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.

Review #2: The manuscript by Huang et al. presents a new monthly, 0.1° × 0.1° gridded NEE dataset for North America spanning 2001–2021, termed MemoryFlux. This dataset was developed using plant-functional-type-specific LSTM models trained with eddy-covariance observations and multiple environmental and satellite-derived predictors. The authors show that MemoryFlux reproduces the strong summertime carbon uptake over the U.S. Midwest Corn Belt, exhibits seasonal variations broadly consistent with those from OCO-2 v10 MIP and CT2022 atmospheric inversions, and estimates a continental-scale annual carbon sink that is substantially weaker than those reported by many existing machine-learning upscaling products. The manuscript further investigates the potential contribution of temporal memory by comparing MemoryFlux with a non-memory configuration and examines NEE anomalies associated with several major drought and flood events. Overall, the dataset represents a potentially valuable contribution to regional carbon-cycle research, and the manuscript is generally well organized. I support publication after minor revision. However, several methodological and interpretive issues should be clarified to further improve the robustness, reproducibility, and appropriate application of the dataset.

Reply: Thank you very much for your interests in our study, quite positive evaluation and valuable comments. We have made our best to tackle the proposed issues as possible.

Major comments:

(1) The manuscript states that the site-level LSTM models were developed using environmental variables derived from EC tower observations, whereas regional

upscaling was performed using ERA5-Land meteorological variables. If tower-based meteorological measurements were indeed used during model training while ERA5-Land variables were used for regional inference, this may introduce a potential predictor-domain mismatch. Differences in air temperature, precipitation, soil moisture, radiation, and VPD between tower measurements and ERA5-Land estimates could propagate into the regional predictions. Please clarify whether ERA5-Land predictors were extracted at flux-tower locations and used to train the final upscaling models, or whether tower-measured meteorological variables were used during model training. Relatedly, the description of the EC-derived target variable requires further detail. The authors should specify the exact FLUXNET/AmeriFlux predictors used, the quality-control criteria applied, the gap-filling procedures, and the minimum monthly data-coverage requirement for retaining observations. These details are important for an ESSD data paper and should be described explicitly rather than relying solely on a previous publication.

Reply: Thank you very much for raising this important issue. We confirm that the site-scale LSTM models in this study were trained using meteorological and environmental variables observed at EC flux-tower sites, whereas regional upscaling was driven by the corresponding 0.1° gridded meteorological variables from ERA5-Land. Our rationale was to first establish and evaluate the relationships between NEE and its environmental drivers using co-located tower observations, thereby making full use of the site-level observational information available during model training. After evaluating model performance at the site scale, the trained PFT-specific LSTM models were applied to regional NEE prediction.

At the regional scale, spatially continuous ground-based meteorological observations are not available across North America, and ERA5-Land was therefore used to provide the gridded meteorological forcing required for upscaling. To reduce inconsistencies between the site-training and regional-prediction stages, we harmonized the ERA5-Land variables with the corresponding EC-site variables in terms of physical definition and units. For example, the 10 m u- and v-components of wind from ERA5-Land were combined to calculate wind speed corresponding to the EC-site WS_F variable; soil water content in the uppermost ERA5-Land soil layer (0–7 cm) was converted to the same unit as the shallowest available site-level SWC measurement; and air temperature, precipitation, downward shortwave radiation, and

VPD were likewise processed to match the definitions and units of the corresponding tower variables. We have clarified these procedures in the revised manuscript.

L286-290: “Key inputs for site-scale model training included meteorological and environmental variables measured at EC flux towers (TA_F, WS_F, P_F, DSR_F, SWC_F_MDS_#, and VPD_F), together with vegetation variables derived from remote sensing datasets (NDVI, LAI, FAPAR, and SIF).”.

L293-296: “The LSTM models were trained using tower-observed meteorological variables rather than ERA5-Land data. This design was intended to make full use of co-located EC-site observations when establishing empirical relationships between NEE and its environmental drivers.”.

L304-305: “To reduce inconsistencies between site-training and regional prediction, ERA5-Land variables were harmonized with the corresponding EC-site variables in terms of physical definitions and units before being used as model inputs.”.

We agree that even after harmonizing variable definitions and units, a potential predictor-domain mismatch may remain between EC-site observations and ERA5-Land gridded data. To further evaluate this issue, we extracted ERA5-Land data at all corresponding flux-tower locations and compared them with concurrent meteorological observations from the same sites (Fig. S10). The results showed strong agreement for air temperature, downward shortwave radiation, VPD, and precipitation, whereas agreement was weaker for wind speed and soil water content. These differences may reflect the contrasting spatial representativeness of point-scale tower observations and 0.1° gridded reanalysis data, as well as differences in local topography, surface roughness, variable definitions, and soil-moisture measurement depth. We therefore acknowledge that this predictor-domain mismatch cannot be completely eliminated and have identified it as a potential source of uncertainty in regional upscaling. Corresponding explanations have been added to the revised manuscript.

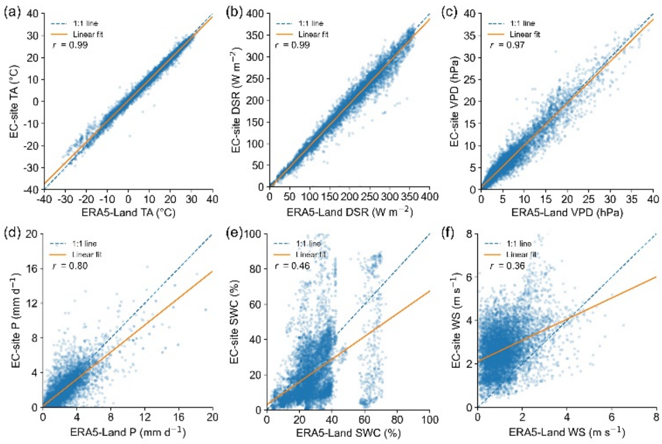

Figure S10. Comparison of ERA5-Land gridded meteorological variables with concurrent observations from EC flux-tower sites. Scatterplots show (a) air temperature (TA), (b) downward shortwave radiation (DSR), (c) vapor pressure deficit (VPD), (d) precipitation (P), (e) soil water content (SWC), and (f) wind speed (WS). Dashed lines indicate the 1:1 relationship, and solid lines indicate linear regression fits.

L680-686: “Second, the site-level models were trained using tower-observed meteorological variables, whereas regional upscaling relied on ERA5-Land gridded predictors, potentially introducing a predictor-domain mismatch despite harmonization of variable definitions and units. Comparisons at flux-tower locations showed strong agreement for TA, DSR, VPD, and P, but weaker agreement for WS and SWC (Fig. S10). These discrepancies may reflect differences in spatial representativeness between point-scale observations and 0.1° gridded reanalysis data, as well as differences in variable definitions, local topography, surface roughness, and soil-moisture measurement depth, and therefore represent an additional source of uncertainty in regional extrapolation.”.

In addition, we have provided further details on the processing of the EC data and target variables. Monthly NEE_VUT_REF data from FLUXNET and AmeriFlux were used as the model target variable, with units of g C m^{-2} d^{-1}. The site-level environmental predictors included TA_F, P_F, WS_F, SW_IN_F, VPD_F, and SWC_F_MDS_#. Because the effective measurement depth of SWC varied among sites, we used the shallowest available SWC observation at each site as the model input to maximize data availability. We used network-provided variables that had undergone quality control and gap filling, and further removed values flagged as −9999 as well as obvious outliers. For sites represented in both the FLUXNET and AmeriFlux databases, we preferentially used the record with the longer time series and more complete valid observations. To ensure sufficient temporal information for model training, only sites with at least 18 consecutive valid monthly observations were retained. We have added corresponding descriptions to the revised manuscript to clarify these procedures.

L286-293: "Key inputs for site-scale model training included meteorological and environmental variables measured at EC flux towers (TA_F, WS_F, P_F, DSR_F, SWC_F_MDS_#, and VPD_F), together with vegetation variables derived from remote sensing datasets (NDVI, LAI, FAPAR, and SIF). For SWC, the shallowest available SWC measurement at each site was used to maximize data availability. We used the quality-controlled and gap-filled variables provided by the flux networks, removed values flagged as −9999 as well as obvious outliers, and retained only sites with at least 18 consecutive valid monthly observations. For sites represented in both FLUXNET and AmeriFlux, the record with the longer and more complete time series was used.".

(2) The manuscript emphasizes that MemoryFlux narrows the gap between bottom-up estimates and atmospheric inversion products such as OCO-2 v10 MIP and CT2022. However, atmospheric inversions and EC-based upscaling approaches do not necessarily represent identical carbon-flux components. Therefore, the manuscript should more carefully establish the comparability between MemoryFlux and atmospheric inversion products before interpreting their agreement as evidence of improved NEE estimation. Indeed, the authors acknowledge that harvest, river transport, fires, and other carbon fluxes may contribute to discrepancies between bottom-up and top-down estimates, and that atmospheric inversions themselves are subject to uncertainties associated with atmospheric transport and prior assumptions. Although fire emissions were removed from the OCO-2 v10 MIP comparison and are reported to be excluded from CT2022, the manuscript should more explicitly define the flux components included in each product and demonstrate that the comparisons are as consistent as possible. In addition, the mean annual NEE values reported for the different products are calculated over different temporal periods, which may influence the apparent similarity among products. I recommend either recalculating key budget

comparisons over common overlapping periods or providing a sensitivity analysis to demonstrate that the main conclusions are not driven by differences in temporal coverage.

Reply: Thank you for this important and constructive comment. We agree that EC-based NEE upscaling and atmospheric inversions do not represent identical carbon-flux components or system boundaries. 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 externally prescribed wildfire-emission component using the NOAA CarbonTracker "fire_flux_imp" estimate. 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.

Nevertheless, we acknowledge that, although we have made every effort to minimize differences related to wildfire emissions in the product comparisons, processes such as harvest, lateral carbon transport, some disturbance-related emissions, and other non-ecosystem carbon fluxes may still contribute differently to the two types of estimates. These adjustments therefore cannot fully eliminate differences in system boundaries between EC-based upscaling and atmospheric inversions. Atmospheric inversions are also subject to uncertainties associated with atmospheric transport, prior flux assumptions, and inversion configurations. To clarify this issue, we revised the manuscript to explicitly define the atmospheric inversion products, the flux components used, and the treatment of wildfire emissions. We also treat the inversions as independent reference estimates rather than ground truth and have revised potentially over-strong expressions such as "more accurate," "overestimate," and "narrows the gap" throughout the manuscript.

L228-232: "Atmospheric inversions infer surface fluxes by combining atmospheric CO_2 observations with atmospheric transport models. This makes them useful for characterizing large-scale seasonal and interannual variations, and they therefore serve as independent references for evaluating bottom-up products.”.

L233-248: “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. 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.”.

L45-47: “MemoryFlux provides an improved bottom-up estimation of North American NEE and shows greater consistency in magnitude with independent atmospheric inversion estimates than several existing EC-based upscaling products.”.

L140-147: “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. We also assess whether MemoryFlux shows greater consistency with inversions than existing bottom-up upscaling products and whether it provides a coherent representation of NEE anomalies associated with selected climate extremes.”.

L147-149: “We anticipate that MemoryFlux, which explicitly incorporates historical ecosystem information within an LSTM framework, provides a robust representation of ecosystem carbon fluxes and serves as a valuable resource for investigating terrestrial carbon dynamics and ecosystem–climate interactions.”.

We also agree that differences in temporal coverage may affect comparisons of mean annual NEE among products. In the original manuscript, the reported long-term means were calculated over the individual periods available for each dataset. To assess whether this affected our conclusions, we recalculated the key carbon-budget comparisons using common overlapping periods wherever possible. Specifically, we recalculated comparisons among MemoryFlux, OCO-2 v10 MIP, CT2022, and FLUXCOM-X over their common 2015–2020 period, as well as comparisons among MemoryFlux, CT2022, EC-MOD2012, NIES2020, the FLUXCOM2020 ensemble products, and FLUXCOM-X over their common 2001–2012 period. These common-period analyses showed that the annual total NEE estimated by MemoryFlux remained closer in magnitude to the atmospheric inversion estimates than several of the EC-based flux-upscaling products, indicating that our main conclusion regarding the relative magnitude of the MemoryFlux carbon sink was not driven by differences in temporal coverage. As shown in Fig. 5 and Table S5, this pattern was also generally evident across the broader study period. To clarify this issue, we have added corresponding text to the revised manuscript.

L411-420: “We further compared annual NEE over common periods. For 2015–2020, mean annual NEE was −1.38 Pg C yr^{-1} for MemoryFlux, −0.83 Pg C yr^{-1} for OCO-2 v10 MIP, −0.69 Pg C yr^{-1} for CT2022, and −1.75 Pg C y^{r-1} for FLUXCOM-X. For 2001–2012, the corresponding estimates were −1.21 Pg C yr^{-1} for MemoryFlux, −0.66 Pg C yr^{-1} for CT2022, −1.81 Pg C yr^{-1} for EC-MOD2012, −3.19 Pg C yr^{-1} for NIES2020, −3.05 and −2.78 Pg C yr^{-1} for the two FLUXCOM2020 ensemble products, and −1.65 Pg C yr^{-1} for FLUXCOM-X. 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.".

(3) The manuscript reports a mean annual NEE of approximately −1.27 ± 0.12 Pg C yr^{-1}. However, the reported ±0.12 value appears to represent the temporal standard deviation of annual estimates rather than an uncertainty estimate for MemoryFlux itself. This distinction is important because the product is generated using PFT-specific neural networks trained with a spatially uneven and, for some PFTs, limited number of EC sites. The manuscript explicitly recognizes substantial observational limitations in boreal and tropical regions and for sparsely sampled PFTs. For a data product, users need to understand not only interannual variability but also the uncertainty associated with model structure and spatial extrapolation. If possible, explicitly reporting such uncertainty information would substantially improve the usability of MemoryFlux.

Reply: Thank you for pointing out this issue. The reported ±0.12 Pg C yr^{-1} represents the interannual standard deviation of annual NEE estimates rather than a formal uncertainty estimate for MemoryFlux. We have therefore revised the manuscript to clearly distinguish interannual variability from model uncertainty and have avoided interpreting this value as an uncertainty interval.

L407-409: "MemoryFlux estimated annual NEE ranging from −1.50 to −1.05 Pg C yr^{-1} over 2001–2021, with a mean of −1.27 Pg C yr^{-1} and an interannual standard deviation of 0.12 Pg C yr^{-1}.".

We agree that information on model and spatial-extrapolation uncertainty would be valuable for users of the dataset. However, the current version of MemoryFlux does not include a formal uncertainty framework that comprehensively accounts for uncertainties arising from model structure, predictor data, and spatial extrapolation. To provide users with additional information on product reliability, we now report the number of flux-tower sites, the number of available monthly observations for each PFT (Tables S2), and the quantitative performance of the LSTM models used to generate MemoryFlux and nonMemoryFlux in the revised manuscript (Tables S6 and S7). These metrics provide some indication of model reliability and the potential uncertainty associated with spatial extrapolation. However, we acknowledge that they are insufficient to fully characterize the overall uncertainty of the dataset. We have explicitly acknowledged this limitation in the revised manuscript and expanded the Discussion to emphasize that estimates for sparsely represented PFTs and regions with limited flux-tower coverage should be interpreted with greater caution.

L154-158: "The study area encompassed the major terrestrial land-cover types across North America, extending from 5° N to 80° N. 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). These observations were combined with various remote sensing and climate reanalysis datasets.".

L470-473: "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).".

L671-676: "Predictions for sparsely sampled PFTs and regions with limited flux-tower coverage should therefore be interpreted with greater caution. Furthermore, although we provide quantitative PFT-level model performance metrics (Tables S6 and S7), the current version of MemoryFlux does not include a spatially explicit uncertainty framework that fully characterizes uncertainties arising from model structure, predictor data, and spatial extrapolation. Developing such a framework will be an important direction for future improvements to the dataset.".

Minor comments:

Line 40: The phrase “all the recent extreme drought and flood events” is overly broad. I suggest revising this expression to “caused by the six selected severe drought and flood events during the study period.

Reply: Thank you for this valuable suggestion. We agree that the original wording was overly broad and have revised it to more accurately reflect the scope of our analysis.

L41-43: “In addition, MemoryFlux captured spatial NEE anomaly patterns associated with six selected severe drought and flood events.”.

Please explicitly define the NEE sign convention at an appropriate point in the manuscript. Although the manuscript implies that negative NEE values indicate carbon uptake, this convention should be clearly stated for users of the dataset.

Reply: Thank you for this helpful comment. We have added a clear definition in the revised manuscript, specifying that negative NEE values indicate net carbon uptake by terrestrial ecosystems, whereas positive NEE values indicate net carbon release to the atmosphere.

L158-160: “The target variable was monthly average NEE (NEE_VUT_REF; g C m^{-2} day^{-1}), with negative NEE values indicating net carbon uptake by terrestrial ecosystems and positive values indicating net carbon release to the atmosphere.”.

Line 133: The manuscript states that 84 flux sites were used, whereas the Code and data availability section indicates that 35 FLUXNET sites and 47 AmeriFlux sites were included, resulting in a total of 82 sites. This inconsistency is confusing and should be carefully checked and corrected.

Reply: Thank you for pointing out these inconsistencies. We apologize for the typographical error in the Code and data availability section. The correct numbers are 35 FLUXNET sites and 49 AmeriFlux sites, for a total of 84 flux sites. We have carefully reviewed and corrected the relevant descriptions throughout the manuscript to ensure consistency.

Table 1 should provide the units for all predictors, and the specific ERA5-Land soil moisture layer used in this analysis should be clearly specified.

Reply: Thank you for this helpful suggestion. We have added the units of all predictors to Table 1. The reported units correspond to those used as model inputs after preprocessing and unit conversion, rather than to the native units of the original datasets. We have also specified the ERA5-Land soil moisture layer used in the analysis.

L174: "soil water content in the first soil layer (0–7 cm; SWC)".

L179-180: "All gridded predictors were converted to the same units as their corresponding variables used for site-level model training.".

L304-305: "To reduce inconsistencies between site-training and regional prediction, ERA5-Land variables were harmonized with the corresponding EC-site variables in terms of physical definitions and units before being used as model inputs.".

Lines 189-192: I agree with the approach used to generate the static PFT map. However, the potential impacts of neglecting land-cover changes should at least be briefly discussed, particularly with respect to croplands, forest disturbances, and fire-affected regions.

Reply: Thank you for this helpful comment. We used a static PFT map to reduce interannual classification noise in the MODIS land-cover product and to maintain consistent PFT assignments throughout the study period. However, this approach cannot capture actual land-cover transitions, such as cropland expansion or abandonment, forest disturbance and regrowth, and fire-induced vegetation changes. These transitions may lead to mismatches between the assigned PFT and the actual land-cover condition in a given year, thereby affecting the corresponding NEE estimates. We have added this limitation to the revised Discussion.

L688-691: "Although the static PFT map used here helps reduce interannual classification noise, it cannot capture actual land-cover transitions, including cropland expansion or abandonment, forest disturbance and regrowth, and fire-induced vegetation changes. Such mismatches may introduce additional uncertainty into regional NEE estimates.".

Line 192: The phrase "major interpolation method" used to describe the land-cover resampling procedure is unclear. I assume that the authors refer to a majority aggregation method. Please specify the exact resampling algorithm used.

Reply: Thank you for pointing this out. We apologize for the unclear wording. The land-cover dataset was resampled using a majority aggregation method. We have revised the corresponding description in the manuscript to explicitly specify the resampling algorithm.

L223-224: "The resulting map was then resampled to a spatial resolution of 0.1° using the majority aggregation method.".

Line 297: The seasonal correlation coefficients (r = 0.96 and 0.97) are high; however, correlations calculated from a 12-month climatological cycle can be strongly influenced by the shared seasonal pattern. I therefore suggest complementing these correlation coefficients with one or two magnitude-based metrics, such as RMSE, mean bias, or differences in seasonal amplitude.

Reply: Thank you for this helpful suggestion. We agree that correlations calculated from a 12-month climatological cycle may be strongly influenced by the shared seasonal pattern. To provide a more comprehensive assessment of agreement among the products, we have therefore added magnitude-based metrics, including RMSE and MAE, to complement the seasonal correlation coefficients.

L366-368: "The estimated NEE seasonal cycle matched well with the OCO-2 v10 MIP (RMSE = 0.11 Pg C month^{-1}, MAE = 0.09 Pg C month^{-1}, $r = 0.96$, $p < 0.001$) and CT2022 inversions (RMSE = 0.13 Pg C month^{-1}, MAE = 0.11 Pg C month^{-1}, $r = 0.97$, $p < 0.001$).".

Lines 334-338: The IAV correlations with CT2022 (r = 0.41, p = 0.07) and OCO-2 v10 MIP (r = 0.70, p = 0.12) are not statistically significant at the p < 0.05 level. Therefore, the statement that MemoryFlux "effectively captured" IAVs should be moderated, particularly for these two comparisons.

Reply: Thank you for this comment. We agree that the original wording was too strong because the correlations with CT2022 and OCO-2 v10 MIP were not statistically significant at the p < 0.05 level. To clarify this point, we have revised the relevant statements in the manuscript accordingly.

L423-428: "MemoryFlux showed positive correlations with OCO-2 v10 MIP ($r = 0.70$, $p = 0.12$), CT2022 ($r = 0.41$, $p = 0.07$), EC-MOD2012 ($r = 0.68$, $p < 0.05$), and FLUXCOM-X ($r = 0.69$, $p < 0.05$), whereas weaker correlations were found with the other datasets. Although the correlations with OCO-2 v10 MIP and CT2022 were not statistically significant, the overall comparison indicated that MemoryFlux, estimated using the LSTM DL approach, captured both the magnitude and IAVs of NEE reasonably well.".