Anonymous Referee #2

The authors of the manuscript 'A differentiable ecosystem modeling framework for large-scale inverse problems: demonstration with photosynthesis simulations' describe the application of the 'differentiable parameter learning'(dPL) framework to the photosynthesis module of FATES model. The framework, and concept, overcomes extrapolation limitations from site-by-site calibration approaches and allows leveraging information content in large-scale datasets towards a global parameterization of photosynthesis models. Neither the concept (Tsai et al., Nature Communications, https://www.nature.com/articles/s41467-021-26107-z, 2021; Bao et al., Authorea, https://www.authorea.com/doi/full/10.1002/essoar.10512186.3, 2022) nor the dPL framework (Tsai et al., 2021; Feng et al., 2022ab) are new. However, the framework is used in the FATES model for the first time and the results would be of interest for further model development, but also to the scientific community at large.

At this point, the experiment focuses on inverting two parameters, V_{cmax25} and B, resulting in that the accuracy of the simulated net photosynthesis rate being slightly improved. The main concerns at this stage relate to apparently incorrect formulations of some key equations, to issues about the validation strategy, to the fact that the forcing data and the experiments are not described sufficiently, challenging the acceptance of the study, while hampering any reproducibility efforts. Please see below for details.

We thank your detailed comments! Wow, this ends up being a 24-page response. As a summary, it seems most of the questions seek clarifications and details about the model. Thank you, and these comments should help us elucidate the model better. We did not find major comments that require computational experiments or major reorganization. There is a question about cross validation, which we have already run. It shows expected and essentially similar results. Moreover, some metrics were requested and we calculated them and reported them in the responses.

We indeed followed our previous differentiable parameter learning paradigm which was first applied in hydrology (Tsai et al., 2021; Feng et al., 2022), as noted in the manuscript, but this is a novel use in the large domain of ecosystem modeling, which is a very large field of study. The system is also different as here we have a nonlinear system of equations while in hydrologic cases we have ordinary differential equations. The mathematical treatment was different. The Julia software solves the system using adjoint solvers, although it is a relatively minor point as we mainly used the PyTorch version for its high parallel efficiency.

We could not have noticed Bao et al., 2022 as it went online after our manuscript did and seems to be undergoing review. Upon some examination, we believe the basic modules are very different. They are using a light-use-efficiency approach and predicted GPP, while our paper focused on photosynthesis using a Farquhar-type model. Hence we don't think there is much overlap between the two.

Major comments:

1. Two key equations are incorrect in the paper:

1) line 140: equation 5, $C_i=C_a-A_n*P_{atm}*(1.4g_s+1.6g_b)/(g_s+g_b)$;
2) line 505: equation A1, $A_c=V_{cmax}*(C_i-\Gamma*)/(C_i+K_c*(1+K_o/O_i))$.
According to the user guide of the FATES model (https://fates-users-guide.readthedocs.io/projects/tech-doc/en/latest/fates_tech_note.html#fundamental photosynthetic-physiology-theory), the equations should be:
1) $C_i=C_a-A_n*P_{atm}*(1.4g_s+1.6g_b)/(g_s*g_b)$;
2) $A_c=V_{cmax}*(C_i-\Gamma*)/(C_i+K_c*(1+O_i/K_o))$.
Since the FATES model is reimplemented in Julia and PyTorch by the authors, the codes might be also wrong. If so, the unit of C_i will be incorrect, leading to errors in the inversion of V_{cmax25} and B. The wrong computation of the effective Michaelis-Menten coefficient ($=K_c*(1+O_i/K_o)$) might only have a slight effect if the temperature is close to 25°C, but should be concerned if the temperature is too low or high (and I do see some points with low leaf temperature in the 'Lin15' database). Thus, I have doubts about the current results and relevant analysis.

Regarding the equations --- we were cautious to adhere to the original FATES equations before implementing it on PyTorch or Julia. Unfortunately, we **realized there were some typos in the manuscript** in line 140 and line 505 in the paper which will then be modified. However, we used the correct equations in our differentiable model as the following:
$C_i=C_a-A_n*P_{atm}*(1.4g_s+1.6g_b)/(g_s*g_b)$;
$A_c=V_{cmax}*(C_i-\Gamma*)/(C_i+K_c*(1+O_i/K_o))$.

No results need to be changed. The code was correct as we compared carefully against the Fortran code in these subroutines as we developed the differentiable versions of the code. We will be publishing the code as the paper gets closer to acceptance so this can be examined in the code. Again, we apologize for the errors in the manuscript.

1. As all the results are validated only once using the temporal holdout data or the random holdout data, the generalizability of the dPL (or NN_B+NN_v) is not clear. If the N-fold or leave-one-out cross-validation can be adopted, the statistical metrics can be more justifiable to reflect the model performance.

Thanks for being rigorous. We believe the randomly selected points were representative, but we already conducted a cross validation (CV) and show the results, as this is trivial. The results are as follows:

(a) Temporal holdout test for the following system (80% train: 20% test)

Runs	Corr		RMSE		Bias		NSE	
	Train	Test	Train	Test	Train	Test	Train	Test
	Corr		RMSE (µmol m-2 s-1)		Bias (µmol m-2 s-1)		NSE	
	Train	Test	Train	Test	Train	Test	Train	Test
Vdef+Bdef	0.565		6.780		1.476		0.041	
Vdef+B	0.632	0.581	6.315	6.088	1.488	0.890	0.182	0.177
V+Bdef	0.758	0.567	4.599	6.148	-0.166	-1.630	0.566	0.161

V+B	0.788	0.766	4.302	4.343	0.104	-0.247	0.62	0.581

(b) Cross Validation (5-fold) test for the following system

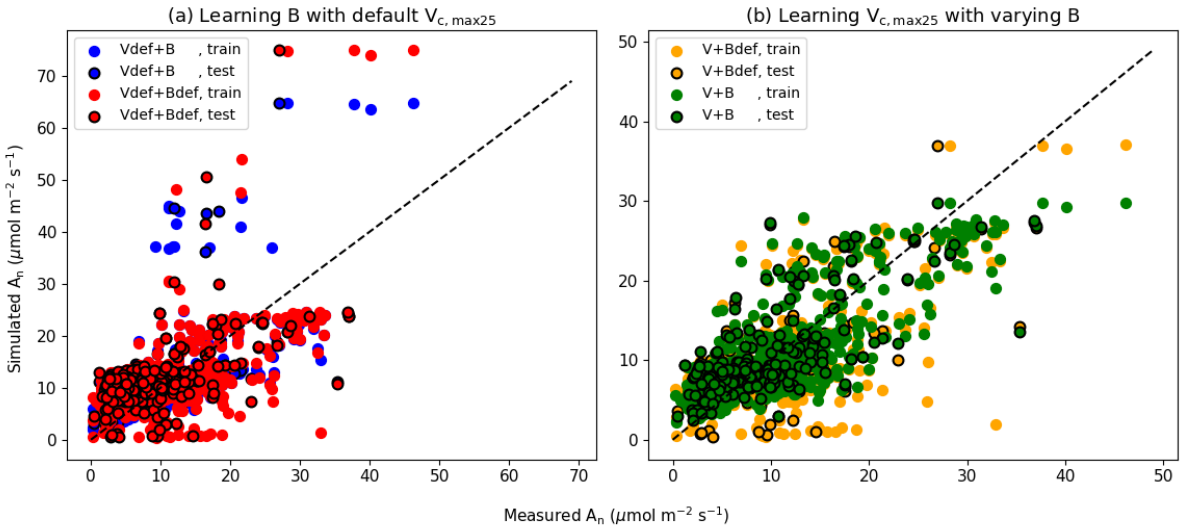
	Corr		**RMSE**		**Bias**		**NSE**	
Runs	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
	Corr		**RMSE (µmol m-2 s-1)**		**Bias (µmol m-2 s-1)**		**NSE**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Vdef+Bdef	0.565		6.780		1.476		0.041	
Vdef+B	0.620	0.618	6.283	6.305	1.456	1.447	0.175	0.171
V+Bdef	0.714	0.707	4.963	5.018	-0.416	-0.407	0.485	0.475
V+B	0.783	0.772	4.308	4.409	0.083	0.094	0.612	0.595

We also provide the metrics for each fold:

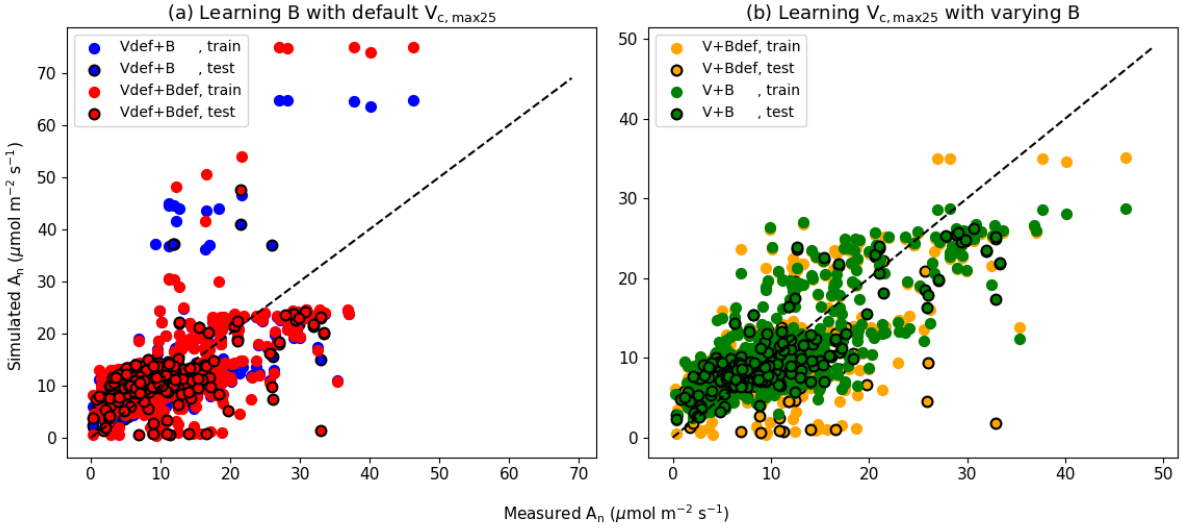
Folds	**COR_test**	**RMS_test**	**BIAS_test**	**NSE_test**	
1	0.726	4.835	0.959	0.495	**V+B**
2	0.834	3.962	-0.228	0.683	
3	0.787	4.512	-0.355	0.617	
4	0.804	4.335	0.027	0.646	
5	0.729	4.318	-0.025	0.509	

This is exactly as we expected in our initial reply posted a few days earlier --- the 5-fold CV results are similar to the previous random results and better than the temporal test results. In addition, we show the train/test A_n values for some random folds:

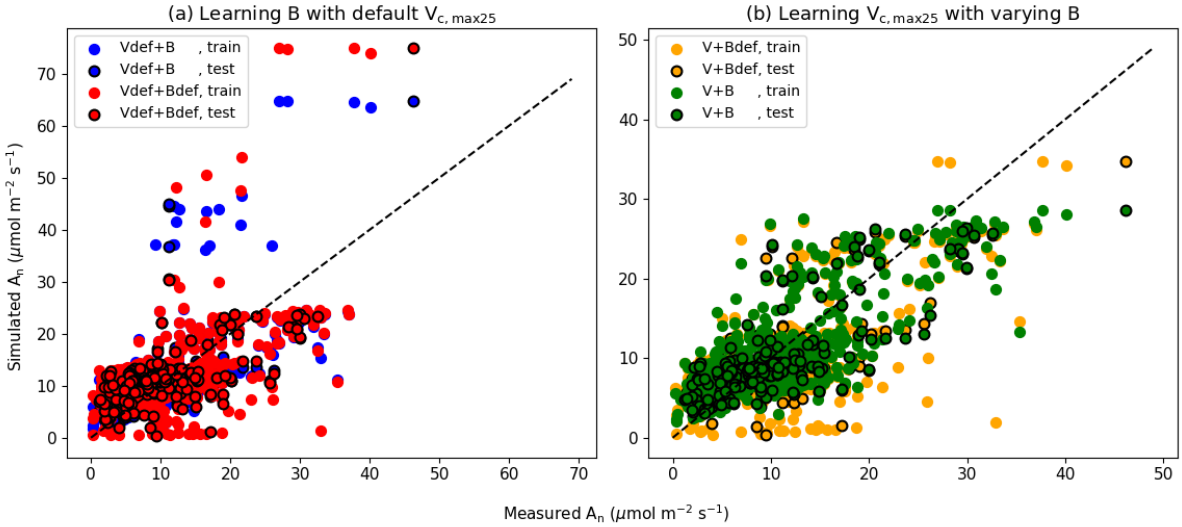
FOLD1 (solid circle indicate test):


(a) Learning B with default $V_{c,max25}$
Vdef+B , train
Vdef+B , test
Vdef+Bdef, train
Vdef+Bdef, test
(b) Learning $V_{c,max25}$ with varying B
V+Bdef, train
V+Bdef, test
V+B , train
V+B , test
Simulated A_n (μmol m^{-2} s^{-1})
Measured A_n (μmol m^{-2} s^{-1})

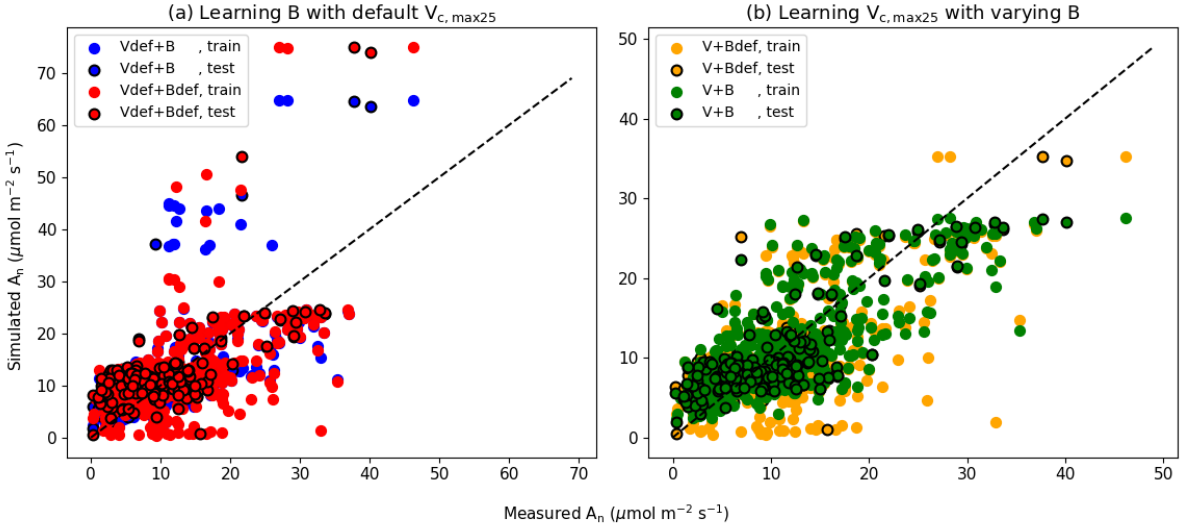
Fold 2:


(a) Learning B with default $V_{c,max25}$
Vdef+B , train
Vdef+B , test
Vdef+Bdef, train
Vdef+Bdef, test
(b) Learning $V_{c,max25}$ with varying B
V+Bdef, train
V+Bdef, test
V+B , train
V+B , test
Simulated A_n (μmol m^{-2} s^{-1})
Measured A_n (μmol m^{-2} s^{-1})

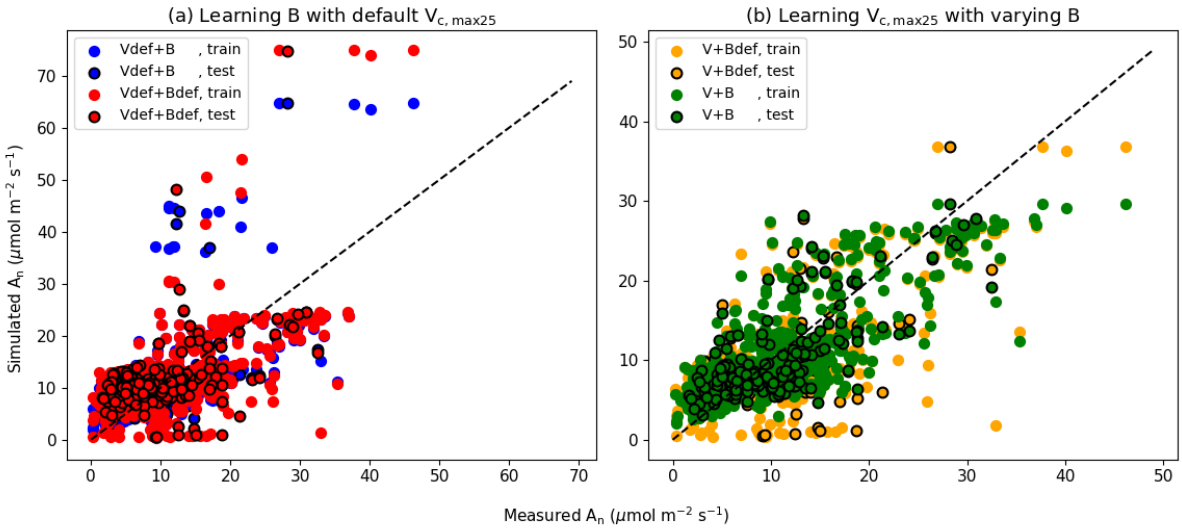
Fold 3:


(a) Learning B with default $V_{c,max25}$
Vdef+B , train
Vdef+B , test
Vdef+Bdef, train
Vdef+Bdef, test
(b) Learning $V_{c,max25}$ with varying B
V+Bdef, train
V+Bdef, test
V+B , train
V+B , test
Simulated A_n (μmol m^{-2} s^{-1})
Measured A_n (μmol m^{-2} s^{-1})

Fold 4:


(a) Learning B with default $V_{c,max25}$
Vdef+B , train
Vdef+B , test
Vdef+Bdef, train
Vdef+Bdef, test
(b) Learning $V_{c,max25}$ with varying B
V+Bdef, train
V+Bdef, test
V+B , train
V+B , test
Simulated A_n (μmol m^{-2} s^{-1})
Measured A_n (μmol m^{-2} s^{-1})

Fold 5:



We believe a spatial test, though, would belong to a different paper as the paper is already getting long. There are many techniques to improve spatial generalization and larger dataset from remote sensing which, if combined with the present content, would just be too much for a first paper. We plan to clarify this point in the paper.

2. The forcing variables and parameters are not clearly differentiated in the paper. For example, is the leaf layer boundary conductance, g_b, a constant parameter across sites or a temporally changing variable? If it is a forcing variable for FATES, where is g_b from? is θ_{ice} a forcing variable or a parameter correlated with temperature and θ_{liq}? Is the C_a a constant value or variable? The model would be different if the spatial and temporal variability of all these factors are considered. If all these are parameters (i.e., scalars), what are the values?

The Lin15 dataset included different forcing variables that we used in our model including:

RH	Relative humidity
T	Air temperature
T_v	Leaf temperature
P_{atm}	Atmospheric pressure
PAR (φ)	Photosynthetic active radiation
g_b	Boundary layer conductance

Concerning (g_b, θ_{ice} and C_a), here are details about how they were considered in the model:

- g_b, the boundary layer conductance values were already available in Lin15 dataset. However, it has some missing values which we then computed using the inverse relationship between g_b and the boundary layer resistance r_b. r_b was approximated by the following equation as documented in CLM5.0 (Lawrence et al., 2019) in section 5.1:

$$r_b = \frac{1}{C_v} * \sqrt{\frac{d_{leaf}}{U_{av}}}$$

Where C_v and d_{leaf} are both constants (0.01 $ms^{-1/2}$ and 0.04 m respectively), while U_{av} is the wind velocity.

- θ_{ice}, the volumetric ice content values were ignored (considered as zero) since both the air and leaf temperatures in our dataset were above the freezing temperature (0 °C or 273.15 K) by at least 5 degrees.

- C_a, the CO_2 partial pressure near the leaf surface values were variable spatially and temporally and they were taken as 0.039% of the atmospheric pressure

We will add the following explanations in the revised manuscript "*We refer to this dataset as Lin15 throughout the rest of this work with 43 sites chosen whose dates and times of measurements were available. Lin15 covered nine different PFT categories including the following: rainfed crop "Crop R", Broadleaf Evergreen Tree Tropical "BET Tropical", Broadleaf Evergreen Tree Temperate "BET Temperate", C3 grass, C4 grass, Needleleaf Evergreen Tree Boreal "NET Boreal", Needleleaf Evergreen Tree Temperate "NET Temperate", Broadleaf Deciduous Tree Temperate "BDT Temperate", and Broadleaf Deciduous Shrub Temperate "BDS Temperate". Measurements were taken on sub-hourly scale but not necessarily on a continuous daily interval. That's why for almost all the sites, data were available on some random days (not necessarily continuous) in one or a few years.*

Lin15 also contained meteorological forcing variables, including air temperature, leaf temperature, atmospheric pressure, relative humidity, radiation and boundary layer conductance. Moreover, we used ERA5 to fill in for any missing forcing variables in Lin15. Both P_{atm} and g_b in equation 4 were used directly from the dataset while C_a is computed as 0.039% of P_{atm}. θ_{ice} in equation 9 was ignored since both the air and leaf temperatures in our dataset were greatly above the freezing temperature"

3. Line 216-218: the reason for replacing saturated soil matric potential (Ψsat) with soil matric potential for closed stomata (Ψc) is not explained. Equation 10 shows that the Ψsat is replaced with soil matric potential for open stomata(Ψo), not Ψc. Furthermore, the Ψi was still calculated using Ψsat in Appendix A (equations A16-A18). I'm confused about which variable was used to calculate Ψi.

Line (216 – 218), we stated the actual equations that we used in for computing ψ_i (in which ψ_{sat} was replaced with ψ_o).

In Appendix A, we kept all the original equations the same whether those related to FATES or to computing the soil water stress function (β_t).

Actual equation used in this study (Line 216 – 218)	Original equation (Appendix A)
$\Psi_i = \Psi_o \times S_i^{-B_i} \geq \Psi_c$	$\Psi_i = \Psi_{sat,i} \times S_i^{-B_i} \geq \Psi_c$

Reasons for this replacement:

In the original CLM4.5 equations, ψ_{sat} is based on empirical functions, percentage of sand *(%sand)*, and fraction of organic matter (F_{om}) (Equations A17 – A18). Using the original Equation 7 for computing ψ_i results in a plant wilting factor w_i equals to one for more than 90% of the data points across different soil layers.

To give the model more flexibility in the computation of ψ_i and thus allow more variability in w_i values, ψ_{sat} was replaced with ψ_o. However, to ensure having w_i values less than or equal 1 as in the original equation 9, we tried to create equation 10 in a way that satisfies this condition using ψ_o. For parameter B (outputted from NN_B), it was restricted to be within the range 0 and 1 to satisfy the same condition as well. Applying those changes, we were able to get ψ_i values within the range of ψ_o and ψ_c while showing more variability in the computed w_i.

Also, we propose adding this paragraph to the **(Model changes)** section for clarification:

"*These changes were implemented to give more flexibility in the computation of the soil matric potential ψ_i. Using the original Equation 7 for computing ψ_i results in a plant wilting factor w_i equals to one for more than 90% of the datapoints across different soil layers. Thus, changing Equation 7 to the form shown in Equation 10 helped to express more variability in w_i and eventually in the computed soil water stress function (β_t).*

The default equations in the Community Land model V4.5 (CLM4.5) for computations of B (Appendix A) show that the parameter B depends on two attributes, %clay and F_{om}, which is why they were used in NN_B. To account for the dependence of ψ_{sat} on %sand (Appendix A) and its replacement by ψ_o (see equations 7 and 10), %sand was also added to NN_B. We also added PFT to NN_B inputs because vegetation may interact with soil moisture constraint and we want to allow relevant factors to be included, rather than restricting the list of inputs to what was previously used in the literature. Since in NN_B, we use quantitative inputs (%sand, %clay, F_{om}) along with categorical inputs (PFT),), we used a one-hot embedding layer in PyTorch"

Here, the point is to calculate photosynthesis. We can see clearly the modified model works very well for photosynthesis. The differentiable modeling approach was specifically designed to enable inspection of various modules and assumptions in the model to update the formula, so more modifications will definitely happen more in the future.

4. Line 218-220: is NN_B used to predict B_i or Ψ_i? B depends on only %clay and F_{om} according to equations A22-A23, while the authors add %sand, which is related to Ψ_{sat} and, therefore, Ψ_i. I didn't find a direct relationship between B_i and %sand according to the original equations in the FATES model. If NN_B is used to predict Ψ_i, I think the equation can be $\Psi_i=\theta_{liq}*NN_B(\%sand,\%clay,PFT,F_{om},T)$, where T represents the factors controlling θ_{ice}, e.g., temperature.

NN_B is used to predict B_i. Indeed, B_i in the original equations depends only on %clay and F_{om}, however due to the changes we implemented to equation 7 (replacement of ψ_{sat} with ψ_o), the %sand was also added to the NN_B. We also added PFT to NN_B inputs because vegetation may interact with soil moisture constraint and we want to allow relevant factors to be included, rather than restricting the list of inputs to what was previously used in the literature. Yes, this is precisely the point of replacing existing equations with NNs --- we can be freed from previous restrictive assumptions that may be faulty. We discussed the incentives for these changes in the previous response

In addition, here, the point is to calculate photosynthesis. We can see clearly the modified model works very well for photosynthesis. The differentiable modeling approach was specifically designed to enable inspection of various modules and assumptions in the model to update the formula, so more modifications will definitely happen more in the future.

Concerning this formula, $\Psi_i=\theta_{liq}*NN_B$(%sand, %clay,PFT,F_{om},T), we would like to thank you for this suggestion and in this regard we can evaluate the model using this formula and the one we suggested and compare the results.

5. I think the neural networks (NN_B and NN_v) need constraints on V_{cmax25} and Ψ_i. Although the authors declared that the predicted V_{cmax25} without any constraints is within a rational range similar to the literature and measurement, the range of the predicted B is not discussed. If the predicted B_i is very large at some point, Ψ_i can be much higher than Ψ_o, leading to w_i being higher than 1 (i.e., exceeding the range defined in equation A15). Besides, the V_{cmax25} is possibly to be inappropriate without any physical constraints at sites not considered in this study.

We did impose some constraints of both NN_B and NN_v in predicting $V_{c,max25}$ and B.
For $V_{c,max25}$:

We constrained the output of NN_v to be between 0 and 1 using a sigmoid activation function for the output layer in the NN. We then rescaled the output to be within a pre-defined range based on literature of minimum value of 20 umol m^{-2} s^{-1} to a maximum value of 150 umol m^{-2} s^{-1}.

For B:
We constrained the output of NN_B to be between [0 , 1] using a sigmoid activation function for the output layer in the NN. Given that the soil wetness S_i (in equation 7 and 10) ranges between [0.01,1] as defined in the original CLM4.5 equations, therefore the term S_i^{-Bi} can have a range of [1, 100] which when multiplied by ψ_o ensures having ψ_i values with a maximum limit of ψ_o, while the condition of $\psi_i >= \psi_c$ was conserved in equation 10 (same as equation 7) for ensuring a minimum limit of ψ_c. Thus, we ensured that ψ_i computed using equation 10 is within the range of ψ_o and ψ_c which resulted in w_i values less than or equal to 1.

This paragraph will be added to section **(Synthetic data and real data experiments)** which states some details in this regard:

"*The MLPs employed were very simple with only three layers; input layer, one hidden layer, and an output layer. To ensure an output value between 0 to 1 for both $V_{c,max25}$ and B parameterizations, sigmoid activation functions were used for both hidden and output layers. $V_{c,max25}$ was then rescaled to be within a pre-defined range based on literature of [20,150] umol m^{-2} s^{-1}. B values were kept between [0,1] and with S_i ranging between [0.01, 1] (see Appendix A), the term S_i^{-Bi} then has a range of [1,100]. This ensured of ψ_o to be the maximum limit of ψ_i, while the condition for a minimum limit of ψ_c was conserved (see equation 10)*"

6. Line 235-236: 'we tested retrieving both V_{cmax25} and B, the latter of which varies spatially and temporally.' If B varies temporally, it should be clarified how the training data is partitioned and how the 'random holdout test' is done. For example, is B changing per year or every N years? how many years/points per site are used to estimate B? Do the training points have to be in sequence or not?

In line 235 – 236: we refer to the synthetic case and since for this case the values for the parameter B were synthesized using the following equation $B = 0.1 * F_{om} + 0.45 * (\%sand + \%clay)$, so B only varies spatially (different static attributes). We will modify this sentence in the next version to be "*we tested retrieving both V_{cmax25} and B, the latter of which varies spatially for different static attributes*".

Moreover, this paragraph will be modified in the section **(Synthetic data and real data experiments)** to explain more about the temporal and the random holdout tests as well as data splitting.

"*For training and testing our candidate models, two different tests were performed with respect to data splitting: random holdout test and temporal holdout test, the latter of which stresses the models' ability to project into the future. In the temporal holdout test, for each PFT in each location, the available dates of measurements were counted where data points measured at the older 80% of these dates were used for training and the other more recent 20% were used for testing. Due to the irregularity of measurement dates at each location (as mentioned previously in section 2.4.1), the temporal periods for the training and testing datasets vary by location. The temporal holdout test was used for both synthetic and real data experiments. For the random holdout test, as the name implies, 80% of the datapoints were randomly selected for training from the available PFT measurements in each location while the rest were used for testing. This test was run only for the real case experiments. We also report results from a 5-fold cross validation where each fold takes turns to be the test fold.*"

7. Line 238-239: 'For simplicity, the computations of B, Ψ_i, w_i, β_t were performed for the top soil layer only.' In the synthetic experiment, only the top soil layer is considered. However, 'B, Ψ_i, w_i' for the other layers are not clarified (=zero or default values in CLM?). Are the other soil layers considered in the real data experiment? If yes, how many 'B' was estimated (i.e., how many soil layers and how many temporally changing B_i)? If not, w_i can only represent the water availability at the top layer. The β_t is equal to w_i and the root distribution, r_i, at the top layer. What is r_i at the top layer (soil depth=0cm according to line 306)?

We will add further explanation for the synthetic case in the Synthetic data and real data experiments section as the following:

"*In the second synthetic case, "$V_{c,max}$ - B", we tested retrieving both $V_{c,max25}$ and B , the latter of which varies spatially for different static attributes. To generate the synthetic data, we assumed $B = 0.1 * F_{om} + 0.45 * (\%sand + \%clay)$, and then the soil matric potential (ψ_i) was calculated using equation 10. The plant wilting factor (w_i) and the soil water stress function (β_t) were calculated using the default equations 9 and 8 respectively. For simplicity, the computations of B, ψ_i, w_i, β_t*

were performed for the topsoil layer only and the other soil layers were ignored. Based on these simplifications, w_1 was equal to β_t with a root distribution value for the top soil layer of one (r_1 = 1). To retrieve B, we used NN_B (see equation 11) but excluded the PFT term."

For the real experiments:
Five soil layers were considered in these experiments with the exact depths described in the (Input and observation datasets) section. NN_B used static attributes (F_{om}, %sand, and %clay) from all soil layers and predicted B values for each layer. So according to B = NN_B (%clay, %sand, PFT, F_{om}), B varies horizontally as well as vertically (static attributes per location), for each soil layer and for each PFT. For better clarification, B equation in the next version could be written as $B_i = NN_{Bi}$ (%$clay_i$, %$sand_i$, PFT, F_{omi})

We will add further explanation for the real case in the Synthetic data and real data experiments section as the following:

"*For both experiments in which B was learnt; V_{def}+B and V+B, the five soil layers (as described in section 2.5.2) were used to estimate B based on the static attributes corresponding to each specific layer. Thus, B varied both horizontally and vertically for each PFT.*"

8. Line 239: 'To retrieve B, we used NNB but exclude the PFT term.'
I think it is not proper if the PFT is excluded from the training but included in the equation. If PFT is excluded, the term should be removed from equation 11. The sentence at line 222 '… along with categorical inputs (PFT), we used…' should be rephrased.

In Line 239, we refer to the synthetic case and as mentioned in the previous comment we synthesized the B parameter values using the following equation B = 0.1 * F_{om} + 0.45 * (%sand + %clay). So, we formulated the NN_B for the synthetic case as B = NN_B (%sand, %clay, F_{om},).

In equation 11, we show the equation used for the real case experiments which included the PFT term as well in NN_B (discussed previously in comment no.3 in the major comments)

9. Line 245: 'The model passing the test of the synthetic case was then applied to a real dataset…'
The same NN was used for synthetic data and real data, but the NN information (layers, neurons activation functions) is not clear. As real data is much more complex, using a different NN structure from the synthetic test might have better performance.

Concerning the NN formation, this paragraph will be added to section **(Synthetic data and real data experiments)** which states some details in this regard:

"*The MLPs employed had three layers; an input layer, one hidden layer, and an output layer. To ensure an output value between 0 to 1 for both $V_{c,max25}$ and B parameterizations, sigmoid activation functions were used for both hidden and output layers. $V_{c,max25}$ was then rescaled to be within a pre-defined range based on literature of [20,150] umol m^{-2} s^{-1}. B values were kept between [0,1] and with S_i ranging between [0.01, 1] (see Appendix A), the term S_i^{-Bi} then has a range of [1,100]. This ensured a maximum limit of ψ_i equal to ψ_o, while the condition of a minimum limit wqual to ψ_c was conserved (see equation 10)*

The quantity of available data posed a limitation and did not permit an extensive hyperparameter tuning experiment with a train/validation/test split. Hence, we employed a lazy trial and error with hyperparameters (learning rates and hidden size) using 70% of the data as training data and 30% as a validation set, just to ensure we had a roughly performing hyperparameter set. We selected a learning rate of 0.01 and a hidden size that is equal to the number of inputs (9 for the NN_V and 8 for the NN_B). We kept the same hyperparameters in the reporting, where we ran 5-fold cross validation with an 80%:20% train: test ratio. In addition, we found that moderately perturbing the hyperparameters resulted in very little change in the performance. This design considered the practical limits of available data, even this study already represents a large-sample study in the domain of ecosystem modeling."

What we mean by "The model passing the test of the synthetic case was then applied to a real dataset…" is that we didn't perform significant changes in the general differentiable model structure when running the synthetic and the real case.

Indeed, it is true that the real case should be more complex than the synthetic case. However, for NN_V we kept it the same for the both cases since in our reference models (CLM4.5, AVIM, BETHY) $V_{c,max25}$ is PFT-dependent parameter and for consistency we didn't make any changes to NN_V (in this paper, as a starting point). For NN_B, we indeed made a slight change between the synthetic and the real case as:

Synthetic Case (one soil layer)	**Real Case (five soil layers)**
B = NN_B (%sand, %clay, F_{om})	B = NN_B (%sand, %clay, PFT, F_{om})

We discussed the reasons for these NN formulations in comment no.3 in the major comments.

Finally, we would like to mention that this study is one of first studies in this field so our purpose is to present the application of dPL framework without necessarily finding the best NNs for learning our target parameters. We can perform more improvements in the parameterization module in the future.

10. Line 266-267: the loss function is very significant to evaluate the NN, but not explained in the paper. Without the equation of the loss function or the NN information, the dPL framework cannot be assessed by others, in other words, the experiment cannot be repeated. I think this doesn't fulfil the requirement of Biogeosciences: 'Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?'.

Concerning the loss function, we discussed its structure in different sections in the manuscript.

$$W = \operatorname{argmin}(L(\delta_{psn}(\theta, \theta_c, F), y^*)) = \underset{W}{\operatorname{argmin}}(L(\delta_{psn}(g^W(R), \theta_c, F), y^*)) \quad (3)$$

In equation 3: we stated that the weights are minimized using the loss function between the simulated target variable y (see Equation 2) and the observed target variable y^*. We then discussed how f1 and f2 equations are reflected on the photosynthesis module in FATES using equation 4 and equation 5. In line 144, we highlighted that the y term is the A_n (the net photosynthesis rate) variable in our problem.

Moreover, figure 1b (new proposed version shown below) shows that the loss function is computed between the simulated and the observed A_n. We mentioned that for the dPL framework, we don't need ground truth for the learnt parameters but for A_n.

Concerning the NN formation, the paragraph we added in response to comment no.9 in the major comments on the last page would further clarify it. Further, our code will be shared upon paper acceptance and the results will be entirely reproducible.

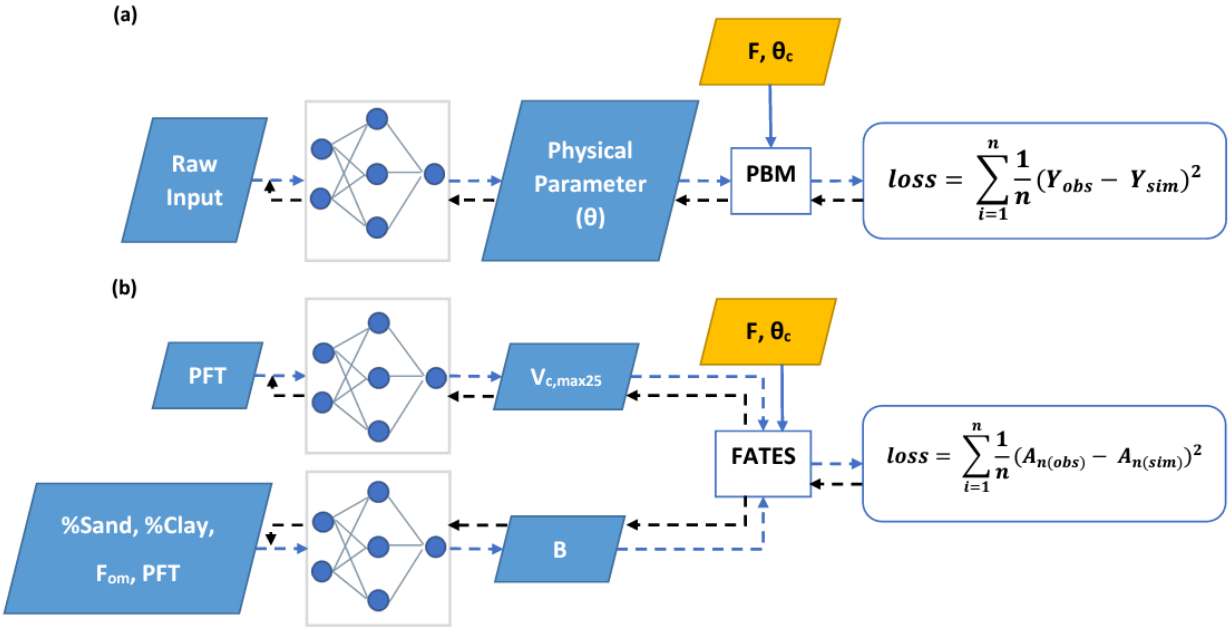


Figure 1. Diagram showing the differentiable parameter learning (dPL) framework which is a hybrid of neural networks and the photosynthesis module in the FATES ecosystem model written on a differentiable platform. (a) The generic workflow: Some raw information is mapped into physical parameters via a neural network. These parameters are sent into a process-based model (PBM), which then outputs variable Y that is compared with observations. Direct supervision for the physical parameters is not required -- we do not need ground truth for these parameters. The loss function is "global" in that it involves all training data points, rather than being computed site-by-site as done in traditional calibration. (b) The workflow for the computational example described in this work. We estimate either $V_{c,max25}$ or the parameter B using neural networks, or both of them at the same time. When they were not estimated from data, default values from the literature were used. Blue arrows show running the neural networks with the PBM in a forward mode, while black arrows indicate backpropagation from the loss function back through the differentiable model equations to the neural networks to update their weights.

11. Line 268-272: the authors 'hope to identify parameters that can generalize well in space', so I think the readers would wonder if the parameters are estimated per site or per PFT. If parameters are estimated per site, how are they aggregated to parameters per PFT in figure 3a and 4a? If estimated per PFT, I'm afraid the spatial variability of the parameters is not fully captured by dPL.

$V_{c,max25}$ values were estimated per PFT since NN_v(PFT) uses just the PFT as input without any static attributes specific to each site. Also, our reference values (used for comparison such as CLM4.5, AVIM, BETHY, and TRY) for $V_{c,max25}$ come for models that define $V_{c,max25}$ per PFT not per site.

'hope to identify parameters that can generalize well in space', by this sentence we that the dPL, contrary to previous site by site calibration, is able to learn from data from all sites simultaneously since the structure of the framework enable it to be trained "globally" in that it involves all training data points, rather than being computed site-by-site as done in traditional calibration. In Tsai et al., 2021 we have already established that casting the parameter problem as parameter learning improves spatial generalization.

Further, we have run some preliminary spatial tests which showed only a small decline of performance when tested in an untrained site. While we obtained a temporal test NSE of 0.581 (80%:20%) train: test ratio, the NSE of a spatial test for the current model is already 0.44, suggesting this model is reasonably well-generalized in space. Unfortunately, we could not identify spatial tests for benchmarking in the ecosystem modeling literature and would appreciate any suggestions with a comparable dataset. As we mentioned earlier, we are working on further improving the spatial generalization with some error mitigation approaches. This will add lots of content and should be for the scope of another paper.

12. Line 292-302: the sources of the soil moisture, stomatal conductance, meteorological forcings and the soil properties are mentioned, but the sources of Ca, gb and Patm are not clear.

Concerning Ca, gb and Patm here are their sources:

- P_{atm} , the atmospheric pressure near the leaf surface, is available in Lin15 dataset so we used them directly and we used ERA5 to fill in for any missing values
- C_a, the CO_2 partial pressure near the leaf surface, is taken as 0.039% of the atmospheric pressure P_{atm}
- g_b, was replied to in comment no.2 in the major comments

A paragraph will be added to explain this (see comment no.2 in the major comments). While these represent simplified treatments, our model's performance suggest that their impacts may be limited. Such simplifications are necessary as we just get started with the different model, and the model can be made more sophisticated later.

13. The data source of 'Lin15' was not specified. I found a database at Lin et al., 2015, bud didn't find the dates information on lines 296-300.

In the supplementary information Lin et al., 2015, page 6:

"Supplementary Table 2: List of data source. The whole database is publicly available and can be downloaded from data repository 40 (https://bitbucket.org/gsglobal/leafgasexchange)."

So, they direct the readers to the database (https://bitbucket.org/gsglobal/leafgasexchange) which have the full parameters list including dates, species, and other forcing variables.

14. Line 304-305: the soil organic carbon content is collected, but the unit is not explained. Does the unit need to be transferred to get the soil organic matter fraction?

Yes, we had to do some unit conversion. According to https://zenodo.org/record/2525553#.Y9Ida-zMKb0 the soil organic carbon is given in 5 g/kg so two conversions were done:

1. Divide by 2 (to convert to %) then divide by 100 (to get a fraction)
2. Multiply by the conventional factor "Van Bemmelen factor" 1.72 (soil organic matter = 1.72 soil organic carbon)

15. Line 410-411: the authors claim that the predicted Vcmax25 'were well correlated with' literature values. However, the correlation coefficient or determination coefficient was never stated in the paper. Too few points are displayed in figure 6b, and the distribution pattern of only four PFT types (crop R, C3 grass, NET Boreal and BDS temperate) is similar to CLM.

First, the point here is that the values we estimated make physical sense, are on the same order of magnitude, and are correlated with the literature values. We expect there to be some correlation but not that high. Higher correlation does not mean it's better. Imagine the extreme case --- if the correlation was 1.0 and every value is the same as literature values, then it would mean the previous values were perfect, which would be surprising and unreasonable. Hence, the precisely correlation value here is not that important. We can calculate the correlation, which is 0.856 with CLM vcmax25, but find it not very relevant to report here.

For Figure 6b, since one point is for a PFT for CLM4.5, and $V_{c,max25}$ is defined on a PFT level, there should be exactly the same number of points as there are PFTs. As a result, the number of data points seemed correct. In the figure below, we show more details about the correlations between the $V_{c,max25}$ learnt by V+B model versus TRY database and other default models. On the other point, we do expect differences from CLM4.5 values.

We attached below a proposal for figure 8 showing the correlation between the $V_{c,max25}$ learnt by V+B model versus TRY database or other default models. As the figure shows, there is high correlation between the estimated $V_{c,max25}$ by V+B model versus CLM4.5 (0.856), BETHY (0.906), and TRY (0.716). However, low correlation exists between the estimated $V_{c,max25}$ by V+B model and AVIM model where the V+B has lower values for BET Temperate, BET Tropical, and BDT Temperate while it shows higher values for BDS Temperate, C3 grass and Crop R. It is difficult to comment which set is better without all models run on the same dataset.

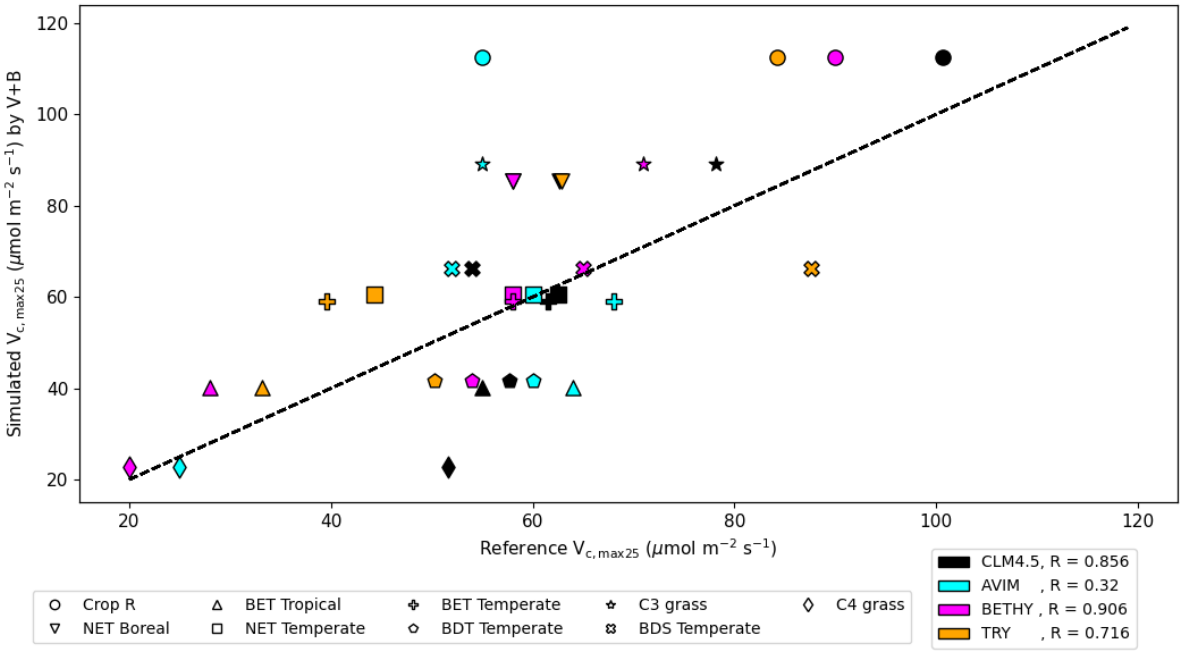


Figure 8. shows the correlation between the $V_{c,max25}$ values estimated by V+B model on y-axis versus $V_{c,max25}$ values from CLM4.5 (black markers), AVIM (cyan markers), BETHY (magenta markers), and TRY database (orange markers). Different markers shape represent different PFTs, while different colours represent different reference sources for $V_{c,max25}$ per PFT. For TRY database we don't values for C3 grass and C4 grass due to the lack of overlap in species between TRY database and our dataset for the two PFTs.

Table 3. $V_{c,max25}$ simulated by V+B model versus observed values from the TRY database (with partial overlap in species with the Lin15 dataset -- the percentage of overlap is provided in the table), and used in different earth system models such as CLM4.5, Atmosphere-Vegetation Interaction Model "AVIM", and the Biosphere Energy Transfer Hydrology scheme "BETHY".

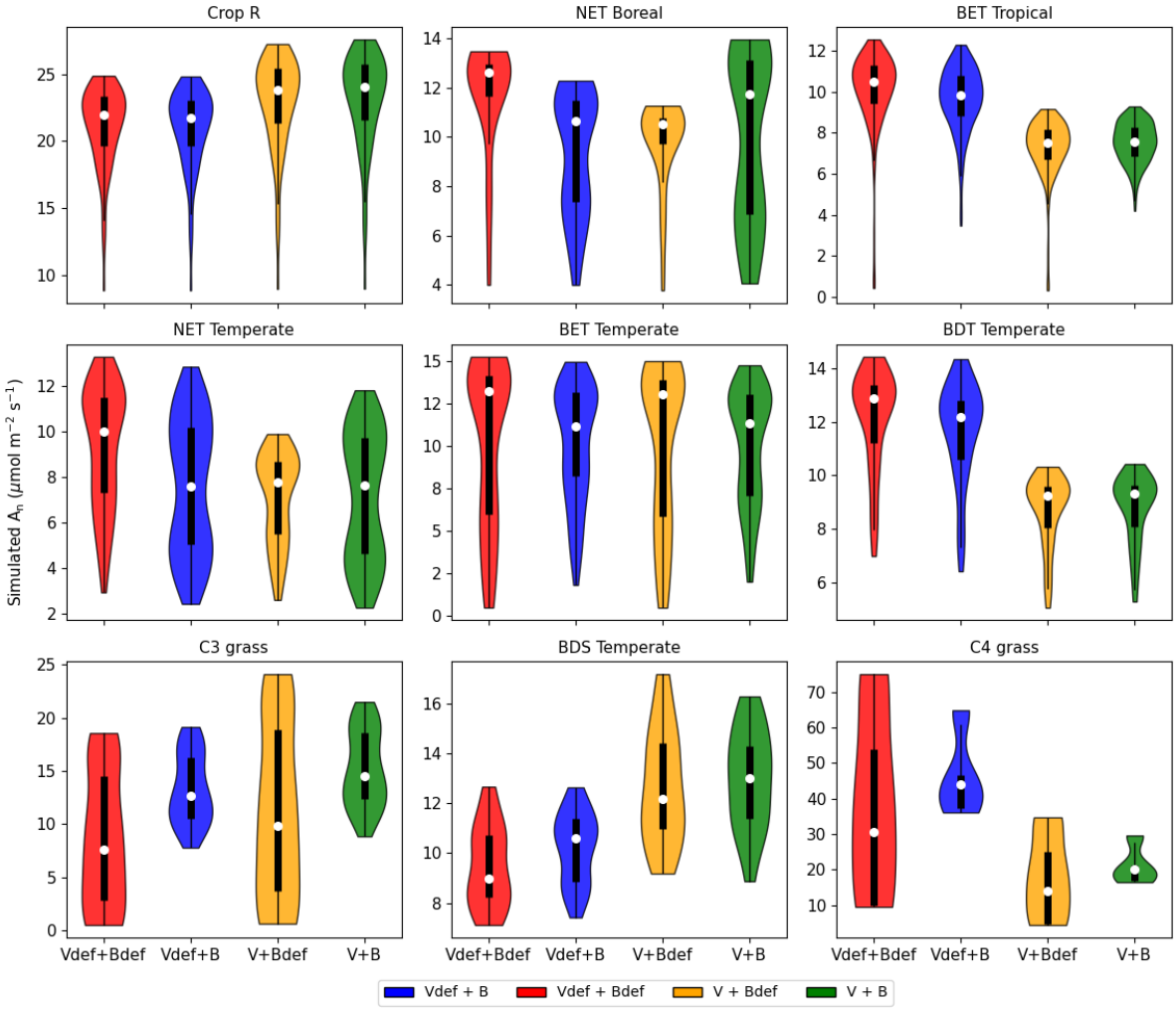
PFT	CLM4.5	AVIM	BETHY	V+B (ours)	TRY (mean / % species overlap)	TRY (std)
BET Temperate	61.5	68	58	59.04	39.54 / 31.3%	4.05
BET Tropical	55	64	28/36	40.07	33.14 / 86.5%	14.09
BDT Temperate	57.7	60	54	41.63	50.27 / 50.0%	21.62
BDS Temperate	54	52	65	66.22	87.61/ 58.3%	11.77
NET Temperate	62.5	60	58	60.64	44.33 / 50.0%	7.13
NET Boreal	62.6	58	58	85.30	62.90 / 100.0%	22.53
C_3 grass	78.2	55/40	71	89.26	-	-
C_4 grass	51.6	25	20	22.86 (limited data points)	-	-
Crop R	100.7	55	90	112.61	84.20 / 60.0%	2.19

16. Line 431-432: I cannot identify the C3 grass at the lower left corner of figure 5b. Maybe a violin plot per PFT can be helpful to show the difference between optimizing B or not for a specific plant type. The figures in the paper only show the net photosynthesis rate across all sites. However, the site-level comparison might be more meaningful to assess the four parameterization strategies: Vdef+B, Vdef+Bdef, V+Bdef, and V+B.

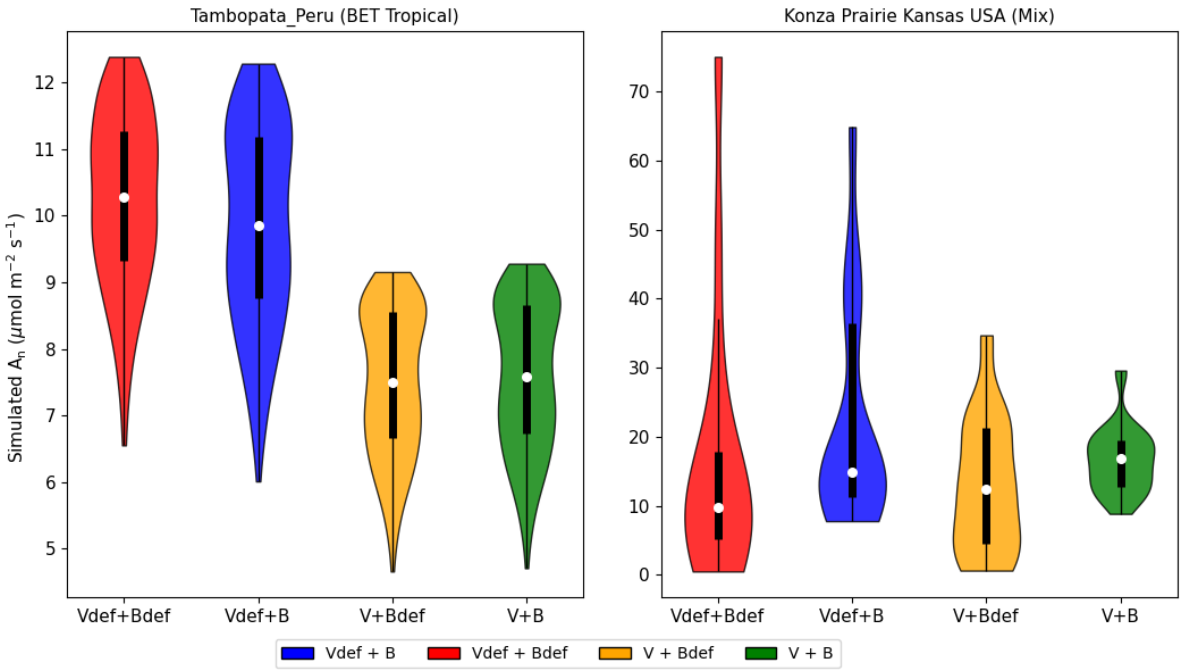
Measurements in Lin15 dataset were taken on sub-hourly scale but not necessarily on a continuous daily interval. for almost all the sites, data were available on some random days (not necessarily continuous) in one or a few years. This means that the data distribution across sites is not balanced some sites have very low amount of data compared to other sites. For this reason, we didn't assess the models using the site level-comparison however we computed the metrics for all sites combined.

Site-level comparison makes more sense when each site has large amount of dataset and dataset amount is uniformly across sites.

We attached the violin plot per PFT (shown below), with the 9 subplots representing the 9 PFTs, and different colors representing different models. Each subplot shows the simulated A_n for all the sites (both training and testing simulations for the temporal test with 80:20 train:test split ratio) with a specific PFT. We will add some of these figures into the revised manuscript.



Violin plot per PFT. Different colors represent different models: Vdef+Bdef, Vdef+B, V+Bdef, and V+B.



Violin plot for two different sites. Different colors represent different models: Vdef+Bdef, Vdef+B, V+Bdef, and V+B. The left panel shows data for Tambopata rainforest in Peru with PFT (Broad leaf Evergreen Tropical). The right panel shows data for Konza Priarie grass ecosystem in USA. We can clearly see different behaviors between V+B and other models.

17. Line 445-450: I didn't see any significant correlation between the estimated Vcmax25 and the PFT-mean from TRY database or other model default values. The authors should provide the scatter plots and the correlation coefficients between the Vcmax values to conclude that the dPL can get parameters correlated with literature values (line 490).

Previously responded to in comment no.15 in the major comments

Minor comments:

1. Line 123: the right part looks very similar to the middle part in equation 3, but the subscript 'W' beside 'argmin' is not explained. As I understand, the 'argmin' in the right part is the same as the 'argmin' in the middle part.

The equation in the next version will be modified to (w below argmin not side subscript)

$$W = \operatorname{argmin}(L(\delta_{psn}(\theta, \theta_c, F), y^*)) = \operatorname*{argmin}_W(L(\delta_{psn}(g^W(R), \theta_c, F), y^*)) \quad (3)$$

By this way, we express that our target is to find the weights of the neural network that minimize the loss function between the observed and the target variable which is the net photosynthesis rate here. So, W here refer to the neural network weights (NN_V and NN_B) in our problem.

2. Line 142: The short name for CO2 partial pressure at the leaf surface is 'Ca', but is 'Cs' in the appendix. Please use a uniform short name across the paper.

C_s and C_a refer to different variables, however, there definitions are close to each other.
C_s : is the CO_2 partial pressure **at** the leaf surface
C_a : is the CO_2 partial pressure **near** the leaf surface

They are correct in the way the equations are written inside the manuscript body or the appendix. At line 140, the definition of C_a will be modified to CO_2 partial pressure **near** the leaf surface.

3. Line 187: equation 11 is cited at line 187 for the first time, but the full equation is placed at line 218. The equation should appear close to the first citation.

Line 187, the equation is written in a more general way as $B = NN_B(R)$, where R refers to the underlying attributes or the raw inputs as mentioned previously in equation 3. In line 218, we tend to show the actual equation that we used for the parameterization in our study. However, we can unify the equation in both appearances to avoid misunderstanding.

4. Line 193: does 'i' represent the soil layer number? I didn't see the explanation around the equation.

Yes, the subscript $_i$ refers to the soil layer number. We will better clarify this in the next version.

5. Line 197: 'across different soil different layers' should be 'across different soil layers'.

Will be modified.

6. Line 203/equation 9: the second line should be $Ti \leq Tf-2$ 'or' $\theta liq \leq 0$.

Will be modified.

7. Line 205: the short name for the physical parameter at the second blue area should be θ but not θc.

Will be modified.

8. Line 218/equation 11: B and Fom should have a subscript, i. Bi=NNBi(%sand,%clay,PFT,Fom,i).

Discussed in no.7 in the major comments.

9. Line 222: the 'one-hot embedding' was already stated at line 183. The definition should be explained where it is mentioned for the first time.

We will move the definition to the first mention of one-hot encoding (from line 222 to line 183)

10. Line 228: the short name for 'differentiable learning framework' is defined but not used.

In Line 228, 'differentiable learning framework' refers to the dPL "differentiable parameter learning framework", we will unify it throughout the whole manuscript to be "differentiable parameter learning framework"

11. Line 310/Figure 2: the full names of the land cover types (e.g., BET tropical) are not explained before or around the figure.

Our dataset included 9 different PFTs categories, a paragraph (see comment no.2 in the major comments with more details about Lin15 dataset will be added to subsection **(Forcing and Photosynthesis rates)** stating the number of PFTs considered plus the full name of each PFT.

12. Line 349: table 2 is mentioned for the first time, but the full table is placed after two pages.

Table 2 was placed two pages after its first mention since we had to place Figure 5 as well after its first mention. We can try to better rearrange the placement of Figure 5 and Table 2 in the next version.

13. Line 384: the CO2 should be CO2(subscript).

Will be modified.

14. Line 390/figure 5: I cannot understand the titles of the subplots. What is the meaning of 'learning B' and 'learning Vcmax25'? The B is not optimized in figure 5a.

Figure 5a subplot shows two models, Vdef+Bdef (red color) and Vdef + B (blue color). So both models agree in using the default $V_{c,max25}$ values corresponding to each PFT that's why subplot (a) title includes **"with default $V_{c,max25}$"**. **"Learning B"** is added to title "a" since B is learnt in Vdef + B model.

Figure 5b subplot shows two models, V+Bdef (yellow color) and V+B (green color). So both models agree in learning $V_{c,max25}$ values corresponding to each PFT that's why subplot (b) title includes **"Learning $V_{c,max25}$"**. "**Varying B"** is added to title "b" since the parameter B is computed from the default equations in CLM4.5 for V+Bdef model, while it is learnt simultaneously with $V_{c,max25}$ for V+B model.

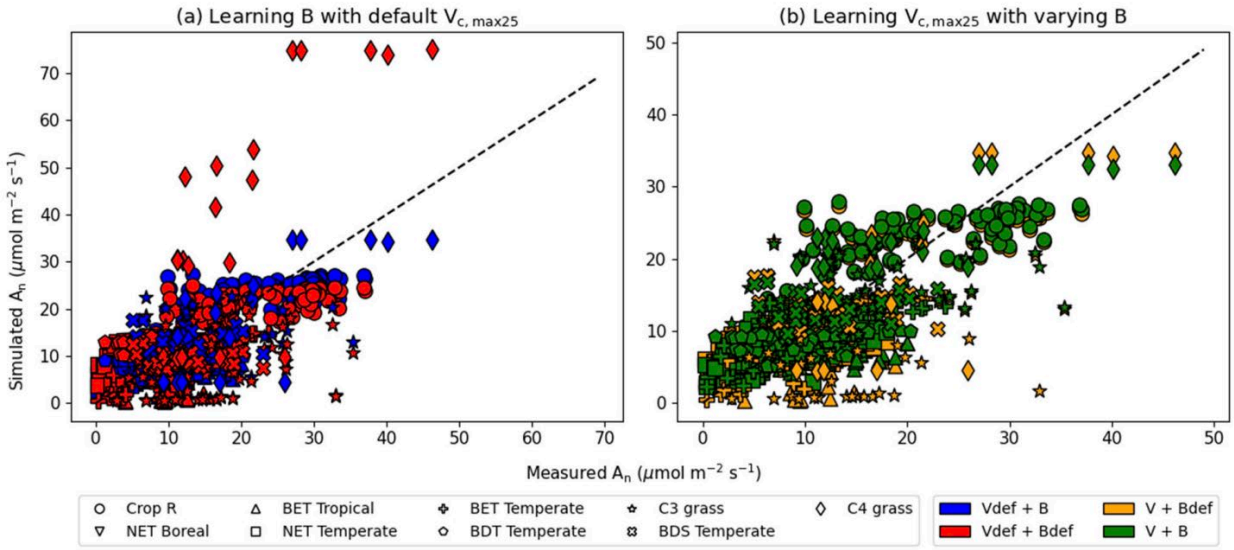


Figure 5. Comparisons of photosynthesis model calibration: mean estimated value of default parameters vs. mean estimated value of best learned parameters vs. observed value for different candidate models. (a) Impact of learning B with default $V_{c,max25}$. (b) Impact of learning $V_{c,max25}$ with varying B. The colors represent the results from the four different models, the shapes indicate the plant functional type (PFT) groups, and the dotted line in each panel indicates the ideal 1:1 relationship.

15. Line 514/equation A7: the Cs is not used.

Cs which refers to the CO2 partial pressure at the leaf surface is used in the model block of equations corresponding to the stomatal conductance computations. Attached below a proposal for figure 2 to be added in the manuscript showing equations corresponding to f1 and f2. The box marked with red color shows the usage of Cs

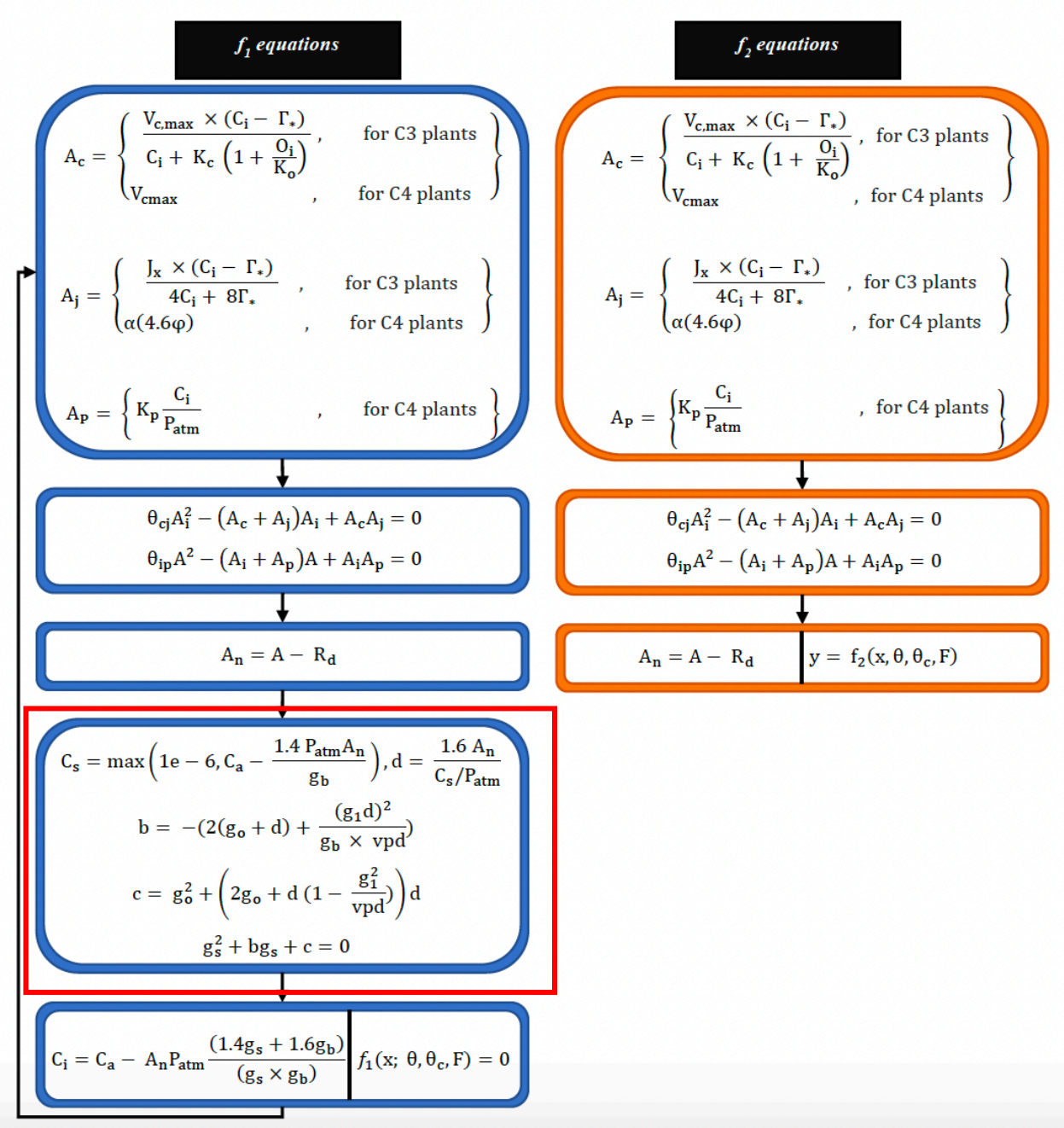


Figure 2. showing the model block of equations corresponding to f1 and f2. Blue boxes refer to equations corresponding to f1. Orange boxes refer to equations corresponding to f2. Further details about the variables and parameters in these equations will be given in a separate table. Once we get the solution for Ci (intercellular leaf CO2 pressure) from f1 equations (nonlinear system), we can run f2 equations to get An (net photosynthesis rate)

16. Line 520-530: the three functions, Φ1, Φ2, and Φ3, need to be clarified.

Φ_1, Φ_2, and Φ_3 refer to the equations or the subroutines that we used to prepare the inputs required to run the FATES photosynthesis module. To run the photosynthesis module, we had to run other correlated subroutine in FATES that provide some crucial inputs required to simulate the photosynthesis.

Φ_1 corresponds to the set of equations in which we used factors from literature or from the Community Land Model (CLM) to map the maximum electron transport rate at 25 °C (J_{max25}), the plant respiration rate at 25 °C (R_{d25}), the initial slope of CO_2 response curve at 25 °C (K_{p25}) from $V_{c,max25}$ as shown below:

$J_{max25} = 1.67\ V_{c,max25}$	Medlyn et al., 2002
$R_{d25} = \begin{cases} 0.015\ V_{c,max25}, & \text{for C3 plants} \\ 0.025\ V_{c,max25}, & \text{for C4 plants} \end{cases}$	Lawrence et al., 2019
$K_{p25} = \begin{cases} 20000\ V_{c,max25}, & \text{for C4 plants} \end{cases}$	

Φ_2 corresponds to the equations responsible for rescaling and adjusting the parameters J_{max25}, K_{p25}, *and* $V_{c,max25}$ for the leaf temperature to output J_{max}, K_p, *and* $V_{c,max}$

Φ_3 corresponds to the the equations responsible for rescaling and adjusting R_{d25} for the leaf temperature to output R_d.

All these equations are well documented in FATES code and in CLM5.0 (Lawrence et al., 2019) in chapter 9 section 9.4.

We will add clarifications in the paper to make this clear.

17. Appendix: the citations of equations are wrong (e.g, lines 503-504, 512, 520, 534…). The equations should be cited using A1-A23.

The citations for all equations in the Appendix will be modified to A[no.] in the new version

References:

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., van den Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty, Journal of Advances in Modeling Earth Systems, 11, 4245–4287, https://doi.org/10.1029/2018ms001583, 2019.

Medlyn, B.E., Dreyer, E., Ellsworth, D., Forstreuter, M., Harley, P.C., Kirschbaum, M.U.F., Le Roux, X., Montpied, P., Strassemeyer, J., Walcroft, A., Wang, K. and Loustau, D. (2002), Temperature response of parameters of a biochemically based model of photosynthesis. II. A review of experimental data. Plant, Cell & Environment, 25: 1167-1179. https://doi.org/10.1046/j.1365-3040.2002.00891.x

Shen, CP. et al., Differentiable modeling in Geosciences to unify machine learning and physical models. https://arxiv.org/abs/2301.04027

Tsai, W-P., K. Fang, X. Ji, K. Lawson, CP. Shen, Revealing causal controls of storage-streamflow relationships with a data-centric Bayesian framework combining machine learning and process-based modeling. Frontiers in Water-Water and Hydrocomplexity, doi:10.3389/frwa.2020.583000 (2020)

Feng, DP., K. Lawson and CP. Shen, Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data, Geophysical Research Letters, doi: 10.1029/2021GL092999 (2021)