REFEREE #1

We sincerely thank the reviewer for the detailed, precise, and pertinent revision of the manuscript, which allowed us to identify limitations in the original paper, correct reporting errors, and improve the clarity, relevance, and robustness of the message conveyed by this study.

MAJOR COMMENTS

MAJOR COMMENT #1 --

The evaluation in Tables 4 and 5 uses all available data, including data the model was calibrated on. The 100-iteration resampling with 70% subsets reduces but does not solve this problem. There is no truly held-out test set anywhere in the study. For simpler formulations this is less of a concern, but GPP5 has 9 free parameters and ER4 has 6, both substantially more than their baselines, and the reported performance improvements for these models cannot be clearly attributed to better generalization rather than better calibration fit. The authors should either implement a proper train/test split, for example by withholding complete site-years per ecosystem type or provide a much more explicit and quantified discussion of overfitting risk for the higher-complexity formulations.

Answer:

Tables 4 and 5 have been revised to report median values of the performance metrics separately for the training and testing splits (see Tables below). A standard random 70%–30% split of data points was retained, rather than withholding complete site-years, to ensure that each of the 100 optimization runs uses the same number of data points in the training and testing subsets. Because data availability varies both between years within the same site and across sites, withholding entire site-years would introduce uncontrolled differences in sample sizes across runs. Maintaining a fixed number of data points makes sure that performance variability is not driven by sample size variation or inconsistencies in data availability. This still allows the training and testing splits to provide a valuable assessment of model calibration and generalization, respectively.

Nevertheless, the approach mentioned based on withholding complete site-years was also tested and the results are shown in Tables 6 and 7 (see Tables below). In this case, Instead of using 70–30% random splits, the data were divided into 6-year training and 2-year testing periods. The full period of study spans a total of 8 years, yielding 28 possible runs instead of 100.

Overall, the two approaches produce comparable results for both GPP and ER. We therefore propose to retain the standard random 70%–30% splits in the revised manuscript and to replace the original Tables 4 and 5 with those presented here. Section 4.4 (Method) will be updated accordingly and Section 5 (Results) will compare median metrics of the testing splits across model formulations.

Please note that Tables 4, 5, 6 and 7 also report metrics for the site-level (temporal-only) evaluation. For Tables 4 and 5, these values correspond to those shown in the original manuscript in Figure 8.

In addition, results for all 25 possible ER formulations (five GPP formulations combined with five ER formulations) will be included in Appendix.

GROSS PRIMARY PRODUCTION TABLES (TABLE 4 vs. TABLE 6)

Spatiotemporal evaluation

GPP	Splits	Upland Tundra			Taiga Forests			Wetlands		
		r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
GPP_{L4C}	training	0.64	1.47	0.41	0.58	1.93	-0.37	0.48	2.19	1.31
	testing	0.64	1.45	0.42	0.58	1.93	-0.38	0.48	2.19	1.31
GPP_1	training	0.56	1.19	-0.32	0.62	1.75	-0.44	0.53	1.33	-0.40
	testing	0.56	1.20	-0.32	0.61	1.76	-0.45	0.53	1.33	-0.40
GPP_2	training	0.62	0.98	-0.01	0.67	1.43	-0.04	0.63	1.04	-0.05
	testing	0.62	0.99	-0.01	0.66	1.44	-0.04	0.63	1.04	-0.05
GPP_3	training	0.65	0.95	-0.00	0.67	1.41	-0.01	0.65	1.01	-0.01
	testing	0.65	0.96	-0.00	0.67	1.42	-0.02	0.65	1.01	-0.01
GPP_4	training	0.75	0.84	-0.01	0.73	1.30	-0.02	0.72	0.92	-0.01
	testing	0.74	0.84	-0.01	0.73	1.30	-0.02	0.72	0.92	-0.01
GPP_5	training	0.77	0.80	-0.02	0.74	1.29	-0.02	0.68	0.99	-0.05
	testing	0.77	0.80	-0.02	0.74	1.29	-0.02	0.67	0.99	-0.05

Site-level evaluation

GPP	Upland Tundra			Taiga Forests			Wetlands		
	r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
GPP_{L4C}	0.62	1.38	0.51	0.63	1.57	-0.36	0.66	1.62	1.27
GPP_1	0.58	1.11	-0.22	0.61	1.72	-0.57	0.65	1.15	-0.62
GPP_2	0.63	0.91	0.05	0.66	1.33	-0.20	0.71	0.85	-0.22
GPP_3	0.65	0.86	0.08	0.70	1.31	-0.14	0.74	0.79	-0.22
GPP_4	0.77	0.71	0.18	0.79	1.15	-0.13	0.79	0.76	-0.23
GPP_5	0.76	0.65	0.04	0.76	1.12	-0.33	0.76	0.82	-0.19

Table 4. Performance of modeled **gross primary production (GPP)** against daily averaged eddy covariance (EC) GPP (GPP_{EC}) for upland tundra, taiga forests, and wetlands during the growing season. GPP_{L4C} refers to outputs from the original L4C model, while GPP_1 through GPP_5 correspond to the five Arctic–Subarctic (AS) adapted formulations (Sections 3 and 4.1). The Pearson correlation coefficient is denoted by r (dimensionless), and ubRMSE and B denote the unbiased root mean square error and bias, respectively (in $\mathrm{gCm^{-2}d^{-1}}$). A positive (negative) B indicates overestimation (underestimation) of GPP_{EC}. **Upper table:** Spatiotemporal performance metrics derived from 100 random splits (70 % training, 30 % testing), accounting for both spatial and temporal variability. The total number of data points used for training (testing) is 1,155 (495) for upland tundra, 3,243 (1,389) for taiga forests, and 2,557 (1,096) for wetlands (Section 4.3). Median values across the 100 splits are reported. **Lower table:** Site-level (temporal-only) performance metrics computed for each EC tower using model outputs obtained with the median parameter values across the 100 splits. Metric values are reported as the median across towers. For this analysis, training and testing data are pooled.

Spatiotemporal evaluation

GPP	Splits	Upland Tundra			Taiga Forests			Wetlands		
		r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
GPP_{L4C}	training	0.64	1.46	0.43	0.58	1.92	-0.37	0.46	2.21	1.31
	testing	0.65	1.45	0.36	0.60	1.89	-0.40	0.54	2.10	1.29
GPP_1	training	0.56	1.20	-0.33	0.62	1.75	-0.44	0.52	1.34	-0.41
	testing	0.56	1.20	-0.30	0.61	1.82	-0.52	0.57	1.25	-0.42
GPP_2	training	0.63	0.98	-0.01	0.67	1.43	-0.04	0.62	1.05	-0.05
	testing	0.62	0.98	0.01	0.66	1.44	-0.11	0.66	0.97	-0.10
GPP_3	training	0.65	0.96	-0.00	0.67	1.41	-0.02	0.64	1.02	-0.01
	testing	0.65	0.97	0.04	0.66	1.42	-0.15	0.69	0.96	-0.04
GPP_4	training	0.75	0.83	-0.01	0.73	1.30	-0.02	0.71	0.94	-0.01
	testing	0.74	0.84	-0.02	0.72	1.30	-0.07	0.77	0.84	-0.05
GPP_5	training	0.77	0.79	-0.02	0.74	1.28	-0.02	0.66	1.00	-0.05
	testing	0.76	0.81	-0.03	0.73	1.29	-0.01	0.72	0.92	-0.17

Site-level evaluation

GPP	Upland Tundra			Taiga Forests			Wetlands		
	r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
GPP_{L4C}	0.62	1.38	0.51	0.63	1.57	-0.36	0.66	1.62	1.27
GPP_1	0.58	1.11	-0.22	0.60	1.73	-0.55	0.65	1.16	-0.63
GPP_2	0.63	0.91	0.06	0.67	1.34	-0.20	0.71	0.85	-0.28
GPP_3	0.65	0.86	0.08	0.70	1.31	-0.17	0.74	0.78	-0.24
GPP_4	0.77	0.71	0.19	0.79	1.15	-0.11	0.79	0.76	-0.22
GPP_5	0.76	0.66	0.08	0.76	1.12	-0.35	0.76	0.81	-0.20

Table 6. Same as Table 4, but instead of 100 runs with 70–30 % training–testing splits, the data were split into 6-year training and 2-year testing periods. The period of study spans a total of 8 years, yielding 28 possible runs instead of 100.

ECOSYSTEM RESPIRATION TABLES (TABLE 5 vs. TABLE 7)

Spatiotemporal evaluation										
ER	Splits	Upland Tundra			Taiga Forests			Wetlands		
		r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
ER_{L4C}	training	0.43	0.99	0.39	0.33	1.71	-0.11	0.23	1.81	1.76
	testing	0.44	0.99	0.37	0.34	1.71	-0.12	0.24	1.80	1.77
ER_1	training	0.52	0.72	-0.10	0.51	1.28	-0.15	0.50	0.80	-0.04
	testing	0.52	0.72	-0.12	0.52	1.28	-0.17	0.49	0.81	-0.05
ER_2	training	0.61	0.63	-0.11	0.53	1.25	-0.12	0.53	0.78	-0.02
	testing	0.61	0.64	-0.12	0.54	1.25	-0.12	0.53	0.78	-0.03
ER_3	training	0.60	0.62	-0.00	0.52	1.23	0.00	0.51	0.79	-0.02
	testing	0.60	0.62	-0.01	0.54	1.23	0.02	0.50	0.80	-0.03
ER_4	training	0.73	0.53	-0.00	0.58	1.17	0.00	0.51	0.79	-0.02
	testing	0.72	0.53	-0.03	0.59	1.17	0.02	0.51	0.80	-0.04
ER_5	training	0.72	0.54	-0.01	0.59	1.17	0.01	0.50	0.80	-0.04
	testing	0.70	0.55	-0.02	0.60	1.17	0.03	0.49	0.81	-0.05

Site-level evaluation									
ER	Upland Tundra			Taiga Forests			Wetlands		
	r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
ER_{L4C}	0.45	0.94	0.35	0.56	1.18	-0.28	0.43	1.10	1.24
ER_1	0.58	0.51	0.06	0.61	0.96	-0.15	0.65	0.56	-0.12
ER_2	0.66	0.46	0.03	0.65	0.95	-0.09	0.66	0.53	-0.09
ER_3	0.57	0.47	0.08	0.65	0.96	0.00	0.67	0.51	-0.18
ER_4	0.58	0.41	-0.01	0.65	1.00	-0.05	0.64	0.48	-0.15
ER_5	0.60	0.42	0.00	0.63	1.01	-0.13	0.63	0.51	-0.16

Table 5. Same as Table 4, but for **ecosystem respiration (ER)**. **GPP_5** was used as the GPP input for ER_1 through ER_5 for upland tundra and taiga forests. For wetlands, **GPP_4** was used instead.

Spatiotemporal evaluation										
ER	Splits	Upland Tundra			Taiga Forests			Wetlands		
		r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
ER_{L4C}	training	0.42	0.99	0.39	0.33	1.71	-0.11	0.23	1.80	1.79
	testing	0.49	0.96	0.38	0.35	1.71	-0.12	0.27	1.79	1.68
ER_1	training	0.52	0.71	-0.10	0.52	1.27	-0.15	0.49	0.80	-0.04
	testing	0.54	0.70	-0.13	0.51	1.28	-0.15	0.50	0.83	-0.02
ER_2	training	0.62	0.63	-0.11	0.53	1.24	-0.11	0.53	0.78	-0.02
	testing	0.65	0.57	-0.15	0.52	1.25	-0.10	0.54	0.78	0.01
ER_3	training	0.60	0.62	-0.00	0.53	1.23	0.00	0.50	0.79	-0.02
	testing	0.60	0.61	-0.04	0.52	1.23	0.02	0.53	0.80	0.00
ER_4	training	0.71	0.55	-0.01	0.59	1.17	0.00	0.50	0.79	-0.02
	testing	0.76	0.48	0.00	0.58	1.17	0.04	0.55	0.81	-0.01
ER_5	training	0.70	0.56	-0.01	0.59	1.17	0.00	0.49	0.80	-0.04
	testing	0.73	0.51	-0.02	0.59	1.16	0.02	0.55	0.81	-0.02

Site-level evaluation									
ER	Upland Tundra			Taiga Forests			Wetlands		
	r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
ER_{L4C}	0.45	0.94	0.35	0.56	1.18	-0.28	0.43	1.10	1.24
ER_1	0.57	0.51	0.11	0.60	0.96	0.02	0.64	0.58	0.02
ER_2	0.66	0.46	0.03	0.64	0.96	0.00	0.66	0.52	-0.10
ER_3	0.58	0.45	0.13	0.64	0.97	0.05	0.67	0.53	-0.04
ER_4	0.58	0.41	-0.08	0.65	1.01	-0.11	0.64	0.48	-0.13
ER_5	0.62	0.41	-0.03	0.63	1.02	-0.13	0.63	0.50	-0.16

Table 7. Same as Table 5, but instead of 100 runs with 70–30 % training–testing splits, the data were split into 6-year training and 2-year testing periods. The period of study spans a total of 8 years, yielding 28 possible runs instead of 100.

MAJOR COMMENT #2 --

There is a deeper conceptual issue worth flagging. The entire calibration and evaluation framework assumes that GPPEC and EREC from flux partitioning are reliable targets. But these quantities are not measured; they are modeled from NEE using methods that themselves rely on temperature and light as primary drivers. This means the AS-adapted formulations are being trained to reproduce outputs of algorithms that share some of the same structural assumptions as the L4C model itself. The strong performance of APAR and GDD adjustments may partly reflect this shared structure rather than genuine independent improvement. The authors touch on this in section 6.5, but do not go far enough. It is worth asking openly whether the model is getting better at capturing carbon dynamics or simply getting better at agreeing with a partitioning algorithm it already resembles.

Answer:

We agree with the reviewer that the issue raised was not enough discussed in the original manuscript. The consistency between flux-partitioning algorithms and modeling frameworks is an essential topic that we intend to address specifically in the coming year.

Section 6.5 will be revised to explicitly acknowledge that

(1) GPP_{EC} and ER_{EC} are not direct measurements but modeled outputs;

(2) there is potential circularity in model evaluation as the AS-adapted formulations may share similar structural assumptions with the flux-partitioning algorithms;

(3) model improvement may reflect better reproduction of flux-partitioning algorithms rather than better representation of carbon dynamics;

(4) there are substantial conceptual differences between flux-partitioning and mechanistic modeling frameworks (such as the L4C model).

A draft of the revised text is provided below:

6.6 Limitations

The reference GPP_{EC} and ER_{EC} used for calibration and evaluation are derived from NEE_{EC} partitioning, meaning they are not direct measurements but modeled outputs based on NEE_{EC} and structural assumptions (Appendix A). This creates a potential circularity in model evaluation, as the AS-adapted formulations may share the same structural assumptions as the flux-partitioning algorithms. Consequently, improvements in modeled ER and GPP may partly reflect the model formulations reproducing the behavior of the flux-partitioning algorithms, rather than independently improving the representation of carbon dynamics. In some cases, GPP_{EC} is constrained to follow a light-response curve, which is why a similar adjustment was tested in GPP_2 (Equation 8). At first glance, this structural similarity likely explains why the adjustment enhanced performance (Section 6.1). However, in other cases, GPP_{EC} is not directly modeled, but derived as the residual between NEE_{EC} and ER_{EC}. Therefore, it is difficult to determine whether improvements from GPP_2 reflect better reproduction of specific flux-partitioning algorithms, a more accurate representation of the true flux dynamics, or a combination of both. In contrast, GDD is not used at all in flux-partitioning algorithms (Appendix A). Therefore, the improvements resulting from the inclusion of GDD in GPP_4 may capture true ecosystem state changes that are also well represented in GPP_{EC}.

Regarding ER, the situation is more complex. ER_{EC} is typically derived from a fitted power-based or exponential-based function (Appendix A). These functions depend solely on temperature and estimate the combined contribution of AR and HR as a single inseparable flux. In contrast, the L4C model and the tested AS-adapted formulations explicitly represent ER as the sum of AR and HR, with each component estimated separately using multiple drivers, including APAR, GDD, MNT, VPD, RZSM, ST, and SSM. This approach relies on assumed linkages between GPP and AR, and between GPP, L_{fall}, SOC, and HR (Kimball et al., 2008), resulting in a more mechanistic, interaction-rich framework than the flux-partitioning algorithms. Consequently, calibrating ER formulations is challenging, because the reference ER_{EC} is obtained using a simpler empirical approach, which may limit model performance. If the ultimate goal is to estimate the CO_2 budget accurately, rather than to predict the underlying GPP and ER components, it may be advantageous to calibrate the L4C model using NEE_{EC} as the reference, rather than relying on GPP_{EC} and ER_{EC} as intermediate targets. However, this approach prevents validating whether the modeled GPP and ER truly reflect the underlying processes and strongly limits the number of free parameters that can be estimated, since only a single reference is available instead of two. For future research, it could also be valuable to partition NEE_{EC} into GPP_{EC} and ER_{EC} using a more mechanistic approach similar to the L4C model, explicitly distinguishing between AR and HR.

In summary, when calibrating TCF models using GPP_{EC} and ER_{EC} as references, one attempts to explain variability in fluxes that originate from a reference framework with a relatively simple structure, a limited number of drivers, and parameters that may vary in space and time (Appendix A). In contrast, TCF models, such as the L4C model, rely on a more complex process representation, a larger set of environmental drivers, and parameters that are assumed to be constant in space and time within a given ecosystem. These fundamental differences inherently complicate model calibration, hinder the interpretation of model performance, and limit our ability to determine whether the underlying processes of ER and GPP are realistically represented when extrapolated to larger spatial and temporal scales.

MAJOR COMMENT #3 ---

Equation 9c appears to contain a typographical error. The denominator of the SRZSM logistic ramp reads g(MNTmin), which looks like a copy-paste error from the SMNT equation directly above it. It should presumably read g(RZSMmin). This needs to be verified and corrected, since it directly affects reproducibility of the RZSM formulation in GPP3.

Answer:

We thank the reviewer for identifying this issue. This was a copy-paste error; $g(MNT_{min})$ will be replaced by $g(RZSM_{min})$ in Equation 9c.

MAJOR COMMENT #4 ---

The scoring system in Equation 14 is difficult to defend as currently described. Ranks are used instead of raw metric values, which throws away information about how much better one formulation is relative to another. A formulation that narrowly beats another gets the same rank benefit as one that beats it by a wide margin. The penalty factor is a simple linear ratio of parameter counts applied as a multiplier on the average rank, with no comparison to established model selection criteria such as AIC or BIC. The authors should either justify this design explicitly or test whether the final formulation selection changes under alternative scoring approaches.

Answer:

Originally, we were uncertain whether the scoring procedure described in Equation 14 was appropriate for model selection. We therefore compared the resulting rankings with those obtained using alternative criteria based on AIC and BIC as recommended:

Rankings for GPP:

(1) Upland tundra

- Equation 14: $\mathbf{GPP_5}$, GPP_4, GPP_3, GPP_2, GPP_1 `
- AIC/BIC: $\mathbf{GPP_5}$, GPP_4, GPP_3, GPP_2, GPP_1

(2) Taiga forests

- Equation 14: $\mathbf{GPP_5}$, GPP_4, GPP_3, GPP_2, GPP_1
- AIC/BIC: $\mathbf{GPP_5}$, GPP_4, GPP_3, GPP_2, GPP_1

(3) Wetlands

- Equation 14: $\mathbf{GPP_4}$, GPP_3, GPP_5, GPP_2, GPP_1
- AIC/BIC: $\mathbf{GPP_4}$, GPP_5, GPP_3, GPP_2, GPP_1

Rankings for ER:

(1) Upland tundra

- Equation 14: $\mathbf{ER_4}$, ER_3, ER_5, ER_2, ER_1 `
- AIC/BIC: $\mathbf{ER_4}$, ER_5, ER_3, ER_2, ER_1

(2) Taiga forests

- Equation 14: $\mathbf{ER_4}$, ER_5, ER_3, ER_2, ER_1
- AIC/BIC: $\mathbf{ER_5}$, ER_4, ER_3, ER_2, ER_1

(3) Wetlands

- Equation 14: $\mathbf{ER_3}$, ER_2, ER_4, ER_5, ER_1
- AIC/BIC: $\mathbf{ER_2}$, ER_3, ER_4, ER_5, ER_1

For GPP, the best-performing formulations were consistent across the two approaches: GPP_5 ranked first for upland tundra and taiga forests, while GPP_4 ranked first for wetlands. The only difference was observed for wetlands, where GPP_3 and GPP_5 ranked second and third using Equation 14, but third and second based on AIC/BIC.

For ER, more differences were observed. However, overall, ER_4 or ER_5 ranked first and second for upland tundra, and taiga forest, whereas ER_2 and ER_3 performed best for wetlands.

The comparison of the scoring approaches revealed that AIC/BIC provide a more continuous assessment of model performance than Equation 14, allowing a clearer evaluation of how close competing formulations are. This is particularly evident for wetlands, where all ER formulations have similar AIC/BIC values. As the reviewer noted, such a result could not be clearly identified using Equation 14.

As a consequence, we will remove the original scoring procedure and replace it with AIC and BIC in the revised manuscript. Sections 4.4 and 5 will be updated accordingly. In addition, ranks will be reported in Appendix (see Tables C1 and D6 below), instead of being omitted (as it was done in the original manuscript).

GPP	Upland Tundra		Taiga Forests		Wetlands	
	ΔAIC	ΔBIC	ΔAIC	ΔBIC	ΔAIC	ΔBIC
GPP_1	-482	-487	-574	-580	-3116	-3122
GPP_2	-1010	-1010	-2036	-2036	-4592	-4592
GPP_3	-1081	-1081	-2134	-2134	-4734	-4734
GPP_4	-1380	-1369	-2682	-2669	-5200	-5188
GPP_5	-1486	-1476	-2737	-2725	-4847	-4836

Table C1. Median differences in Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) between GPP_1 through GPP_5 and GPP_{L4C}. GPP_{L4C} refers to GPP outputs from the original L4C model, while GPP_1 through GPP_5 represent the five Arctic–Subarctic (AS) adapted formulations (Sections 3 and 4.1).

ER	Upland Tundra		Taiga Forests		Wetlands	
	ΔAIC	ΔBIC	ΔAIC	ΔBIC	ΔAIC	ΔBIC
			GPP_1 as input			
ER_1	-331	-336	-25	-31	-5138	-5144
ER_2	-977	-982	-1363	-1369	-5973	-5978
ER_3	-1174	-1184	-1916	-1929	-5786	-5797
ER_4	-1511	-1516	-2246	-2252	-5775	-5781
ER_5	-1484	-1489	-2337	-2344	-5600	-5606
			GPP_2 as input			
ER_1	-742	-747	-1393	-1400	-5700	-5706
ER_2	-1015	-1020	-1732	-1738	-6032	-6038
ER_3	-1187	-1197	-1990	-2002	-5849	-5861
ER_4	-1509	-1514	-2290	-2296	-5873	-5879
ER_5	-1510	-1515	-2380	-2386	-5783	-5789
			GPP_3 as input			
ER_1	-811	-816	-1433	-1439	-5740	-5746
ER_2	-1056	-1061	-1744	-1750	-5922	-5928
ER_3	-1210	-1221	-1977	-1990	-5835	-5846
ER_4	-1540	-1546	-2277	-2283	-5899	-5905
ER_5	-1553	-1558	-2367	-2373	-5830	-5836
			GPP_4 as input			
ER_1	-729	-734	-1621	-1627	-5853	-5859
ER_2	-1062	-1067	-1931	-1937	-6007	-6013
ER_3	-1210	-1220	-2122	-2135	-5944	-5955
ER_4	-1602	-1607	-2395	-2401	-5936	-5942
ER_5	-1606	-1611	-2498	-2504	-5868	-5873
			GPP_5 as input			
ER_1	-881	-886	-1844	-1850	-5774	-5780
ER_2	-1185	-1190	-2030	-2036	-5976	-5982
ER_3	-1269	-1279	-2146	-2158	-5863	-5874
ER_4	-1613	-1619	-2455	-2461	-5840	-5846
ER_5	-1577	-1582	-2482	-2488	-5742	-5748

Table D6. Same as Table C1, but for **ecosystem respiration (ER)**.

MAJOR COMMENT #5 --

For wetlands, the temporal correlation of NEE drops from 0.50 in the original L4C model to 0.33 in the AS-adapted formulation. This is not a small degradation. It means the adapted model tracks wetland carbon exchange less accurately in time than the model it is supposed to improve. The authors mention error compensation in section 6.5 but do not examine which sites or years drive this result, or whether the scoring-based selection of GPP4 and ER3 for wetlands is itself contributing to the problem. This finding needs a dedicated discussion, not a brief mention.

Answer:

We sincerely thank the reviewer for pointing out this drop in correlation. While investigating the cause of this discrepancy, we identified an error in the code used to compute the NEE configurations. This issue did not affect only the correlation of NEE_{AS} for wetlands, but the entire row in Table 6 (of the original manuscript) reporting NEE_{AS} performance was incorrect (r, ubRMSE, and B for all three ecosystems). Specifically, we had incorrectly paired the GPP and ER components, using:

$NEE_{i,j} = ER_i(GPP_j) - GPP_{\mathbf{i}}$

instead of

$NEE_{i,j} = ER_i(GPP_j) - GPP_{\mathbf{j}}$

where $ER_i(GPP_j)$ denotes the i^{th} ER formulation using the j^{th} GPP formulation as input.

As a result, NEE_{AS} was computed as $ER_4(GPP_5) - GPP_4$ for upland tundra and taiga forests, and as $ER_3(GPP_4) - GPP_3$ for wetlands. This inconsistency was not apparent for upland tundra and taiga forests, as GPP_5 and GPP_4 are broadly similar. However, there is a clear difference in performance between GPP_3 and GPP_4 due to the inclusion of GDD, which primarily affects correlation. Consequently, GPP_3 was not in phase with $ER_3(GPP_4)$, leading to an inconsistent configuration that is not physically meaningful. The reported values should have been as follows:

NEE	Upland Tundra			Taiga Forests			Wetlands		
	r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
NEE_{L4C}	0.55 (0.44)	0.86 (0.90)	-0.03 (0.06)	0.55 (0.54)	1.17 (1.04)	0.26 (0.40)	0.47 (0.50)	0.99 (0.79)	0.45 (0.73)
NEE_{AS}	0.64 (0.64)	0.73 (0.68)	0.01 (0.06)	0.63 (0.59)	1.09 (0.96)	0.02 (0.04)	0.48 (0.59)	0.93 (0.72)	0.01 (0.17)

Table 6. Performance of modeled **net ecosystem CO_2 exchange (NEE)** against daily-averaged eddy covariance (EC) NEE (NEE_{EC}) in upland tundra, taiga forests, and wetlands during the growing season. Modeled NEE is computed as the difference between modeled ecosystem respiration (ER) and gross primary production (GPP). NEE_{L4C} denotes NEE from the original L4C model, while NEE_{AS} refers to outputs from the Arctic–Subarctic (AS) adapted formulations. For upland tundra and taiga forests, $NEE_{AS} = ER_4 - GPP_5$, while for wetlands $NEE_{AS} = ER_3 - GPP_4$. The formulation selection was based on their performance scores (Section 4.4). Refer to Sections 3 and 4.1 for a description of each model formulation. ubRMSE and B denote the unbiased root mean squared error and bias, respectively, expressed in $\mathrm{gCm^{-2}d^{-1}}$. A positive (negative) B indicates that the model formulation overestimates (underestimates) NEE_{EC}. r is the Pearson correlation coefficient (dimensionless). Values reported outside parentheses represent **spatiotemporal performance**, accounting for both spatial and temporal variability. This evaluation includes all available NEE_{EC} data points for each ecosystem type after filtering (1,650 for upland tundra, 4,632 for taiga forests and 3,653 for wetlands; Section 4.3). Values in parentheses represents **temporal performance**, shown as the median metrics across EC towers.

Performance for upland tundra and taiga forests does not change much and the correlation is no longer degraded for wetlands. Please refer to Major Comment #8 for additional details on the performance of the NEE configurations.

MAJOR COMMENT #6 --

ER1 and ER2 use a single SOC pool rather than the original three, because reference SOC data were not available for recalibration. This is a reasonable practical decision but it means the comparison between ER1/ER2 and ERL4C is not a clean test of the Lfall allocation scheme. It simultaneously tests a different pool structure. The authors should acknowledge this confounding more clearly in the methods and in the discussion of section 6.2.

Answer:

Section 4.3 will be modified to highlight the reduction of SOC pools as an additional adjustment:

4.3 Model formulation calibration

The AS-adapted GPP and ER formulations were calibrated separately for each ecosystem type (upland tundra, taiga forests, and wetlands), using GPP_{EC} and ER_{EC} as reference targets. The optimization framework for calibration used least-squares minimization via the MATLAB lsqcurvefit function (MathWorks, Inc., 2023), which minimizes the sum of squared differences between the model outputs and target values. This is an unconstrained optimization, with no additional penalty terms applied. To mitigate overfitting, the optimization process was repeated 100 times using different random training (70 %) and testing (30 %) subsets of the data for each ecosystem type: 1,155 (495) data points for upland tundra, 3,243 (1,389) for taiga forests and 2,557 (1,096) for wetlands. Final model parameter values were taken as the median across all runs. This approach aimed to capture diverse data combinations and promote a more stable and representative parameterization by smoothing out the influence of outliers or any individual biased subset. A 70 % subset size was arbitrarily chosen to balance between providing sufficient data for robust model calibration and retaining enough data variability across the 100 iterations (Martinez Molera, 2025). Contrary to the global calibration of the original L4C model, no reference SOC data were used to constrain the recalibration over the AS environments. As a result, in ER_1 and ER_2, only the SOC_1 pool was modeled to avoid potential parameter compensation and to prevent unrealistic SOC distribution across the original three pools. This reduction in the number of SOC pools constitutes an additional adjustment for ER_1 and ER_2 that is confounded with the change in L_{fall} estimation scheme (Section 4.1). An initial guess was necessary for SOC_1 on March 31, 2015, to explicitly solve the SOC dynamics, since the system is formulated recursively and requires a starting value to iterate forward in time (Equation 4). March 31, 2015, was chosen as the start of the simulation because it precedes the first date of the period of study. The initial guess was set to 0 gCm^{-2} for the first spin-up iteration, providing a neutral starting point to avoid biasing the early simulation. It was subsequently updated using the SOC value on March 31, 2022, corresponding to the last March 31 within the study period. A total of 20 spin-up iterations were performed.

Section 6.2 will also be modified to underline the confounding effects of the SOC pool reduction, litterfall scheme, and GPP input:

6.2 Comparison of SOC-based and empirical approaches for ER modeling

Updating the allocation of mean annual NPP to L_{fall} from a constant to a LAI-based formulation to represent SOC dynamics (ER_1 vs. ER_2; Equations 5 and 6) improves ER model performance. Both the spatiotemporal evaluation and the median metrics across EC towers indicate higher r and lower ubRMSE and B, with stronger improvements for upland tundra and weaker improvements for taiga forests and wetlands (Table 5). The benefits are limited relative to the added model complexity,

especially in taiga forests and wetlands, compared with the simpler approach that replaces SOC dynamics with a single constant R_{base} (ER_3, Equation 12). Based on the spatiotemporal evaluation, introducing temporal variability in R_{base} (ER_4, Equation 13) leads to improved performance in upland tundra and taiga forests (Table 5). However, the median metrics across EC towers do not indicate a clear improvement across the three ecosystems (Table 5). Compared to upland tundra and taiga forests, all ER formulations are highly similar for wetlands, regardless of the metrics considered (r, ubRMSE, B, ΔAIC, and ΔBIC) and the type of evaluation (spatiotemporal vs. site-level; Tables 5 and D6).

Overall, using SOC dynamics with the L_{fall} estimation scheme from the L4C model version 8 to model ER appears to be the most suitable approach, as it performs better than version 7 and is physically grounded and mechanistically interpretable compared with the two empirical approaches. Unfortunately, the improved performance of ER_1 and ER_2 compared to ER_{L4C} is difficult to interpret, as it reflects the combined effects of changes in the L_{fall} estimation scheme, SOC pool structure, and GPP input (Sections 4.1, 4.3). Therefore, this comparison does not constitute a clean test of the added-value of the L_{fall} allocation scheme version 8 compared to version 7 for the study regions.

Continuing to explore alternative ways to estimate L_{fall} may be a promising direction for future research. However, the assumption that mean annual NPP can serve as a proxy for the magnitude of L_{fall} may not be realistic (Sierra et al., 2022). In addition, the timing of NPP allocation to L_{fall} may not accurately represent litter production dynamics, particularly given the large uncertainties in LAI and FPAR retrievals at high northern latitudes (Xu et al., 2018; Pu et al., 2023). Furthermore, because NPP is derived from modeled GPP, any inaccuracies in GPP propagate directly into modeled L_{fall}, SOC, and ultimately ER. Finally, recent work in Alaska has shown that implementing vertical SOC transport to simulate depth-dependent L_{fall}, SOC distribution, and corresponding HR rates may further improve ER estimates (Yi et al., 2020).

MAJOR COMMENT #7 ---

GDD is normalized annually per site using each year's own minimum and maximum values. For data points early in the growing season, this normalization requires information from later in the same year. The model's seasonal shape scalar for April implicitly uses what happened in August. The authors should clarify whether this creates an information leakage issue in the evaluation and whether a climatological or cross-year normalization was considered.

Answer:

We acknowledge that this use of GDD requires the full annual air temperature cycle to be known, introducing potential information leakage and a one-year lag in the computation of model outputs, which makes it unsuitable for real-time forecasting. However, the primary objective of this study is not prediction in an operational setting, but rather the retrospective reconstruction of CO_2 fluxes and the estimation of CO_2 budgets. In this context, we consider the use of full-year information to be appropriate.

Nevertheless, we tested an alternative normalization where GDD is normalized using long-term minimum and maximum values, which avoids within-year information leakage. We found that model performance was comparable, or only slightly lower, under this approach. These results suggest that the choice of GDD normalization has a limited influence on the performance improvements associated with adding GDD as an input into GPP modeling.

This test will be incorporated in the revised manuscript in Appendix (see Table C2 below), and Section 6.1 will be updated to explicitly discuss this point:

Spatiotemporal evaluation

GPP	Splits	Upland Tundra			Taiga Forests			Wetlands		
		r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
GPP_4	training	0.73	0.86	-0.01	0.73	1.32	-0.01	0.72	0.92	-0.00
	testing	0.73	0.86	0.00	0.72	1.33	-0.02	0.72	0.92	-0.00
GPP_5	training	0.74	0.83	-0.03	0.73	1.30	-0.01	0.67	0.99	-0.04
	testing	0.74	0.84	-0.01	0.73	1.30	-0.02	0.67	0.99	-0.05

Site-level evaluation

GPP	Upland Tundra			Taiga Forests			Wetlands		
	r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
GPP_4	0.77	0.76	0.16	0.78	1.19	-0.12	0.76	0.77	-0.22
GPP_5	0.76	0.73	0.02	0.76	1.17	-0.34	0.74	0.82	-0.19

Table C2. Same as Table 4, but here GPP_4 and GPP_5 use growing degree days (GDD) as inputs, which were normalized by the long-term minimum and maximum across all years rather than by annual values (Section 6.1).

Finally, it is noteworthy that GDD was normalized using annual minimum and maximum values, implying that early-season GDD values implicitly depend on values from later in the same year. This approach therefore requires the full annual air temperature cycle to be known, introducing a one-year lag in the computation of model outputs that is not appropriate for real-time forecasting. However, the primary objective of this study is not prediction in an operational setting, but rather the retrospective reconstruction of CO_2 fluxes and the estimation of CO_2 budgets. For potential forecasting applications, GDD can instead be normalized using long-term minimum and maximum values across all years, resulting in comparable or slightly lower performance (see GPP_4 and GPP_5 in Table 4 vs. Table C2). This alternative approach avoids any within-year information leakage and suggests that GDD normalization has a limited influence on the overall performance improvements associated with adding GDD as an input.

6.2 Comparison of SOC-based and empirical approaches for ER modeling

MAJOR COMMENT #8 ---

GPP and ER formulations are selected independently based on their individual performance scores and then combined to produce NEE. But minimizing GPP error and minimizing ER error separately does not guarantee minimizing NEE error, since the two error terms can be correlated or offsetting. The authors acknowledge this briefly in section 6.5 but do not test whether selecting formulations based directly on NEE performance would produce different combinations or better results, particularly for wetlands where the current approach visibly underperforms.

Answer:

As recommended by the reviewer, we evaluated all 25 possible configurations for NEE. Overall, we found that the best-performing GPP and ER formulations, once combined, give the best, or among the best, NEE configurations. NEE performance is more driven by the GPP formulation than the ER formulation with GPP_4 and GPP_5 leading to higher r and lower ubRMSE. In contrast no clear and consistent pattern emerges regarding the impact of ER formulations on NEE performance, with the exception of ER_1, that lead to increased B and ubRMSE.

In the revised manuscript, Section 4.4 will be updated to explicitly indicate that the 25 ER and NEE configurations were tested and the corresponding results will be added in Appendix. Results in Section 5.3 will also be updated. Table 6 will be removed and replaced by Figure 8 (see below), that displays summarized performance for all NEE configurations (instead of just one initially in Table 6 of the orignial manuscript). Finally, the discussion regarding NEE will be updated in Section 6.5.

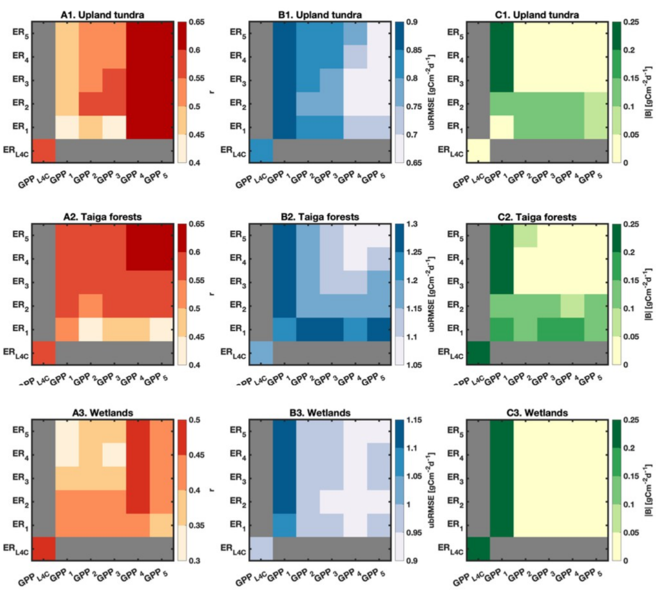


Figure 8. Spatiotemporal performance of modeled **net ecosystem CO_2 exchange (NEE)** against daily-averaged eddy covariance NEE (NEE_{EC}) for **upland tundra**, **taiga forests**, and **wetlands** during the growing season. NEE is computed as the difference between ecosystem respiration (ER) and gross primary production (GPP). GPP_{L4C} and ER_{L4C} refer to outputs from the original L4C model, while GPP_1 through GPP_5 and ER_1 through ER_5 correspond to the Arctic–Subarctic (AS) adapted formulations (Sections 3 and 4.1). **The (ER_i, GPP_j) grid cell corresponds to the $NEE_{i,j}$ configuration**. The Pearson correlation coefficient is denoted by r (dimensionless), and ubRMSE and |B| denote the unbiased root mean square error and absolute bias, respectively (in $\mathrm{gCm^{-2}d^{-1}}$). Spatiotemporal metrics were derived from 100 random splits (70 % training, 30 % testing), accounting for both spatial and temporal variability. The total number of data points used for training (testing) is 1,155 (495) for upland tundra, 3,243 (1,389) for taiga forests, and 2,557 (1,096) for wetlands (Section 4.3). Reported values correspond to median metrics computed over the testing splits (Tables E1-E5).

MINOR COMMENTS

MINOR COMMENT #1 --

Comment: The term "AT" appears in section 6.4 referring to air temperature but is not defined in the abbreviation table and does not appear anywhere else in the paper. The manuscript consistently uses MNT for minimum air temperature throughout. This should be made consistent.

Answer: In the revised manuscript, the acronym “AT” will be introduced in Section 4.1 when presenting the GPP_4 formulation. AT corresponds to the daily mean air temperature used to calculate growing degree days (GDD), in contrast to MNT, which represents the minimum daily air temperature used in the GPP instantaneous temperature response (S_{MNT}). The definition of “AT” will be added to Table 2.

MINOR COMMENT #2 --

Comment: Figure 8 is cited repeatedly in sections 5.1 and 5.2 without specifying which panel is being referred to. With six subpanels across three ecosystem types, the reader is left guessing. Panel-level citations, for example Figure 8A1 or Figure 8B2, would help significantly.

Answer:

Results initially presented in Figure 8 will be moved to Tables 4 and 5 (see Major Comment #1). Figure 8 will still exist but will be used for another purpose (NEE performance; see Major Comment #8).

MINOR COMMENT #3 --

Comment: line 452. "does not provides any benefits neither" should read "does not provide any benefits either."

Answer:

This error will be corrected in the revised manuscript.

MINOR COMMENT #4 --

Comment: the 10th percentile threshold used to exclude shoulder-season data in section 4.2 is described by the authors themselves as arbitrary. A brief note on sensitivity to this threshold, or at least a reference to comparable choices in prior work, would add confidence.

Answer:

The growing season is more commonly defined using fixed fractions of annual maximum GPP rather than percentile-based thresholds. However, a previous study has shown that the growing season identification is sensitive to the chosen fraction of annual maximum GPP (Panwar 2023). Therefore we chose to use a distribution-based threshold as this is less sensitive to extreme values and inter-annual variability. For instance, two years with the same growing season timing (i.e., identical start and end dates) but different photosynthetic peaks are treated identically, which would not be the case when using the annual maximum GPP. Although the 10^{th} percentile threshold may not be optimal for each time series, it is consistent with the structure of the mean seasonal cycle of GPP, where it broadly separates winter and shoulder seasons from the growing season (see Figure B2 below). Similar separations are obtained when using thresholds within the 5th–15th percentile range, indicating low sensitivity to the chosen percentile. We therefore opted for a common threshold for the three ecosystem types.

As noted, the 10th percentile threshold was described as arbitrary in the original manuscript because different percentiles could have been selected with little influence on the resulting filtering.

Section 4.2 will be updated in the revised manuscript to clarify the choice of the threshold (see below) and Figure B2 will be added in Appendix for transparency.

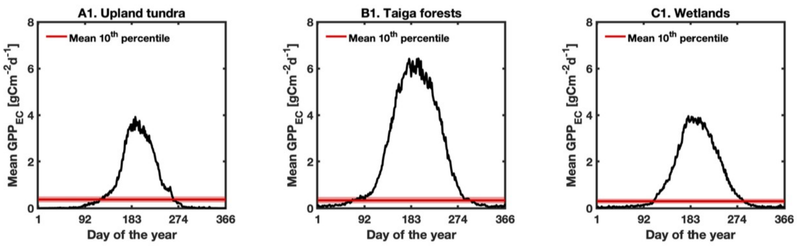


Figure B2. Mean seasonal cycle of daily averaged eddy covariance (EC) gross primary production (GPP_{EC}) for upland tundra, taiga forests and wetlands. The horizontal red line indicates the estimated separation between dormant and transitional periods with active photosynthetic periods (i.e., the growing season) using a percentile-based GPP_{EC} threshold (Section 4.2). This threshold corresponds to the mean 10^{th} percentile computed across EC towers, with the 5^{th} and 15^{th} percentiles shown as variability bounds.

4.2 Growing season timing and data filtering

GPP_{EC} was used as an indicator to identify the growing season (Gonsamo et al., 2013). For each EC tower, GPP_{EC} values below the noise threshold of 0.05 $\mathrm{gCm^{-2}d^{-1}}$ were first attributed to the winter season and removed. Among the remaining values, those below the 10^{th} percentile were considered part of the shoulder seasons (i.e., transitional periods between fully frozen and fully thawed states) and were excluded. The growing season is more commonly defined using fixed fractions of annual maximum GPP_{EC} rather than percentile-based thresholds (Panwar et al., 2023; Luo et al., 2025). However, such approaches rely on the assumption that the annual maximum GPP_{EC} is robustly captured each year, which may not hold due to data gaps for instance. In addition, a previous study has shown that the growing season identification is sensitive to the chosen fraction of annual maximum GPP_{EC}, indicating a lack of consensus on an optimal threshold (Panwar et al., 2023). In contrast, the percentile-based approach used here provides a distribution-based criteria that is less sensitive to extreme values and inter-annual variability, ensuring consistency across years. For instance, two years with the same growing season timing (i.e., identical start and end dates) but different photosynthetic peaks are treated identically, which would not be the case when using the annual maximum GPP_{EC}. Although the 10^{th} percentile threshold may not be optimal for each GPP_{EC} time series, it is consistent with the structure of the mean seasonal cycle of GPP_{EC}, where it separates winter and shoulder seasons from the growing season (Figure B2). In addition, similar separations are obtained when using thresholds within the 5^{th}–15^{th} percentile range, indicating low sensitivity to the chosen percentile. Finally, GPP_{EC} and ER_{EC} values above the 99^{th} percentile were treated as outliers and removed.

Complementary filtering flags were applied to ensure biophysical plausibility of root-level soil activity and photosynthesis during the growing season from a modeling perspective. The specific criteria for these flags are as follows:

- $ST_{10cm} \geq 275.15$ K (i.e., 2 °C above freezing)
- $ST_{20cm} \geq 275.15$ K
- $ST_{39cm} \geq 275.15$ K (applied to ENF sites only, see Table 1)
- $MNT \geq 275.15$ K

ST_{20cm} and ST_{39cm} refer to soil temperature at 20 and 39 cm depths, respectively, and were retrieved from the SMAP SPL4SMGP product. For the remainder of this study, ST refers to ST_{10cm} as ST_{20cm} and ST_{39cm} are not used further.

MINOR COMMENT #5 --

Comment: The description of EC flux-partitioning methods across lines 97 to 111 runs quite long for a modeling paper. Most of this is standard material. Trimming it or moving essential detail to supplementary information would improve the flow of the introduction.

Answer:

Some of the information on the EC measurements and flux-partitioning methods described in Section 2 will be moved in Appendix. Consequently, Section 2 will be shorter in the revised manuscript.

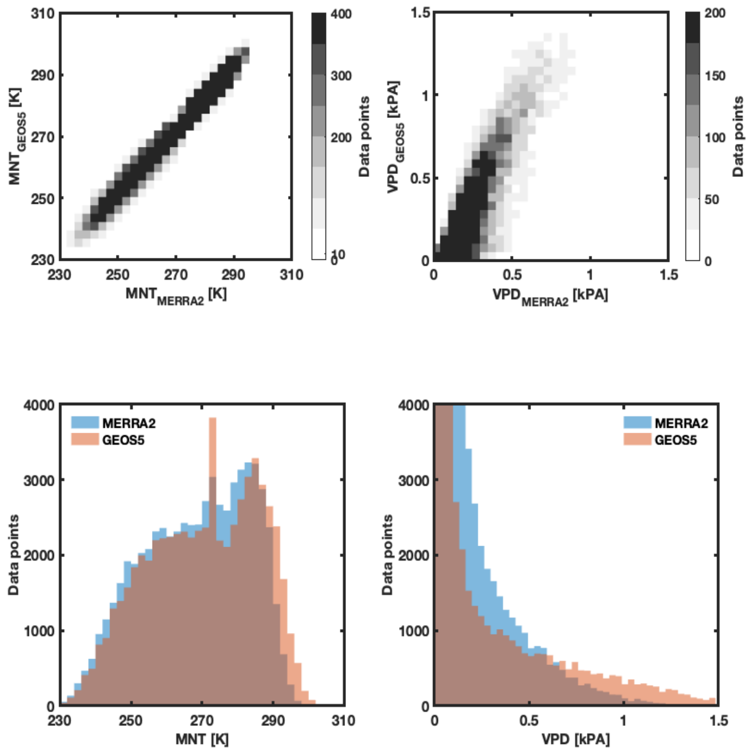
MINOR COMMENT #6 ---

Comment: the justification for using MERRA-2 instead of GEOS-5 FP for VPD and MNT is reasonable, but a brief confirmation that the two products agree closely at the study sites would close a potential concern about whether this substitution introduces systematic differences between the AS-adapted models and the operational L4C configuration.

Answer:

The GEOS-5 FP product was downloaded, and minimum air temperature (MNT) and vapor pressure deficit (VPD) were compared against MERRA-2 at the study sites (see screenshot below, where all sites were pooled). The two products show good overall agreement for MNT but larger differences are observed for VPD (see figure below). However, this is of limited concern because, in the tested model formulations for upland tundra and wetlands, VPD does not constrain GPP estimates. In contrast, VPD has a stronger influence in taiga forests, but most data points fall within the 0–0.5 kPa range, where the corresponding stress scalar (S_{VPD}) roughly ranges from 1 to 0.75, indicating lower constraint (see panels B3–F3 in Figures 2–4 of the original manuscript).

Although the use of the GEOS-5 FP product would have ensured closer consistency for comparison purposes, MERRA-2 was selected because it is better constrained by observations, aligning more with the broader objective of the project to provide more representative estimates of $C0_2$ fluxes in Arctic and Subarctic regions.



Caption: Comparison of minimum air temperature (MNT) and vapor pressure deficit (VPD) between GEOS-5 and MERRA-2 at the study sites, shown as scatter plots and histograms. Sites from the three ecosystems were pooled together.

MINOR COMMENT #7 ---

Comment: The abbreviation table is a helpful addition given the notation density of the paper, but "AT" and "SOC pool" structure (labile, structural, recalcitrant) are referred to in the text without full entries in the table. A quick check for completeness would be worthwhile.

Answer:

The abbreviated terms for the the three SOC pools (labile, structural, recalcitrant) along with the “AT” term will be added to Table 2 in the revised manuscript.

Abbreviation	Definition
AS	Arctic-Subarctic
L4C model	Soil Moisture Active Passive Level-4 Terrestrial Carbon Flux model
EC	Eddy covariance
PFT	Plant functional type
NEE	Net ecosystem CO_2 exchange $[\mathrm{gCm^{-2}d^{-1}}]$
GPP	Gross primary production $[\mathrm{gCm^{-2}d^{-1}}]$
ER	Ecosystem respiration $[\mathrm{gCm^{-2}d^{-1}}]$
AR	Autotrophic respiration $[\mathrm{gCm^{-2}d^{-1}}]$
HR	Heterotrophic respiration $[\mathrm{gCm^{-2}d^{-1}}]$
NPP	Net primary production $[\mathrm{gCm^{-2}d^{-1}}]$
PAR	Photosynthetically active radiation $[\mathrm{MJm^{-2}d^{-1}}]$
FPAR	Canopy-intercepted fraction of absorbed photosynthetically active radiation $[\mathrm{dim.}]$
APAR	Canopy-absorbed photosynthetically active radiation $[\mathrm{MJm^{-2}d^{-1}}]$
MNT	Minimum air temperature $[\mathrm{K}]$
AT	Mean air temperature $[\mathrm{K}]$
VPD	Vapor pressure deficit $[\mathrm{kPa}]$
RZSM	Rootzone soil moisture $[\mathrm{m^3m^{-3}}]$
SSM	Surface soil moisture $[\mathrm{m^3m^{-3}}]$
ST	Soil temperature $[\mathrm{K}]$
SOC	Soil organic carbon $[\mathrm{gCm^{-2}}]$
SOC_1	Labile soil organic carbon pool $[\mathrm{gCm^{-2}}]$
SOC_2	Structural soil organic carbon pool $[\mathrm{gCm^{-2}}]$
SOC_3	Recalcitrant soil organic carbon pool $[\mathrm{gCm^{-2}}]$
L_{fall}	Litterfall $[\mathrm{gCm^{-2}}]$
GDD	Normalized growing degree days $[\mathrm{dim.}]$
S_x	Stress scalar corresponding to the environmental variable x $[\mathrm{dim.}]$
LAI	Leaf area index $[\mathrm{dim.}]$
NDVI	Normalized difference vegetation index $[\mathrm{dim.}]$

Table 2. Summary of frequently used abbreviations.