Response to comments

Note: In the following, we reply to the comments with texts in blue, and the referred line numbers are according to the tracked change document.

Reviewer #1: This study introduces MemoryFlux, an LSTM-based monthly NEE dataset for North America at 0.1° spatial resolution covering the period from 2001 to 2021. By incorporating multiple vegetation and environmental variables with historical information from a six-month look-back window, the authors aim to improve the representation of ecosystem temporal memory in regional carbon flux upscaling. The resulting product reproduces several well-known spatial and seasonal features, including strong growing season carbon uptake in the U.S. Midwest, and shows greater agreement with selected atmospheric inversion products than several existing empirical upscaling products in terms of seasonal amplitude and continental carbon sink magnitude. The comparison between MemoryFlux and nonMemoryFlux, together with the analyses of major drought and flood events, provides an interesting demonstration of the potential importance of temporal information in data-driven carbon cycle modelling. I consider the dataset suitable for publication in ESSD after minor revision; however, several aspects related to the attribution of improvements to "memory effects" and the benchmarking strategy should be strengthened and interpreted more cautiously.

Reply: Thank you for your positive evaluation and constructive comments. We have carefully revised the manuscript accordingly and provide point-by-point responses to each comment below.

The details are as follows:

The central scientific claim of this study is that incorporating memory effects improves the representation of NEE IAVs and climate-extreme-related anomalies. The manuscript states that MemoryFlux incorporates information from previous time steps, whereas nonMemoryFlux relies only on synchronous explanatory variables. However, the definition and implementation of nonMemoryFlux are not described in sufficient detail. Is the temporal look-back structure the only difference between the two model configurations? In addition, variables such as NDVI, LAI, FAPAR, and SIF inherently contain temporal persistence and may themselves reflect legacy effects.

Therefore, the implication that one model incorporates ecosystem memory while the other contains no memory may be overly simplistic or potentially misleading. Please clarify this issue and more clearly demonstrate that the differences between MemoryFlux and nonMemoryFlux can be specifically attributed to the representation of temporal memory effects. It would also be valuable to report quantitative differences in model performance between MemoryFlux and nonMemoryFlux at the site or PFT level, rather than relying primarily on representative time-series examples shown in Fig. S6.

Reply: Thank you for your valuable comments. In our experimental design, nonMemoryFlux uses the same predictor variables, model architecture, data partitioning, nested cross-validation framework, grid-search procedure, and early-stopping strategy as MemoryFlux. The only intentional difference between the two configurations is their temporal input structure: MemoryFlux uses a six-month look-back window that explicitly incorporates historical predictor information, whereas nonMemoryFlux uses only predictor values from the current time step. As noted in the manuscript, variables such as NDVI, LAI, FAPAR, and SIF exhibit temporal persistence and may therefore implicitly retain information related to antecedent ecosystem conditions. Consequently, both model configurations may contain some implicit legacy information through contemporaneous vegetation-state predictors, while MemoryFlux additionally incorporates historical predictor sequences through its temporal look-back structure. Accordingly, the comparison between MemoryFlux and nonMemoryFlux is more appropriately interpreted as an assessment of the added value of explicitly incorporating historical predictor information, rather than as a strict contrast between a “memory” and a completely “non-memory” model. To clarify this point, we have revised the relevant wording throughout the manuscript.

L444-455: “To evaluate whether explicitly incorporating historical climate and environmental information improved the representation of NEE IAVs, we generated two LSTM-based NEE datasets, MemoryFlux and nonMemoryFlux. The two configurations were identical with respect to predictors, model architecture, data partitioning, and nested cross-validation, differing only in their temporal input structure. MemoryFlux updated its internal state using information from previous time steps, thereby modelling temporal dependencies and potential memory effects between antecedent environmental conditions and current NEE. In contrast, nonMemoryFlux was driven only by synchronous explanatory variables. Both configurations may retain some implicit legacy information through contemporaneous vegetation-state predictors (e.g., NDVI and LAI), whereas MemoryFlux additionally incorporated historical predictor sequences explicitly through its temporal look-back structure. Therefore, the comparison of the two datasets provided a controlled assessment of the added value of explicitly incorporating historical predictor information.”.

Because all other model settings were held constant, differences in predictive performance between the two configurations can be more directly associated with the additional historical information introduced through the look-back structure. To further strengthen this assessment, we added quantitative comparisons of model performance between MemoryFlux and nonMemoryFlux at the PFT level in the revised manuscript. These additional results allow the performance gains associated with explicitly incorporating historical predictor information to be evaluated more systematically across vegetation types. We also added relevant wording to clarify that the observed differences between MemoryFlux and nonMemoryFlux are primarily associated with the explicit incorporation of historical predictor information through the temporal look-back structure.

L470-484: “We further compared, at the PFT level, the predictive performance of the LSTM models used to generate MemoryFlux and nonMemoryFlux (Fig. S6). Overall, R^2 increased from 0.63 for nonMemoryFlux to 0.79 for MemoryFlux, while RMSE decreased from 1.12 to 0.83 g C m^{-2} d^{-1} and MAE decreased from 0.69 to 0.52 g C m^{-2} d^{-1}. Improvements were observed across all PFTs, with particularly pronounced gains for ENF, for which R^2 increased by 0.39 (Tables S6 and S7). We also evaluated whether the temporal input structure improved the representation of seasonal and interannual variations in carbon fluxes at the site scale by comparing NEE estimates from MemoryFlux and nonMemoryFlux with long-term flux-tower observations across PFTs (Fig. S7). Clear differences were evident in their representation of temporal variability. MemoryFlux generally reproduced the observed seasonal cycle and IAVs more closely than nonMemoryFlux, particularly at lower NEE values, whereas nonMemoryFlux showed weaker IAVs. Together, these results suggested that explicitly incorporating historical predictor information improved both overall predictive performance and the representation of temporal variations in NEE.”.

A major narrative throughout the paper is that MemoryFlux “narrows the gap” between bottom-up estimates and atmospheric inversions products. While this result is interesting, numerical agreement with an inversion product should not automatically be interpreted as evidence of higher accuracy. EC-based upscaling does not explicitly account for harvest, some disturbance emissions, and other processes that may be included in top-down net land–atmosphere carbon flux estimates. Therefore, the difference between the approximately −1.27 Pg C yr-1 estimated by MemoryFlux and the approximately −0.7 to −0.8 Pg C yr-1 estimated by atmospheric inversion products should not be interpreted solely as model error; part of this discrepancy may arise from differences in flux definitions and system boundaries. I recommend that the authors precisely define the flux components extracted from CT2022 and OCO-2 MIP and describe in sufficient detail how fire emissions are removed. Furthermore, statements such as “more accurate” or “overestimate” should be replaced with more appropriate terms, such as “closer to,” “larger uptake than,” or “more consistent with,”.

Reply: Thank you for raising this important point. As you noted, we agree that numerical agreement with an atmospheric inversion product should not automatically be interpreted as evidence of higher accuracy. Although atmospheric inversions provide valuable independent references for evaluating seasonal and interannual variations and large-scale carbon budgets, they are not direct measurements of NEE and are subject to uncertainties associated with atmospheric transport, prior flux assumptions, observational coverage, and inversion methodology. Given the limited availability of large-scale direct flux observations, comparisons with atmospheric inversions provide a practical complementary means of evaluating regional NEE estimates. To clarify this point, we have revised the manuscript accordingly.

L244-248: “Importantly, closer agreement between MemoryFlux and the atmospheric inversions was not interpreted as direct evidence of higher accuracy. Instead, the inversions were treated as independent references based on fundamentally different methodologies, providing complementary information for assessing the consistency of regional NEE estimates in the absence of large-scale direct observations.”.

Furthermore, EC-based upscaling does not explicitly account for processes such as harvest, some disturbance-related emissions, and other carbon transfers that may influence top-down estimates of net land–atmosphere carbon exchange. Therefore, part of the remaining discrepancy between MemoryFlux and these inversions may reflect differences in flux definition and system boundaries rather than model error alone. We have added relevant discussion to the revised manuscript to clarify this issue.

L598-604: “Our slightly higher NEE estimates compared to those from previous top-down inversions may partly reflect differences in system boundaries and flux components represented by EC-based upscaling and atmospheric inversions. For example, fluxes associated with harvested products, fires, inland waters, and other lateral or disturbance-related carbon transfers may not be fully represented by tower-based ecosystem NEE but can influence atmospheric inversion estimates (Ballantyne et al., 2021).".

To more clearly explain the sources of differences among these products, we revised the manuscript to define the atmospheric inversion products and the specific flux components used in our analysis. For OCO-2 v10 MIP, we used the ensemble-mean posterior land flux from the "LNLGIS" experiment ("land"), which assimilates OCO-2 Land Nadir and Land Glint XCO_2 retrievals together with in situ CO_2 measurements. This land component represents the posterior terrestrial land–atmosphere CO_2 flux. To improve comparability with ecosystem NEE, we adjusted the inversion estimate by removing the externally prescribed wildfire-emission component ("fire_flux_imp") provided by NOAA CarbonTracker. For CT2022, we used the optimized terrestrial biosphere flux ("bio_flux_opt"), which represents land-biosphere CO_2 exchange excluding wildfire emissions. In CT2022, wildfire emissions are provided separately from the optimized biosphere component, together with fossil-fuel and ocean flux components. We have clarified these flux definitions and the treatment of wildfire emissions in the revised manuscript.

L233-244: "For OCO-2 v10 MIP, we used the ensemble-mean posterior land flux from the "LNLGIS" experiment ("land"), which assimilates OCO-2 Land Nadir and Land Glint XCO_2 retrievals together with *in situ* CO_2 measurements and provides 1° estimates for 2015–2020. To improve comparability with EC-based ecosystem NEE, we removed the prescribed wildfire-emission component using the National Oceanic and Atmospheric Administration (NOAA) CarbonTracker "fire_flux_imp" estimate. We also included CarbonTracker2022 (CT2022), the latest version of CarbonTracker developed by NOAA (Jacobson et al., 2023; Peters et al., 2007). For CT2022, we used the optimized terrestrial biosphere flux ("bio_flux_opt"), which provides 1° global land–atmosphere CO_2 exchange for 2001–2020, with wildfire emissions represented separately from the optimized biosphere flux.".

In summary, although we have made every effort to reduce differences between atmospheric inversion products and EC-based estimates arising from flux definitions and system boundaries, some discrepancies inevitably remain because both approaches are subject to inherent limitations and uncertainties. Atmospheric inversions should therefore be regarded as independent reference estimates rather than as ground truth. Accordingly, we have revised the interpretation of inter-product comparisons throughout the manuscript and moderated statements that could imply superior accuracy. We have also revised the previous “narrows the gap” narrative and now interpret the agreement between MemoryFlux and atmospheric inversions primarily as consistency between independent estimation approaches rather than as evidence of greater accuracy.

L264-267: “Comparisons with these datasets allowed us to assess the consistency and differences among MemoryFlux, previously reported regional NEE patterns, existing bottom-up products, and independent top-down inversion estimates, as well as to evaluate how effectively MemoryFlux represents regional-scale seasonal, interannual, and extreme-event-related variability in NEE.”.

L557-560: “Overall, among the NEE datasets considered, MemoryFlux provided a coherent representation of regional carbon flux anomaly patterns across the selected extreme events, including relatively moderate events such as the 2013 California drought (Fig. 8).”.

L385-386: “Even the recently updated FLUXCOM-X showed relatively strong JJA carbon uptake, second only to NIES2020.”.

L415-420: “These common-period comparisons showed that MemoryFlux was closer in magnitude to the atmospheric inversions than several regional and global EC-based flux-upscaling products, including EC-MOD2012, NIES2020, the FLUXCOM2020 ensemble products, and FLUXCOM-X.”.

The evaluation of extreme event performance is largely qualitative and should be supported by quantitative metrics. Figure 7 is visually compelling; however, the conclusion that MemoryFlux provides “more accurate constraints” on drought- and flood-related anomalies appears to rely primarily on visual agreement among spatial patterns and predefined event regions. Given the strength of this conclusion, additional quantitative evaluation would be preferable. Such analyses would transform the extreme-event section from a predominantly qualitative demonstration into a more robust assessment of the product’s performance. In addition, anomaly baselines are calculated using the individual temporal coverage of each product. Because the products have different temporal periods, this choice may influence anomaly magnitudes and complicate inter-product comparisons. A common reference period should be adopted for products with overlapping temporal coverage.

Reply: Thank you for pointing out this important issue. We agree that, in the original manuscript, the evaluation of NEE anomalies during extreme climate events was primarily based on the consistency of spatial anomaly patterns among different products and within the affected regions. Therefore, describing MemoryFlux as providing “more accurate constraints” on drought- and flood-related NEE anomalies was indeed too strong, particularly in the absence of independent regional-scale NEE observations for direct validation. We have therefore revised the relevant description in the manuscript accordingly.

L557-560: “Overall, among the NEE datasets considered, MemoryFlux provided a coherent representation of regional carbon flux anomaly patterns across the selected extreme events, including relatively moderate events such as the 2013 California drought (Fig. 8).”.

L616-621: “In this study, the regional NEE anomaly patterns represented by MemoryFlux showed broad consistency with those from atmospheric inversion products and were spatially coherent with concurrent anomalies in SM and SIF. This agreement suggested that MemoryFlux provided a reasonable representation of regional NEE anomaly patterns associated with the selected extreme events.”.

L646-647: “Overall, MemoryFlux showed a more coherent representation of subcontinental NEE anomaly patterns across the selected events than nonMemoryFlux.”.

L517-520: “The spatial correspondence among MemoryFlux NEE anomalies, environmental proxy anomalies, and the documented extent of the selected extreme events suggested that MemoryFlux provided a coherent representation of regional NEE anomaly patterns associated with these events.”.

In this study, we primarily evaluated the ability of different NEE products to represent the spatial patterns of NEE anomalies associated with extreme events at the grid scale by comparing them with independent proxies of vegetation activity and moisture conditions, namely solar-induced chlorophyll fluorescence (SIF) and soil moisture (SM) anomalies, within the regions affected by selected extreme climate events. The extreme events and their major affected regions shown in Figs. 7 and 8 were selected mainly with reference to Byrne et al. (2019). That study similarly compared spatial patterns of NEE anomalies with independent proxy variables, including SIF, drought indices, and soil temperature, and used rectangular boxes to delineate major climate-anomaly regions, thereby assessing the ability of carbon-flux products to represent spatial anomalies associated with extreme events. Accordingly, our analysis largely followed this diagnostic framework based on spatial anomaly consistency.

"In the northern extra-tropics, the observational coverage of GOSAT is highly seasonal, so we limit our analysis of anomalies in the northern extra-tropics to the summer (JJA) when observational coverage is the best (Liu et al., 2014; Byrne et al., 2017). Figure 3 shows the anomalies for the proxies, FLUXCOM NEE, and $GC_{2\times2.5\text{–}200\%}$ NEE across the Northern Hemisphere for JJA 2010–2013. The proxies and FLUXCOM generally show high coherence in anomalies. Events for which FLUXCOM NEE gives enhanced emissions to the atmosphere also show reduced SIF, increased scPDSI, and increased Tsoil. We have highlighted (with boxes) major climate anomalies over this time period: the 2010 Russian heat wave, the 2011 drought in Mexico and the southern USA, the 2012 North American drought, and the 2013 California drought." (Byrne et al. 2019)

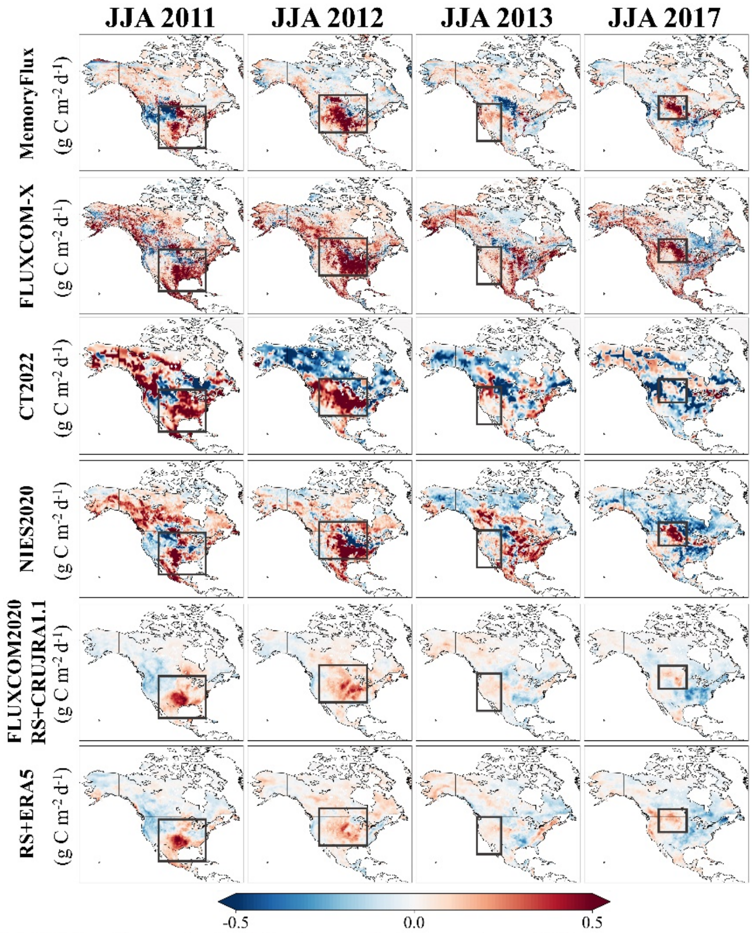
We further recognize that, given the current data constraints, there are clear limitations to conducting a rigorous quantitative evaluation of NEE anomalies during extreme events. Continuous regional-scale NEE observations that could serve as independent ground truth are unavailable, whereas SIF and SM, although reflecting vegetation photosynthetic activity and moisture conditions, respectively, and providing independent information on ecosystem responses to extreme events, are not direct observations of NEE. Therefore, even if the relationships between NEE and SIF or SM were quantified further, such metrics would primarily characterize spatial consistency among different variables rather than directly quantify the prediction error or accuracy of the NEE products. For this reason, we did not extend the revised analysis into a formal quantitative performance evaluation. Instead, we redefined the objective of this section as assessing whether MemoryFlux can reasonably represent the spatial patterns of NEE anomalies associated with selected extreme climate events and whether these anomalies show overall consistency with ecosystem responses reflected by concurrent SIF and SM anomalies. Accordingly, we have explicitly stated in the Discussion that this type of analysis, based on spatial anomaly patterns and consistency with proxy variables, is primarily qualitative and cannot replace rigorous quantitative validation against independent NEE observations. Through these revisions, we have limited the conclusions of the extreme-event analysis to a diagnostic assessment of the ability of different NEE products to represent spatial anomaly patterns and their consistency with independent ecological proxy variables, rather than interpreting the results as direct evidence that MemoryFlux has higher quantitative accuracy during extreme events.

L496-498: "The SIF and SM datasets were used as supporting ecological and hydroclimatic indicators and were not intended as direct proxies or independent observations of NEE.”.

L513-515: “The spatial extent of the NEE anomalies represented by MemoryFlux was also broadly consistent with the reported extent of the 2019 flood, with widespread flooding across the Midwest and more limited impacts farther south (Fig. S8).”.

L622-627: “However, this assessment was primarily qualitative. Because continuous regional-scale NEE observations that could serve as an independent reference were unavailable, rigorous quantitative evaluation of model performance during extreme events remains difficult. Moreover, SM and SIF were not direct observations of NEE and therefore provided supporting ecological and hydroclimatic information rather than direct validation of carbon-flux estimates. Accordingly, the present analysis should be interpreted as a diagnostic assessment of spatial consistency rather than as evidence of superior quantitative accuracy.”.

In the original manuscript, anomalies for different products were calculated using their respective periods of data availability, which may influence anomaly magnitudes and complicate direct inter-product comparisons. To improve comparability, we therefore recalculated the anomalies using a common 2011–2017 reference period for products with overlapping temporal coverage, including MemoryFlux, FLUXCOM-X, CT2022, NIES2020, and the FLUXCOM2020 ensemble products. The recalculated results showed that the spatial extent and overall patterns of the regional NEE anomalies remained broadly similar to those obtained using the original product-specific reference periods, although the anomaly magnitudes changed slightly. This suggests that the identification of the main anomaly regions was generally insensitive to the choice of reference period, whereas the anomaly magnitudes were affected to some extent. We have therefore revised the manuscript to explicitly acknowledge the influence of the reference period on anomaly magnitude and to clarify that a common 2011–2017 baseline was used to improve inter-product comparability.



JJA 2011: *Drought in Mexico and the southern USA*

JJA 2012: *North American drought* **JJA 2013:** *California drought*

JJA 2017: *Flash drought across the northern U.S and Canadian Prairies*

JJA 2019: *Midwest American floods*

JJA 2021: *Southwestern North American drought and associated wildfires*

Figure S9. Spatial anomalies of JJA NEE associated with major drought and flood events in North America, calculated using a common 2011–2017 reference period for MemoryFlux, CT2022, NIES2020, FLUXCOM2020 RS+CRUJRA1.1, FLUXCOM2020 RS+ERA5, and FLUXCOM-X. Black rectangles indicate the regions most strongly affected by the selected extreme climate events.

L553-557: “To improve comparability, we recalculated the anomalies using a common 2011–2017 reference period for products with overlapping temporal coverage, including MemoryFlux, FLUXCOM-X, CT2022, NIES2020, and the FLUXCOM2020 ensemble products (Fig. S9). The recalculated results showed that the spatial extent and overall patterns of the regional NEE anomalies remained broadly similar to those obtained using the original product-specific reference periods, although the anomaly magnitudes changed slightly.”.

References:

Byrne, B., Jones, D. B. A., Strong, K., Zeng, Z.-C., Deng, F., and Liu, J.: Sensitivity of CO_2 Surface Flux Constraints to Observational Coverage, J. Geophys. Res.-Atmos, 112, 6672–6694, https://doi.org/10.1002/2016JD026164, 2017.

Byrne, B., Jones, D. B. A., Strong, K., Polavarapu, S. M., Harper, A. B., Baker, D. F., and Maksyutov, S.: On what scales can GOSAT flux inversions constrain anomalies in terrestrial ecosystems? Atmos. Chem. Phys., 19, 13017–13035, https://doi.org/10.5194/acp-19-13017-2019, 2019.

Liu, J., Bowman, K.W., Lee, M., Henze, D. K., Bousserez, N., Brix, H., Collatz, G. J., Menemenlis, D., Ott, L., Pawson, S., Jones, D. B. A., and Nassar, R.: Carbon monitoring system flux estimation and attribution: impact of ACOS-GOSAT XCO_2 sampling on the inference of terrestrial biospheric sources and sinks, Tellus B, 66, 22486, https://doi.org/10.3402/tellusb.v66.22486, 2014.

Specific comments

L136: "terrestrial vegetation biomes"? These classes are actually IGBP land-cover/PFT categories. Please consider using more precise terminology throughout the manuscript.

Reply: Thank you for pointing out this issue. We agree that the classes used in this study are based on the IGBP land-cover classification and are treated as plant functional types (PFTs) within the modeling framework. We have revised the wording at L154 and carefully reviewed the manuscript to ensure consistent use of the relevant terminology throughout.

L154-155: "The study area encompassed the major terrestrial land-cover types across North America, extending from 5° N to 80° N.".

L160-161: "The database covered 10 major plant functional types (PFTs) in North America (Fig. 1)".

L268-270: The manuscript repeatedly emphasizes that MemoryFlux captures the strongest carbon uptake in the Midwest Corn Belt during the growing season.

However, on an annual basis, MemoryFlux identifies the strongest carbon sink in the southeastern U.S., whereas the atmospheric inversion products still indicate stronger uptake in the Midwest. Please explain more clearly how the successful reproduction of the peak growing-season Corn Belt signal can be reconciled with the disagreement in annual spatial patterns relative to the inversions.

Reply: Thank you for this insightful comment. Previous studies have reported exceptionally strong growing-season carbon uptake in the Midwest Corn Belt, whereas forests in the southeastern United States have been identified as an important annual carbon sink. Our results show that MemoryFlux reproduces both features: pronounced peak growing-season uptake in the Midwest and strong annual carbon uptake in the southeastern United States. These seasonal and annual spatial patterns are therefore not contradictory but instead reflect different temporal characteristics of the regional carbon cycle.

We agree that the annual spatial pattern derived from the atmospheric inversions differs from this pattern, with stronger carbon uptake remaining centered over the Midwest. We consider this discrepancy an important result and report it explicitly rather than implying agreement among products. The difference may reflect methodological and system-boundary differences, as well as uncertainties inherent in both EC-based upscaling and atmospheric inversion approaches. Our study aims to evaluate both the consistency and differences among MemoryFlux, previously reported regional NEE patterns, and existing products in terms of spatial distribution, seasonality, and interannual variability. Compared with other products, MemoryFlux shows greater consistency with previously reported regional patterns in representing the spatial distribution of carbon sinks at both growing-season and annual scales. Accordingly, we have revised the manuscript to clarify that MemoryFlux is more consistent with previously reported regional patterns and with most of the published NEE products in reproducing the contrast between strong growing-season uptake in the Midwest and strong annual uptake in the southeastern U.S., while interpreting the disagreement with atmospheric inversions more cautiously.

L140-143: “We evaluate the consistency and differences among MemoryFlux, previously reported empirical knowledge, existing bottom-up products, and independent top-down inversion estimates, with particular attention to seasonal, interannual, and spatial variability.”.

L331-342: “At the annual scale, MemoryFlux and the other ML-based flux-upscaling products consistently identified the southeastern United States as the region of strongest carbon uptake, whereas the OCO-2 v10 MIP and CT2022 inversions showed stronger annual uptake centred over the Midwest Corn Belt. This discrepancy may reflect methodological differences and uncertainties inherent in both EC-based upscaling and atmospheric inversion approaches. During the peak growing season, MemoryFlux reproduced pronounced seasonal carbon uptake over the Midwest U.S. croplands, a feature also evident in OCO-2 v10 MIP, CT2022, and EC-MOD2012. In contrast, other global flux-upscaling products showed their strongest seasonal uptake in the southeastern United States. These results indicated that the seasonal and annual spatial patterns represented by MemoryFlux were more consistent with previous empirical evidence and with most published estimates.".

L499: Several PFT-specific models are based on a very limited number of sites. Table S1 includes only two sites for SAV, with similarly sparse representation for MF, WSA, and CSH. Although the manuscript already discusses SAV as an example, I think it is necessary to report either the number of sites or the number of monthly NEE observations for each PFT, as this information is critical for assessing the reliability of the regional upscaling results.

Reply: Thank you for your valuable suggestion. In the revised manuscript, we have added the number of flux sites and the corresponding number of monthly NEE observations for each PFT in Table S2.

L155-157: "We used observations from 84 flux sites (Table S1), totaling7,711 monthly NEE records (Table S2), drawn from FLUXNET2015 (Pastorello et al., 2020) and the AmeriFlux network (Novick et al., 2018).".

L250-252: Please clarify whether the gridded SAV model uses a different set of predictors from the other PFT-specific models and, if so, whether this difference affects the comparability of model performance among PFTs.

Reply: Thank you for pointing out this issue. The gridded SAV model indeed used a slightly different predictor set from the other PFT-specific models. Specifically, SWC was excluded from the site-level SAV model because reliable SWC observations were unavailable at the SAV flux sites. To maintain consistency between site-level model training and regional prediction, SWC was therefore also excluded from the corresponding gridded SAV model. In our research, each PFT-specific model was trained and evaluated independently using the predictor information available for that PFT, and all models showed good predictive performance within their respective PFTs. The SAV model also performed well (R^2 = 0.83). Therefore, we do not expect this difference to substantially affect the overall comparability of model performance among PFTs. The performance metrics remain informative for evaluating model skill within each PFT. Nevertheless, we believe that the absence of SWC may limit further improvement in SAV model performance. We have added a corresponding discussion in the revised manuscript to clarify this issue.

L310-314: “For the SAV PFT, SWC was not included as a predictor because reliable site-level SWC observations were unavailable. To maintain consistency between site-level model training and regional upscaling, SWC was also excluded from the corresponding gridded SAV model. Despite the reduced predictor set, the SAV LSTM model still demonstrated strong predictive performance (R^2 = 0.83; Fig. S3).”.

L669-672: “For the SAV PFT, SWC was excluded from the predictor set. Although the SAV model still showed good predictive performance, the reduced predictor set may limit further improvements in model performance. Predictions for sparsely sampled PFTs and regions with limited flux-tower coverage should therefore be interpreted with greater caution.”.

Please avoid interpreting standard deviations across years as uncertainty intervals. In Fig. 4 and Tables S2–S3, please clearly specify whether the shading and error bars represent interannual variability, model ensemble spread, or another source of variation.

Reply: Thank you for pointing out this issue. We agree that the standard deviations across years represent interannual variability rather than model uncertainty or model ensemble spread. We have therefore clarified the meaning of the shading and error bars in Fig. 4 and Tables S3–S4.

L380: “Error bars indicate the standard deviation across years, representing interannual variability.”.

“Table S4. Mean and standard deviation of NEE over the season scale (Pg C season-1) estimated by seven products (EC-MOD2012 data unavailable) mentioned in this study in whole North America during 2015–2017. The standard deviation represents interannual variability.”.

L549-550: The increasing carbon sink trend is not statistically significant and should therefore not be described as an established long-term increase.

Reply: Thank you for this comment. We have revised the manuscript to describe this result more cautiously as a non-significant tendency toward increasing carbon uptake.

L746-748: “The annual total NEE exhibited a non-significant tendency toward increasing carbon uptake in North America ($k = -0.009$ Pg C yr^{-2}, $p = 0.09$).”

The authors acknowledge in the Discussion that atmospheric inversions are not direct measurements of NEE. In this context, statements such as “overestimate JJA carbon uptake” and “overestimate carbon uptake in recent years” may be overly strong, as they implicitly treat the reference estimates as ground truth. Please revise similar wording throughout the manuscript. Furthermore, the phrase “monitoring climate extremes” may be somewhat too strong for these model-generated NEE products. Terms such as “representing” or “capturing” may be more appropriate.

Reply: Thank you for this helpful comment. We agree that atmospheric inversion estimates should not be treated as ground truth and that expressions such as “overestimate JJA carbon uptake,” “overestimate carbon uptake in recent years,” and “monitoring climate extremes” are therefore overly strong when referring to model-derived NEE products. We have accordingly revised similar expressions throughout the manuscript.

L203-204: “This product provides long-term information on global vegetation dynamics.”.

L385-386: “Even the recently updated FLUXCOM-X showed relatively strong JJA carbon uptake, second only to NIES2020.”.

L457-458: “In particular, nonMemoryFlux tended to exhibit a stronger increasing trend (−0.013 Pg C yr^{-2}) and estimated stronger carbon uptake in recent years (since 2014),”.

L485: “Capturing climate extremes impacts on regional NEE”.

L486-487: “We further evaluated the ability of the LSTM-based model to represent NEE IAVs at the grid scale by examining its representation of spatial NEE anomalies associated with selected climate extremes.”.

The manuscript refers to the “2020–2021 southwestern North American drought,” whereas the event analysis presented in the figures appears to focus on a specific summer period. Please clarify which summer or summers were used to calculate the anomaly and provide the rationale for this choice.

Reply: Thank you for pointing this out, and we apologize for the lack of clarity. For the 2020–2021 drought in southwestern North America, the NEE anomaly shown in our analysis was calculated for JJA 2021 relative to the corresponding climatological baseline. Although the drought began developing in summer 2020 and persisted through 2021, we selected summer 2021 because it represented a more mature and severe phase of the event. We have clarified this rationale in the revised manuscript and now explicitly state that the 2020–2021 southwestern North American drought is represented by the JJA 2021 anomaly in our analysis.

L508-510: “For the 2020–2021 drought in southwestern North America, the NEE anomaly was evaluated for JJA 2021 relative to the corresponding climatological baseline, as this period represents a more mature and severe phase of the event.”