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 C3 and D7 (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 retained the standard random 70%–30% splits in the revised manuscript and replaced the original Tables 4 and 5 with those presented here. Section 4.4 (Method) has be updated accordingly and Section 5 (Results) now compares median metrics of the testing splits across model formulations. Nevertheless, Section 6.6 has been added to the revised manuscript to explicitly highlight that the influence of temporal autocorrelation on model performance was evaluated and Tables C3 and D7 are presented in the Appendix.

Please note that Tables 4, 5, C3 and D7 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 using the standard random 70%–30% splits) have been included in Appendix.

GROSS PRIMARY PRODUCTION TABLES (TABLE 4 vs. TABLE C3)

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 C3. 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 D7)

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 D7. 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.

649 **6.6 Influence of temporal autocorrelation on model performance**

650 The current model validation approach uses random 70 %–30 % training–testing subsets for model calibration and evaluation
651 (Sections 4.3 and 4.4). As a result, model performance may be partly influenced by temporal autocorrelation, since GPP_{EC} and
652 ER_{EC} data points from the same site may be separated by only a few days while belonging to the training and testing subsets,
653 respectively. Consequently, a complementary sensitivity analysis of model performance was conducted using a validation
654 approach that provides a stronger temporal separation between the data subsets used for calibration and evaluation. Instead of
655 using random 70 %–30 % training–testing subsets, complete years were withheld from model calibration. For each ecosystem,
656 the available data were assigned to 6-year training and 2-year testing periods. Because the study period spans 8 years, this
657 yielded 28 possible train–test runs. The sizes of the training and testing subsets vary among runs because data availability differs
658 among years and sites. This alternative validation approach reduces the potential influence of temporal autocorrelation but
659 introduces variability in sample size among runs, which may itself affect performance metrics. Nevertheless, results obtained
660 with this complementary approach were comparable to those obtained with the random 70 %–30 % splits, indicating that the
661 main findings of this study are robust to the validation strategy employed (Tables C3 and D7).

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:

Partitioned GPP_{EC} and ER_{EC} were selected as reference in this work because they currently constitute the most widely used reference datasets for model calibration. Most modeling products have been developed and evaluated using these references (e.g., SMAP L4C, FLUXCOM-X https://doi.org/10.5194/bg-21-5079-2024).

Nevertheless, 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 has been 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).

The revised text is provided below:

6.7 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; $g(MNT_{min})$ has been 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: **GPP_5**, GPP_4, GPP_3, GPP_2, GPP_1 `
- AIC/BIC: **GPP_5**, GPP_4, GPP_3, GPP_2, GPP_1

(2) Taiga forests

- Equation 14: **GPP_5**, GPP_4, GPP_3, GPP_2, GPP_1
- AIC/BIC: **GPP_5**, GPP_4, GPP_3, GPP_2, GPP_1

(3) Wetlands

- Equation 14: **GPP_4**, GPP_3, GPP_5, GPP_2, GPP_1
- AIC/BIC: **GPP_4**, GPP_5, GPP_3, GPP_2, GPP_1

Rankings for ER:

(1) Upland tundra

- Equation 14: **ER_4**, ER_3, ER_5, ER_2, ER_1 `
- AIC/BIC: **ER_4**, ER_5, ER_3, ER_2, ER_1

(2) Taiga forests

- Equation 14: **ER_4**, ER_5, ER_3, ER_2, ER_1
- AIC/BIC: **ER_5**, ER_4, ER_3, ER_2, ER_1

(3) Wetlands

- Equation 14: **ER_3**, ER_2, ER_4, ER_5, ER_1
- AIC/BIC: **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, the original scoring procedure has been removed and replaced with AIC and BIC in the revised manuscript. Sections 4.4 and 5 have been updated accordingly. In addition, ranks have been 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:

$NEE_{i,j} = ER_i(GPP_j) - GPP_i$

instead of

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

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 $gCm^{-2}d^{-1}$. 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 has been 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 has also been 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 has been incorporated in the revised manuscript in Appendix (see Table C2 below), and Section 6.1 has been 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).

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

557 **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 has been updated to explicitly indicate that the 25 ER and NEE configurations were tested and the corresponding results have been added in Appendix. Results in Section 5.3 has also been updated. Table 6 has been 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 original manuscript). Finally, the discussion regarding NEE has been updated in Section 6.5.

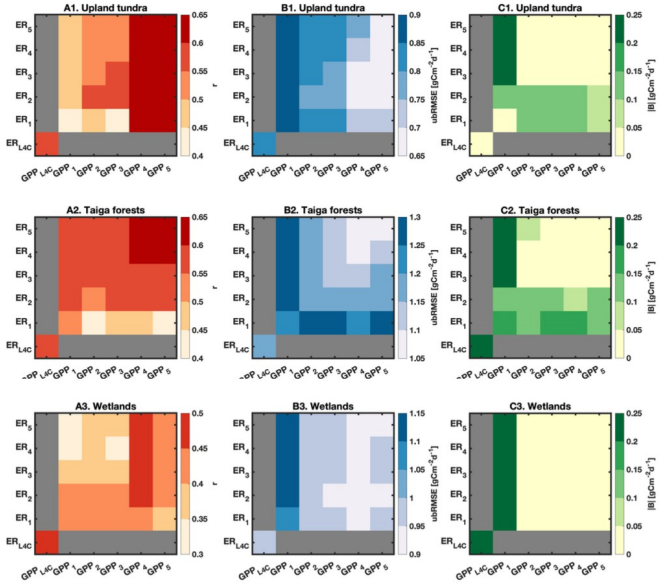


Figure 8. Spatiotemporal performance of modeled **net ecosystem CO_2 exchange (NEE)** against daily-averaged eddy covariance NEE (NEE_{EC}) for **upland tundra**, **taiga forests**, and **wetlands** during the growing season. NEE is computed as the difference between ecosystem respiration (ER) and gross 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 $gCm^{-2}d^{-1}$). 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" has been 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" has been 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 have been moved to Tables 4 and 5 (see Major Comment #1). Figure 8 still exists but is now 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 has been 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 has been updated in the revised manuscript to clarify the choice of the threshold (see below) and Figure B2 has been added in Appendix for transparency.

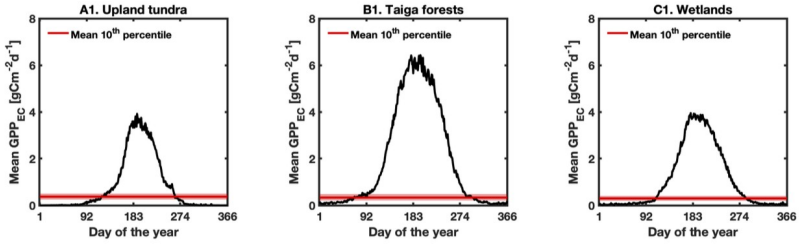


Figure B2. Mean seasonal cycle of daily averaged eddy covariance (EC) gross primary production (GPP_{EC}) for upland tundra, taiga forests and wetlands. The horizontal red line indicates the estimated separation between dormant and transitional periods with active photosynthetic periods (i.e., the growing 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 $gCm^{-2}d^{-1}$ 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 has been moved in Appendix. Consequently, Section 2 is shorter in the revised manuscript.

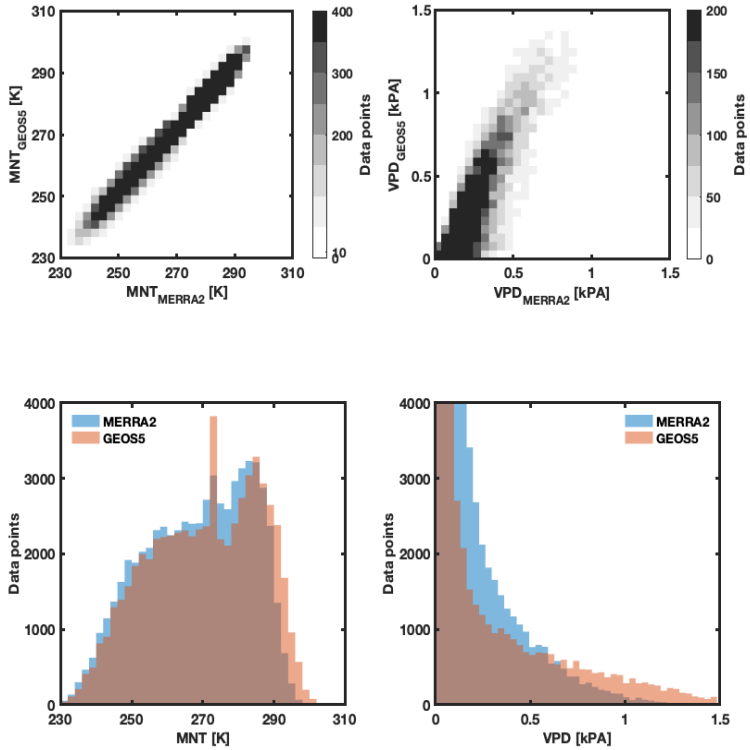
MINOR COMMENT #6 ----------

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 have been 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 [$gCm^{-2}d^{-1}$]
GPP	Gross primary production [$gCm^{-2}d^{-1}$]
ER	Ecosystem respiration [$gCm^{-2}d^{-1}$]
AR	Autotrophic respiration [$gCm^{-2}d^{-1}$]
HR	Heterotrophic respiration [$gCm^{-2}d^{-1}$]
NPP	Net primary production [$gCm^{-2}d^{-1}$]
PAR	Photosynthetically active radiation [$MJm^{-2}d^{-1}$]
FPAR	Canopy-intercepted fraction of absorbed photosynthetically active radiation [dim.]
APAR	Canopy-absorbed photosynthetically active radiation [$MJm^{-2}d^{-1}$]
MNT	Minimum air temperature [K]
AT	Mean air temperature [K]
VPD	Vapor pressure deficit [kPa]
RZSM	Rootzone soil moisture [m^3m^{-3}]
SSM	Surface soil moisture [m^3m^{-3}]
ST	Soil temperature [K]
SOC	Soil organic carbon [gCm^{-2}]
SOC_1	Labile soil organic carbon pool [gCm^{-2}]
SOC_2	Structural soil organic carbon pool [gCm^{-2}]
SOC_3	Recalcitrant soil organic carbon pool [gCm^{-2}]
L_{fall}	Litterfall [gCm^{-2}]
GDD	Normalized growing degree days [dim.]
S_x	Stress scalar corresponding to the environmental variable x [dim.]
LAI	Leaf area index [dim.]
NDVI	Normalized difference vegetation index [dim.]

Table 2. Summary of frequently used abbreviations.

REFEREE #2

We thank the reviewer for the comments, which have helped us improve the clarity of the manuscript and make it more accessible to a broader audience, including readers who are not familiar with the topic or the modeling framework.

MAJOR COMMENTS

MAJOR COMMENT #1 ----------

[…] a major concern I have, which is not addressed or mentioned at all, is the spatial-temporal auto-correlation. The work is based on the assumption that one or multiple proposed model formulations showed improved accuracy compared to SMAP L4C product. The evaluation results indeed support this assumption. However, it seems the training and evaluation data were randomly spitted as 70%:30% for training:testing. Given that the authors take each unique combination of EC measurement and day as a data point (Lines 110-111), there is high spatial and temporal auto-correlation between the training and evaluation dataset. Because of this, the output accuracy metrics are impacted by this auto-correlation, suggesting the accuracy we are seeing is more like training accuracy rather than testing accuracy. For our own work, I ever saw testing accuracy dropped from R2 = 0.72 to R2= 0.15 after removing the impact of auto-correlation. This suggests the improved accuracy metrics from the proposed model formulations might be a result of this auto-correlation rather than real model improvement compared to the SMAP L4C baseline. The authors need to clarify this.

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). Standard random 70%–30% splits were used 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, maintaining a fixed number of data points makes sure that performance variability is not driven by sample size variation or inconsistencies in data availability.

Another approach, more robust to autocorrelation, was tested by withholding complete site-years during calibration (see Referee #1 Major Comment #1). The results are shown in Tables C3 and D7 (see Tables below). Instead of using 70–30% training–testing 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. With this approach, the training and testing subsets vary in size across runs, meaning that performance variability may be influenced by differences in sample size or inconsistencies in data availability.

Overall, the two approaches produce comparable results for both GPP and ER. We therefore retained the standard random 70%–30% splits in the main text of the revised manuscript and replaced the original Tables 4 and 5 with those presented here. Section 4.4 (Method) has been updated accordingly and Section 5 (Results) now compares median metrics of the testing splits across model formulations. Nevertheless, Section 6.6 has been added to the revised manuscript to explicitly highlight that the influence of temporal autocorrelation on model performance was evaluated and Tables C3 and D7 have been included in Appendix.

Please note that Tables 4, 5, C3 and D7 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 using the standard random 70%–30% splits) have been included in Appendix.

GROSS PRIMARY PRODUCTION TABLES (TABLE 4 vs. TABLE C3)

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 C3. 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 D7)

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 D7. 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.

649 **6.6 Influence of temporal autocorrelation on model performance**

650 The current model validation approach uses random 70 %–30 % training–testing subsets for model calibration and evaluation
651 (Sections 4.3 and 4.4). As a result, model performance may be partly influenced by temporal autocorrelation, since GPP_{EC} and
652 ER_{EC} data points from the same site may be separated by only a few days while belonging to the training and testing subsets,
653 respectively. Consequently, a complementary sensitivity analysis of model performance was conducted using a validation
654 approach that provides a stronger temporal separation between the data subsets used for calibration and evaluation. Instead of
655 using random 70 %–30 % training–testing subsets, complete years were withheld from model calibration. For each ecosystem,
656 the available data were assigned to 6-year training and 2-year testing periods. Because the study period spans 8 years, this
657 yielded 28 possible train–test runs. The sizes of the training and testing subsets vary among runs because data availability differs
658 among years and sites. This alternative validation approach reduces the potential influence of temporal autocorrelation but
659 introduces variability in sample size among runs, which may itself affect performance metrics. Nevertheless, results obtained
660 with this complementary approach were comparable to those obtained with the random 70 %–30 % splits, indicating that the
661 main findings of this study are robust to the validation strategy employed (Tables C3 and D7).

MAJOR COMMENT #2 ----------

[…] It is also unclear what data were used to evaluate and compare the NEE estimation between NEEL4C and NEEAS (Table 6). The author mentioned in Lines 350-351 without specifying what data was used. The 70%:30% split approach was used to optimize model parameters of the proposed formulations, after which there seems no independent data left to evaluate NEE with optimized GPP and ER formulations.

Answer:

To clarify the overall framework, three interdependent reference datasets were used in this study: GPP_{EC} , ER_{EC} , and NEE_{EC}, with $NEE_{EC} = ER_{EC} - GPP_{EC}$.

For model development, a 70%–30% split was applied to GPP_{EC} and ER_{EC}: 70% of the data were used to calibrate the AS-adapted GPP and ER formulations; the remaining 30% were used for evaluation.

NEE_{AS} is then computed as the difference between the calibrated AS-adapted ER and GPP and evaluated against NEE_{EC}. Therefore, the NEE_{AS} evaluation was conducted on data that were not used during the calibration of the GPP and ER formulations.

We retained the same 70%–30% training-testing split for consistency across GPP, ER, and NEE evaluations, ensuring that NEE_{EC} values in each subset are properly paired with their corresponding GPP_{EC} and ER_{EC}.

We have revised the manuscript to make this workflow clearer.

4.4 Model formulation evaluation

For each ecosystem type, the AS-adapted GPP and ER formulations were evaluated spatiotemporally, capturing the combined effects of spatial and temporal variability, with GPP_{EC} and ER_{EC} used as reference targets. For each of the 100 optimization runs (Section 4.3), the Pearson correlation coefficient (r), the unbiased root mean square error (ubRMSE), and the bias (B) were computed separately for the training and testing subsets, enabling the assessment of model calibration and generalization, respectively. Median values across the 100 runs were reported. The trade-off between goodness of fit and model complexity was quantified using the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC), which were applied on the training subsets. The best-performing ER and GPP formulations were identified based on the lowest ΔAIC and ΔBIC values, defined as AIC - AIC_{L4C} and BIC - BIC_{L4C}, respectively. Median values across the 100 runs were reported. As a complementary analysis, site-level (temporal-only) performance was evaluated for each EC tower by computing r, ubRMSE,

and B using model outputs obtained with the median parameter values across the 100 runs. For this analysis, training and testing data were pooled, and median metrics across EC towers were reported.

As the ER formulations require GPP as an input, all 25 possible ER configurations were systematically evaluated (ER_1 through ER_5 combined with GPP_1 through GPP_5). This approach also yields 25 corresponding NEE configurations, as NEE is defined as the difference between ER and GPP (Equation 3). Evaluating all combinations enables assessment of whether the best-performing GPP formulation is associated with the best-performing ER formulation, and whether the combination of the best-performing ER and GPP formulations also results in the most accurate NEE configuration. It additionally enables identification of cases where GPP and ER formulations that perform less well individually (relative to ER_{EC} and GPP_{EC}) nevertheless yield a more accurate NEE configuration (relative to NEE_{EC}). Such behavior may arise because errors in modeled ER and GPP can either compensate or accumulate when computing NEE.

The 25 NEE configurations were evaluated following the same approach as for the GPP and ER formulations, using NEE_{EC} as the reference target. Although NEE configurations were not directly calibrated, the same training and testing subsets were retained for consistency with the GPP and ER formulation evaluations, ensuring that each NEE_{EC} value was paired with its corresponding GPP_{EC} and ER_{EC} values.

Because the evaluation of 25 ER formulations produces many metrics, only ER_1 through ER_5 using the GPP input that yields the best performance are presented in the main text. Results for all configurations are provided in Appendix D. This avoids bias arising from the choice of GPP input when comparing ER formulations. For NEE, summarized performance for all configurations is presented in the main text, while detailed results are provided in Appendix E. Hereafter, $NEE_{i,j}$ denotes modeled NEE computed as ER_i - GPP_j.

MINOR COMMENTS

MINOR COMMENT #1 ----------

I encourage the authors to provide line number for each line. Also there are too many abbreviations making the paper difficult to follow. Are abbreviations like B for bias and Lfall for Litterfall necessary?

Answer:

Line numbers have been added throughout the revised manuscript.

We agree that the number of abbreviations may hinder clarity. To address this, Table 2 was specifically included to summarize all abbreviations used in the manuscript. We have also revisited the text in the revised manuscript to reduce unnecessary abbreviations where possible. However, we have retained certain standard abbreviations such as B (bias) for consistency with commonly used metrics (e.g., r for correlation and ubRMSE for unbiased root mean square error) and to maintain readability in figures and tables where space is limited. We hope these revisions improve the clarity of the manuscript.

MINOR COMMENT #2 ----------

Lines 348-350 fit better for Section 4.3 Model formulation calibration. I was wondering how NEE was calculated when reading 4.3

Answer:

The content included in lines 348–350 has been revised. Overall, Sections 4.3 and 4.4 have also been revised to improve clarity.

While Equation 1 in Section 1 defines NEE as the difference between ER and GPP, we agree that this relationship was not sufficiently reiterated later in the manuscript, which may have caused confusion when reading Section 4.3. We also recognize that it was not clearly stated that both the L4C model and the tested formulations rely on this definition to compute NEE. To improve clarity, we have updated Equation 3 in the revised manuscript to explicitly state: NEE=ER−GPP.

137 The L4C model runs at a daily time step and is defined as follows:

138 $$NEE(t) = ER(t) - GPP(t) \quad (3a)$$

139 $$GPP(t) = \epsilon_{max} \cdot APAR(t) \cdot S_{MNT}(t) \cdot S_{VPD}(t) \cdot S_{RZSM}(t) \quad (3b)$$

140 $$ER(t) = AR(t) + HR(t) = \alpha \cdot GPP(t) + [k_1 \cdot SOC_1(t) + (1-\eta) \cdot k_2 \cdot SOC_2(t) + k_3 \cdot SOC_3(t)] \cdot S_{ST}(t) \cdot S_{SSM}(t) \quad (3c)$$

MINOR COMMENT #3 ----------

Lines 341 equation 14, the nmin should be a fixed number based on the description, please specify the number.

Answer:

In the original manuscript, n_{min} was set to 6 for both GPP and ER formulations (see Table 3). However, in the revised manuscript, the scoring system has been changed and is now based on the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC), which are more widely established model selection criteria. As a result, Equation 14 has been removed. This change is motivated by Referee #1 (Major Comment #4), who noted that Equation 14 relies on metric ranks rather than raw metric values, thereby discarding information about how much better one formulation is relative to another. For example, a formulation that narrowly outperforms another gets the same rank as one that beats it with a much larger improvement. By adopting AIC and BIC, the revised approach now provides a more informative and quantitative assessment of the relative performance of competing formulations.

MINOR COMMENT #4 ----------

Lines 345-346, the authors quickly mentioned the evaluation on temporal performance without enough details, is the median across EC tower were used for Figure 8? Also the authors were mentioning spatio-temporal performance all the time and the temporal performance several times in results and discussions, but the whole manuscript did provide any temporal dynamics of the GPP/ER/NEE predictions? This type of figure would greatly help readers to understand the research.

Answer:

Site-level (temporal-only) performance was evaluated for each EC tower by computing metrics using model outputs obtained with the median parameter values across the 100 splits. In this analysis, training and testing data were pooled, and median metrics across EC towers were reported in Figure 8. In the revised manuscript, median metrics from Figure 8 have been moved to Tables 4 and 5 to provide all performance metrics in a single place. The evaluation workflow described in Section 4.4 has been updated and clarified accordingly in the revised manuscript (see Major Comment #2).

We recognize that including time-series figures could help readers better understand the study; however, we chose not to include them for several reasons. First, space constraints are a concern, as the manuscript is already quite long. In the original version, we aimed to remain as concise as possible by limiting the number and size of tables and figures while still providing a comprehensive overview of the study. Second, the objective of the paper is not to overly emphasize promoting a model formulation as superior to the original SMAP L4C model, which could be unintentionally reinforced by including time-series comparisons. Rather, our goal is to provide suggestions for the modeling community to explore and test. Finally, given the large number of formulations evaluated, selecting which time series to present would be somewhat arbitrary. As noted in your Comment #9 (as well as Referee #1 Major Comments #4 and #8), arbitrariness in model selection may be a concern. To address this, we have included additional performance results in Appendix in the revised manuscript and therefore propose not to add time-series figures.

MINOR COMMENT #5 ----------

Lines 354 – 355: are these two lines necessary given the section titles like 5.1, and 5.1.1-5.1.3 are making the content pretty clear already.

Answer:

These lines have been removed in the revised manuscript, as the section titles (e.g., 5.1 and 5.1.1–5.1.3) already provide sufficient structure and clarity. While some co-authors initially preferred to briefly restate the content, we agree that this was redundant.

MINOR COMMENT #6 ----------

Discussions

Lines 509-512: I believe you need to add references there to support your discussion and the following recommendation.

Answer:

Elmendorf (2023), *Limits on phenological response to high temperature in the Arctic*, has been added as a reference in addition to Fu et al. (2014).

MINOR COMMENT #7 ----------

Lines 523-525: you need to add references as the comparison between V8 and V7 is apparently not from this work

Answer:

The differences in litterfall scheme between V7 and V8 are described in lines 170–180 and detailed in Equations 5 and 6. The comparison between these two approaches is conducted within this study: the ER_1 formulation implements the V7 scheme, while the ER_2 formulation follows the V8 approach. We have clarified in the revised manuscript Section 4.1 that ER_1 and ER_2 differ only in their litterfall scheme.

Regarding the ER modeling, we tested different formulations for the HR component while leaving the AR component unchanged. The ER formulations, labeled ER_1 through ER_5, are described below (Table 3):

- ER_1: Rather than using the the L_{fall} estimation scheme from the baseline L4C model version 8, ER_1 instead adopts the one from version 7 (Equation 5). The L4C model transitioned from version 7 to version 8 over the course of the present study was conducted, during which the L_{fall} estimation scheme was modified. The version 7 scheme was retained to enable comparison with the one introduced in version 8. Additionally, HR response to SSM (S_{SSM}) is redefined as a logistic ramp, analogous to S_{RZSM} in GPP_3 (Equation 9c and Figure B1.A2). The original linear response (Equation 7d) was directly replaced because the logistic ramp can reproduce a linear behavior if the relationship between SSM and HR is actually linear. The ER response function to ST is unchanged but was recalibrated (Equation 7e).
- ER_2 (defined as ER_1 with additional adjustments): The L_{fall} estimation scheme is reverted to that of the baseline L4C model version 8. Hence, ER_1 and ER_2 differ only in the L_{fall} estimation scheme (version 7 vs. 8), isolating its impact.

MINOR COMMENT #8 ----------

Lines 527-529: not sure why aboveground biomass (AGB) is mentioned here, as the whole paper didn't give any context to discuss about AGB.

Answer:

This sentence has been revised in the updated manuscript, and the reference to aboveground biomass (AGB) has been removed to maintain a clear causal chain in the model description: GPP → NPP → Lfall → SOC → ER.

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 NPP, 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).

MINOR COMMENT #9 ----------

Lines 534-535: the writing is confusing and contradicts the Table 4, as Table 4 clearly suggests GPP3 is better than GPP2. I understand the difference of metrics are not that big, but the authors kind of rely on those small differences to pick up the better formulation between GPP4 and GPP5.

Answer:

In the original manuscript, Table 4 does not clearly indicate that GPP_3 performs better than GPP_2; this may result from a misinterpretation between different pairwise comparisons (e.g., GPP_2 vs. GPP_1, or GPP_4 vs. GPP_3).

However, we acknowledge that the differences between GPP_4 and GPP_5 are indeed small, and that GPP_5 is not clearly superior to GPP_4. In the revised manuscript, the scoring system has been updated (see Referee #1 Major Comment #4). Although GPP_5 is still selected for upland tundra and taiga forests, we now include results for all 25 possible ER formulations (five GPP formulations combined with five ER formulations) in Appendix. This addition improves transparency and highlights that several formulations perform very similarly.

In addition, the same approach has been applied to NEE, where all 25 possible formulations have been evaluated. These results have been provided in Appendix, while only a summary figure is presented in the main text (see new Figure 8 below).

We believe these changes reduce potential arbitrariness in formulation selection and improve the transparency of the analysis.

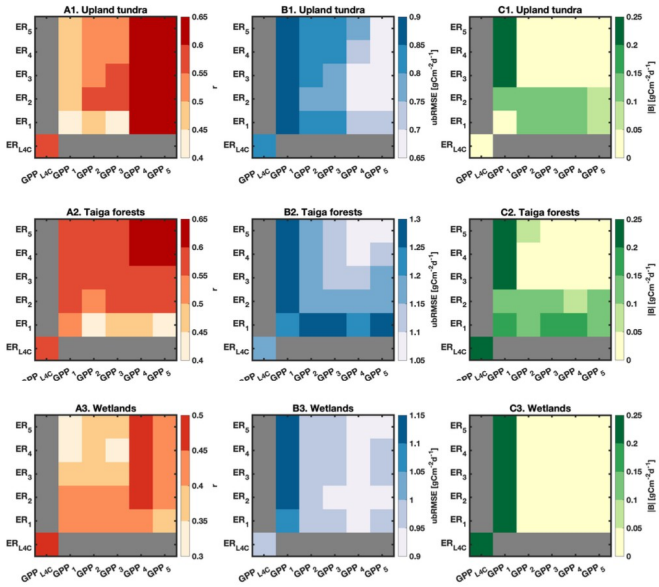


Figure 8. Spatiotemporal performance of modeled **net ecosystem CO_2 exchange (NEE)** against daily-averaged eddy covariance NEE (NEE_{EC}) for **upland tundra**, **taiga forests**, and **wetlands** during the growing season. NEE is computed as the difference between ecosystem respiration (ER) and gross 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 $gCm^{-2}d^{-1}$). 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 COMMENT #10 ----------

Lines 537-538: the writing 'but the added value may not justify the increased complexity required to implement this adjustment' is not straightforward, are you trying to see the model is too complicated and over-fitted? I suggest the authors to increase clarity and conciseness throughout the paper. Another example where the clarity can be improved is in Lines 549-550: 'clear evidence is lacking to suggest that GPP in wetlands does not exhibit diminishing returns under high RZSM conditions'

Answer:

The message intended in lines 537–538 was that, given the relatively limited performance gains associated with introducing nonlinear ramps for modeling GPP responses to environmental drivers, this adjustment appears secondary in importance compared to the implementation of a nonlinear light-response curve and the incorporation of GDD (Section 6.1). We agree that the original wording may have been unclear and potentially misleading, and it has been revised.

Regarding lines 549–550, the intention was to highlight that the literature does not provide clear evidence on whether GPP in wetlands continues to scale or level off under high RZSM conditions. This point has been reformulated in the revised manuscript for improved clarity.

More broadly, these and other potentially unclear sentences have been revised throughout the manuscript.

MINOR COMMENT #11 ----------

Conclusions

Lines 630-634, the importance of winter and should seasons is only mentioned in Conclusions, which is not good writing practice. I recommend authors to add this into the section 6.5 Limitations to acknowledge this research didn't focus on these seasons. The authors then can briefly mention this in Conclusions.

Answer:

In the revised manuscript, a discussion of the importance of winter and shoulder seasons has been added to Section 6.6 (see below; Section 6.5 in the original manuscript). Section 7 - Conclusions has been updated to briefly state that future work will aim to improve the representation of these periods in the L4C model.

Several studies have also shown that GPP responds to the ratio of leaf-internal to ambient CO_2 concentration (Wang et al., 2014b, 2017). Although this ratio is regulated by environmental conditions such as temperature and VPD, neither the original L4C model nor the tested AS-adapted formulations and flux-partitioning algorithms explicitly accounts for the response of GPP to changes in ambient CO_2 concentration. Because ambient CO_2 varies over time and may continue to increase in the future, this omission may limit the ability of the L4C model to accurately predict GPP over long temporal scales.

Finally, it is important to note that this study focuses exclusively on the growing season, whereas the ultimate objective of improving the L4C model for the North American AS is to better estimate the full annual CO_2 budget over recent years by integrating modeled NEE since 2015. Although CO_2 flux magnitudes are highest during the growing season, GPP and AR become minimal or absent during the shoulder seasons and winter, while HR persists, even under snow-covered and frozen soil conditions. Several studies have shown that the winter and shoulder seasons play a critical role in shaping the annual CO_2 budget (Kim et al., 2013; Natali et al., 2019), as they primarily constitute a CO_2 source due to the dominance of HR. Therefore, future work will focus on improving the L4C model for these periods to provide more reliable year-round estimates of NEE, GPP, and ER, as well as more representative annual CO_2 budgets.

7 Conclusions