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 6 and 7 (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 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 $gCm^{-2}d^{-1}$). 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 ----------

[…] 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 will revise 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 will be 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 will also revisit the text in the revised manuscript to reduce unnecessary abbreviations where possible. However, we will retain 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 will 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 will be revised. Overall, Sections 4.3 and 4.4 will be 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 will update 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 will be changed and be 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 will be 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 will provide 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 will be moved to Tables 4 and 5 to provide all performance metrics in a single place. The evaluation workflow described in Section 4.4 will be 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 will include 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 will be 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*, will be 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 will clarify 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 will be revised in the updated manuscript, and the reference to aboveground biomass (AGB) will be 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 will be updated (see Referee #1 Major Comment #4). Although GPP_5 will still be selected for upland tundra and taiga forests, we will include results for all 25 possible ER formulations (five GPP formulations combined with five ER formulations) in Appendix. This addition will improve transparency and highlight 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 will be provided in Appendix, while only a summary figure will be presented in the main text (see new Figure 8 below).

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

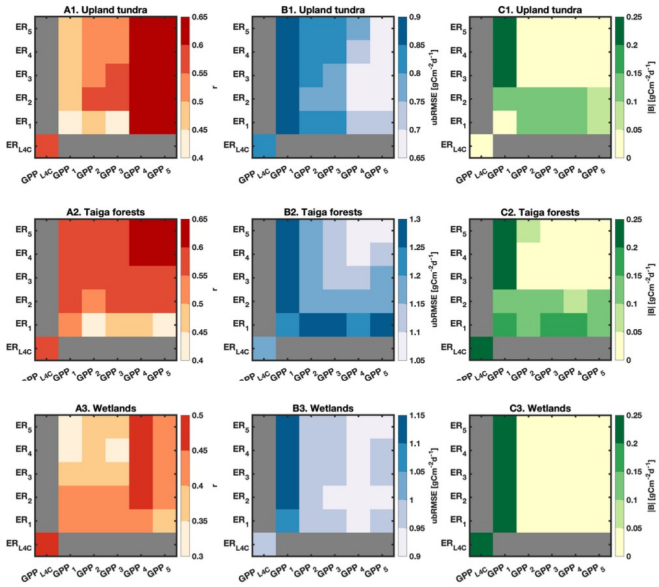

Figure 8. Spatiotemporal performance of modeled **net ecosystem CO_2 exchange (NEE)** against daily-averaged eddy covariance NEE (NEE_{EC}) for **upland tundra**, **taiga forests**, and **wetlands** during the growing season. NEE is computed as the difference between ecosystem respiration (ER) and gross 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 will be revised accordingly.

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 will be reformulated in the revised manuscript for improved clarity.

More broadly, these and other potentially unclear sentences will be 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 will be added to Section 6.6 (see below; Section 6.5 in the original manuscript). Section 7 - Conclusions will be updated to briefly state that future work will aim to improve the representation of these periods in the L4C model.

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.

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.