REFEREE #2

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

The authors used ‘leave year out’, instead of ‘leave site-year out’ approach as they claimed; Consequently, the concern of temporal auto-correlation has been addressed, but I didn’t see any change addressing the concern caused by spatial auto-correlation.

Answer:

A second alternative validation approach was tested to reduce the potential influence of spatial autocorrelation on model performance. In this approach, the 70--30% training--testing subsets were applied to the set of eddy covariance tower sites within each ecosystem rather than to the individual data points, such that all data from a given site were assigned exclusively to either the training or the testing subset. Because the number of sites differs among ecosystems, the number of possible unique train--test runs was 5, 21, and 28 for upland tundra, taiga forests, and wetlands, respectively. While this validation approach reduces the potential influence of spatial autocorrelation, it also introduces large variability in subset size among runs, and provides a stricter assessment of model ability to generalize to unseen sites.
Overall, results obtained with this complementary approach were comparable to those obtained with the random 70--30% splits applied to the individual data points and the leave-year-out validation approach (see below Tables 4, C3 and C4 for GPP, and Tables 5, D7, and D8 for ER). The most notable difference was that the absolute bias was higher for the testing splits in some cases. However, this does not alter the main findings of the study. This difference may reflect variations in the spatial representativeness of the training subsets, in subset size, in site composition among runs, or a combination of these factors. Nevertheless, these effects cannot be disentangled with the present dataset, particularly for upland tundra, where only five sites were available.
Section 6.6 has been updated in the revised manuscript to present this complementary analysis and (see below) Tables C4 and D8 have been added in Appendix.

650 **6.6 Influence of temporal and spatial autocorrelation on model performance**

651 The current model validation approach used random 70–30 % training–testing subsets for model calibration and evaluation
652 (Sections 4.3 and 4.4). The 70–30 % splits were applied to individual data points. Consequently, model performance may
653 be partly influenced by temporal autocorrelation, since GPP_{EC} and ER_{EC} data points from the same EC tower site may be
654 separated by only a few days while belonging to the training and testing subsets, respectively. Model performance may also be
655 influenced by spatial autocorrelation, as data from the same site are included in both the training and testing subsets.
656 Consequently, two complementary sensitivity analyses were conducted using alternative validation approaches that provide
657 stronger temporal and spatial separation, respectively, between the data subsets used for calibration and evaluation.
658 The first alternative validation approach consisted of withholding complete years from model calibration instead of using
659 random 70–30 % training–testing subsets. For each ecosystem, the available data were assigned to 6-year training and 2-year
660 testing periods. Because the study period spans 8 years, this yielded 28 possible unique train–test runs. The sizes of the training
661 and testing subsets varied among runs because data availability differs among years and sites. This validation approach reduces
662 the potential influence of temporal autocorrelation but introduces variability in subset size among runs, which may itself affect
663 performance metrics. Nevertheless, results obtained with this complementary approach were comparable to those obtained

664 with the random 70–30 % splits applied to the individual data points, indicating that the main findings of this study are robust
665 to temporal autocorrelation (Tables C3 and D7 vs. Tables 4 and 5).
666 In the second alternative validation approach, the 70–30 % training–testing subsets were applied to the set of sites within
667 each ecosystem rather than to the individual data points, such that all data from a given site were assigned exclusively to either
668 the training or the testing subset. Because the number of sites differs among ecosystems, the number of possible unique train–
669 test runs was 5, 21, and 28 for upland tundra, taiga forests, and wetlands, respectively. This validation approach reduces the
670 potential influence of spatial autocorrelation while providing a stricter assessment of model ability to generalize to unseen sites.
671 However, as in the first alternative approach, the sizes of the training and testing subsets varied among runs. Overall, results
672 obtained with this complementary approach were comparable to those obtained with the other two validation approaches
673 (Tables C4 and D8 vs. Tables 4 and 5 and vs. Tables C3 and D7). The most notable difference was that absolute B was
674 higher for the testing splits in some cases. However, this does not alter the main findings of the study. This difference may
675 reflect variations in the spatial representativeness of the training subsets, in subset size, in site composition among runs, or a
676 combination of these factors. Nevertheless, these effects cannot be disentangled with the present dataset, particularly for upland
677 tundra, where only five sites were available (Table 1).

GPP model performance comparison:

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 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.

Spatiotemporal evaluation										
GPP	**Splits**	**Upland Tundra**			**Taiga Forests**			**Wetlands**		
		r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
GPP_{L4C}	training	0.63	1.48	0.40	0.59	1.90	-0.34	0.48	2.16	1.30
	testing	0.62	1.38	0.51	0.64	1.81	-0.45	0.56	1.89	1.34
GPP_1	training	0.56	1.19	-0.32	0.62	1.74	-0.44	0.55	1.32	-0.41
	testing	0.57	1.08	-0.16	0.60	1.79	-0.56	0.57	1.28	-0.33
GPP_2	training	0.64	0.97	-0.01	0.67	1.44	-0.04	0.63	1.04	-0.05
	testing	0.62	0.94	0.10	0.62	1.43	-0.19	0.63	0.94	0.02
GPP_3	training	0.66	0.94	-0.00	0.69	1.39	-0.01	0.65	1.01	-0.01
	testing	0.64	0.90	0.14	0.66	1.40	-0.06	0.64	0.93	0.05
GPP_4	training	0.76	0.82	-0.01	0.74	1.29	-0.01	0.73	0.92	-0.01
	testing	0.76	0.77	0.19	0.73	1.25	-0.10	0.71	0.85	0.00
GPP_5	training	0.77	0.79	-0.02	0.75	1.27	-0.01	0.69	0.97	-0.04
	testing	0.74	0.73	0.19	0.73	1.29	-0.13	0.65	0.92	-0.04

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	1.12	-0.19	0.61	1.71	-0.65	0.65	1.16	-0.53
GPP_2	0.62	0.89	0.08	0.67	1.37	-0.04	0.71	0.85	-0.32
GPP_3	0.65	0.86	0.11	0.70	1.31	-0.27	0.74	0.78	-0.27
GPP_4	0.78	0.71	0.18	0.79	1.16	-0.16	0.78	0.77	-0.26
GPP_5	0.76	0.65	0.03	0.77	1.13	-0.31	0.77	0.82	-0.30

Table C4. Same as Table 4, but the 70–30 % training-testing splits were applied to the set of eddy covariance (EC) tower sites rather than to the individual data points, such that all data from a given site were assigned to either the training or the testing split. Because the number of EC tower sites differs among ecosystems, this yielded 5, 21, and 28 possible runs for upland tundra, taiga forests, and wetlands, respectively, instead of 100.

ER model performance comparison:

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.

Spatiotemporal evaluation										
ER	**Splits**	**Upland Tundra**			**Taiga Forests**			**Wetlands**		
		r	ubRMSE	B	r	ubRMSE	B	r	ubRMSE	B
ER_{L4C}	training	0.42	1.00	0.40	0.34	1.71	-0.10	0.24	1.78	1.74
	testing	0.45	0.94	0.35	0.34	1.65	-0.16	0.35	1.51	1.83
ER_1	training	0.53	0.70	-0.11	0.52	1.28	-0.17	0.51	0.80	-0.04
	testing	0.49	0.53	0.25	0.49	1.29	-0.18	0.55	0.74	-0.16
ER_2	training	0.63	0.61	-0.10	0.54	1.22	-0.11	0.54	0.79	-0.01
	testing	0.65	0.44	-0.02	0.51	1.25	-0.20	0.57	0.73	-0.03
ER_3	training	0.58	0.62	-0.00	0.53	1.23	0.00	0.53	0.79	-0.01
	testing	0.64	0.43	0.14	0.51	1.23	-0.06	0.55	0.76	-0.05
ER_4	training	0.70	0.56	-0.03	0.58	1.17	0.00	0.53	0.78	-0.01
	testing	0.71	0.52	0.03	0.57	1.21	-0.08	0.52	0.77	-0.22
ER_5	training	0.71	0.55	-0.01	0.60	1.15	0.00	0.53	0.80	-0.04
	testing	0.69	0.46	0.04	0.57	1.21	-0.09	0.53	0.78	-0.17

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	0.08	0.61	0.97	-0.22	0.65	0.57	-0.06
ER_2	0.66	0.44	0.04	0.65	0.95	-0.12	0.66	0.52	-0.10
ER_3	0.61	0.44	0.10	0.64	0.96	0.02	0.66	0.53	0.10
ER_4	0.58	0.40	0.09	0.66	1.01	0.03	0.64	0.51	0.12
ER_5	0.61	0.42	-0.00	0.66	1.04	0.06	0.63	0.51	-0.12

Table D8. Same as Table 5, but the 70–30 % training-testing splits were applied to the set of eddy covariance (EC) tower sites rather than to the individual data points, such that all data from a given site were assigned to either the training or the testing split. Because the number of EC tower sites differs among ecosystems, this yielded 5, 21, and 28 possible runs for upland tundra, taiga forests, and wetlands, respectively, instead of 100.

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

As 30% of testing is mentioned in both 4.3 Model calibration and 4.4 Model evaluation, it is unclear whether the 30% testing data was used for selecting the parameters or the 70% training data was used for selecting the parameters. If the 30% testing data was used for selecting the model parameters, it could be problematic, as there is no independent testing data in this case because model calibration process already see this testing data.

Answer:

Section 4.3 has been revised to explicitly state that model parameters were estimated using the training subsets, and that final parameter values were defined as the median of the parameter estimates across all optimization runs.

311 **4.3 Model formulation calibration**

312 The AS-adapted GPP and ER formulations were calibrated separately for each ecosystem type (upland tundra, taiga forests,
313 and wetlands), using GPP_{EC} and ER_{EC} as reference targets. The optimization framework for calibration used least-squares
314 minimization via the MATLAB lsqcurvefit function (MathWorks, Inc., 2023), which minimizes the sum of squared differences
315 between the model outputs and target values. This is an unconstrained optimization, with no additional penalty terms applied.
316 To mitigate overfitting, the optimization process was repeated 100 times using different random training (70 %) and testing
317 (30 %) subsets of the data for each ecosystem type: 1,155 (495) data points for upland tundra, 3,243 (1,389) for taiga forests
318 and 2,557 (1,096) for wetlands. In each optimization run, model parameters were estimated using the corresponding training
319 subset, and final parameter values were defined as the median of the parameter estimates across all 100 runs. This approach
320 aimed to capture diverse data combinations and promote a more stable and representative parameterization by smoothing out
321 the influence of outliers or any individual biased subset. A 70 % subset size was arbitrarily chosen to balance between providing

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

For my minor comment 4, the authors' response is not very convincing: 1 figure worth 1000 words, and the figure could be in supplementary file too.

Answer:

Figure B3 has been added in Appendix in the revised manuscript (see below). It compares the mean seasonal cycle of GPP, ER and NEE derived from the eddy covariance reference, the original L4C model, and one of the tested model formulation (GPP_4, ER_4, and $NEE_{4,4}$).

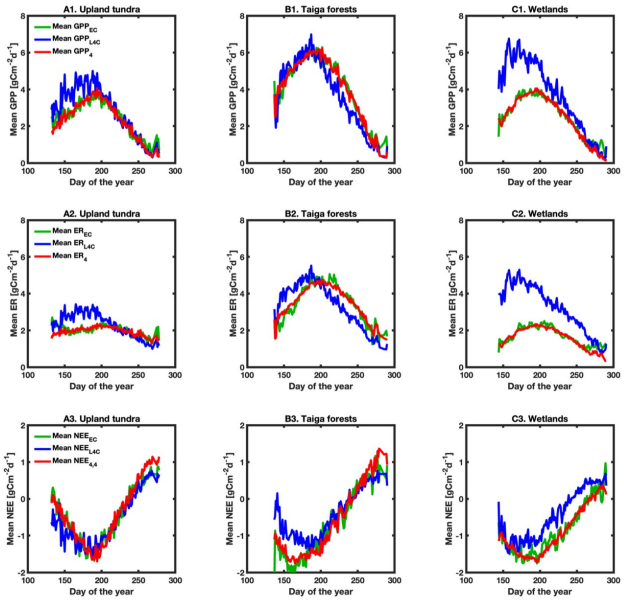

Figure B3. Mean seasonal cycle of daily averaged gross primary production (GPP), ecosystem respiration (ER), and net ecosystem CO_2 exchange (NEE) across eddy covariance tower sites for upland tundra, taiga forests, and wetlands. GPP_{EC}, ER_{EC}, and NEE_{EC} denote eddy covariance data. GPP_{L4C}, ER_{L4C}, and NEE_{L4C} represent outputs from the original L4C model, whereas GPP_4, ER_4, and $NEE_{4,4}$ (i.e., $ER_4 - GPP_4$) correspond to outputs from the Arctic–Subarctic adapted formulations using test case 4 for both GPP and ER (Sections 3 and 4.1). The data shown represent the growing season after applying the filtering criteria described in Section 4.2.