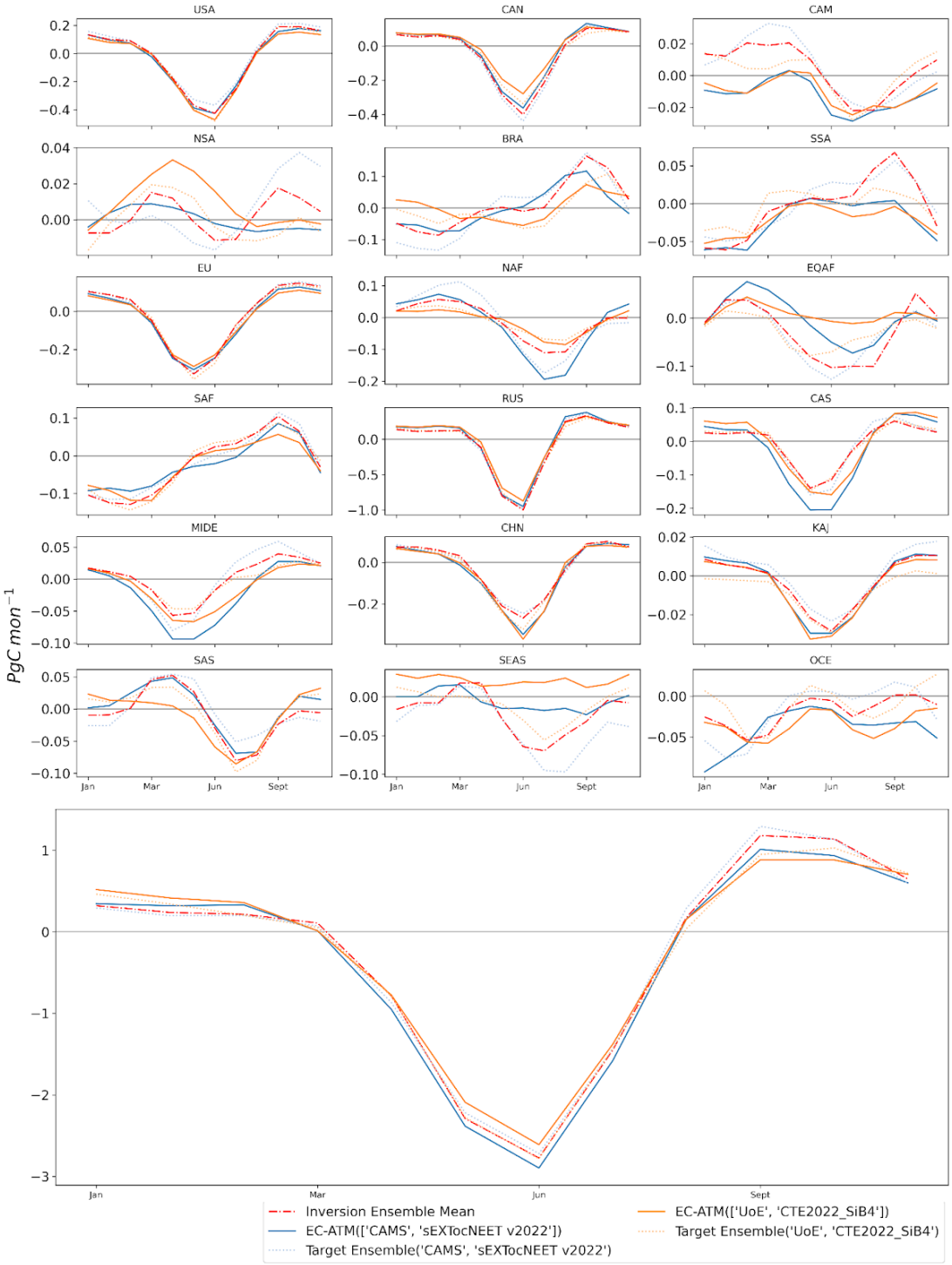

Figure R6: The MSC results for the ‘one-against-many’ training runs, with EC-ATM models optimized against two inversion systems. The solid lines are the EC-ATM models trained using the named pair of inversion systems, and the dotted lines are MSC of the named pair’s mean regional and global integrals. For all panels, the x-axis is the yearly cycle, and the y axis is PgC mon-1.

References:

Baker, D. F., Law, R. M., Gurney, K. R., Rayner, P., Peylin, P., Denning, A. S., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fung, I. Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Masarie, K., Prather, M., Pak, B., Taguchi, S., and Zhu, Z.: TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO2 fluxes, 1988–2003, Global Biogeochemical Cycles, 20, https://doi.org/10.1029/2004GB002439, 2006.

Crisp, D., Dolman, H., Tanhua, T., McKinley, G. A., Hauck, J., Bastos, A., Sitch, S., Eggleston, S., and Aich, V.: How Well Do We Understand the Land-Ocean-Atmosphere Carbon Cycle?, Reviews of Geophysics, 60, e2021RG000736, https://doi.org/10.1029/2021RG000736, 2022.

Ciais, P., Rayner, P., Chevallier, F., Bousquet, P., Logan, M., Peylin, P., and Ramonet, M.: Atmospheric inversions for estimating CO2 fluxes: methods and perspectives, Climatic Change, 103, 69–92, https://doi.org/10.1007/s10584-010-9909-3, 2010.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., Le Quéré, C., Luijkx, I. T., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Alkama, R., Arneth, A., Arora, V. K., Bates, N. R., Becker, M., Bellouin, N., Bittig, H. C., Bopp, L., Chevallier, F., Chini, L. P., Cronin, M., Evans, W., Falk, S., Feely, R. A., Gasser, T., Gehlen, M., Gkritzalis, T., Gloege, L., Grassi, G., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Jain, A. K., Jersild, A., Kadono, K., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Landschützer, P., Lefèvre, N., Lindsay, K., Liu, J., Liu, Z., Marland, G., Mayot, N., McGrath, M. J., Metzl, N., Monacci, N. M., Munro, D. R., Nakaoka, S.-I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P. I., Pan, N., Pierrot, D., Pocock, K., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Rodriguez, C., Rosan, T. M., Schwinger, J., Séférian, R., Shutler, J. D., Skjelvan, I., Steinhoff, T., Sun, Q., Sutton, A. J., Sweeney, C., Takao, S., Tanhua, T., Tans, P. P., Tian, X., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G. R., Walker, A. P., Wanninkhof, R., Whitehead, C., Willstrand Wranne, A., et al.: Global Carbon Budget 2022, Earth System Science Data, 14, 4811–4900, https://doi.org/10.5194/essd-14-4811-2022, 2022.

Gaubert, B., Stephens, B. B., Basu, S., Chevallier, F., Deng, F., Kort, E. A., Patra, P. K., Peters, W., Rödenbeck, C., Saeki, T., Schimel, D., Van der Laan-Luijkx, I., Wofsy, S., and Yin, Y.: Global atmospheric CO2 inverse models converging on neutral tropical land exchange, but disagreeing on fossil fuel and atmospheric growth rate, Biogeosciences, 16, 117–134, https://doi.org/10.5194/bg-16-117-2019, 2019.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Köhler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., and Reichstein, M.: Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 17, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020, 2020.

Kaminski, T. and Heimann, M.: Inverse Modeling of Atmospheric Carbon Dioxide Fluxes, Science, https://doi.org/10.1126/science.294.5541.259a, 2001.

Ruehr, S., Keenan, T. F., Williams, C., Zhou, Y., Lu, X., Bastos, A., Canadell, J. G., Prentice, I. C.,

Kendall, A., Gal, Y., and Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 7482–7491, 2018.

Sitch, S., and Terrer, C.: Evidence and attribution of the enhanced land carbon sink, Nat Rev Earth Environ, 1–17, https://doi.org/10.1038/s43017-023-00456-3, 2023.

Schimel, D., Schneider, F. D., and JPL Carbon and Ecosystem Participants: Flux towers in the sky: global ecology from space, New Phytologist, 224, 570–584, https://doi.org/10.1111/nph.15934, 2019.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Valentini, R., Arneth, A., Bombelli, A., Castaldi, S., Cazzolla Gatti, R., Chevallier, F., Ciais, P., Grieco, E., Hartmann, J., Henry, M., Houghton, R. A., Jung, M., Kutsch, W. L., Malhi, Y., Mayorga, E., Merbold, L., Murray-Tortarolo, G., Papale, D., Peylin, P., Poulter, B., Raymond, P. A., Santini, M., Sitch, S., Vaglio Laurin, G., van der Werf, G. R., Williams, C. A., and Scholes, R. J.: A full greenhouse gases budget of Africa: synthesis, uncertainties, and vulnerabilities, Biogeosciences, 11, 381–407, https://doi.org/10.5194/bg-11-381-2014, 2014.