Response to Anonymous Referee #2

We sincerely thank the reviewer for her/his effort and the very useful comments. We have revised our manuscript *P-model v1.0: An optimality-based light use efficiency model for simulating ecosystem gross primary production* and have addressed all points raised by the reviewer.

Below, we provide a point-by-point response to all the comments. Text by the reviewer is in blue and indented. Our response is in black. *New text is green, italic*. *Existing (unchanged) manuscript text is black italic*.

> Stocker et al present, calibrate, and evaluate a GPP model that is built off a previously published P-model. From the way the manuscript is structured it seems that the new additions to the P-model are the temperature dependence and the soil moisture stress (but this isn't clear from the abstract or the introduction). The authors should be commended for developing the model into an R package, using the FLUXNET2015 data in a way that recognizes the GPP is modeled product with considerable uncertainty (i.e., by analyzing multiple partitioning methods), and for providing throughout background for the modeling framework. Overall, the manuscript is well-written.

We greatly appreciate this positive assessment and the recognition of the value of our open-access model implementation. In order to clarify the new additions and relation of the P-model to earlier publications early on, we added the following sentence to the abstract:

[...] The model builds on the theory developed in Prentice et al. (2014) and Wang et al. (2017a) and is extended to include low temperature effects on the intrinsic quantum yield and an empirical soil moisture stress factor. [...]

This is clarified further by the text in the introduction that we have added in response to referee 1:

The purpose of this paper is [...] (iv) to introduce a robust and pragmatic solution to resolving model bias under dry and cold conditions.

> My recommendations for improvements are:
>
> Be clearer about what is new to this version of the P-model. Based on the components of the model that are in the introduction vs. methods, I would assume that the temperature dependence and the soil moisture stress are the new components.

This is addressed by added text in the abstract and introduction as described above.

> Since the modeling framework depends on the SPLASH model, more detail is needed about that model. How many parameters does it require and how where the SPLASH parameters determined? Any parameters that it requires should be added to Table A6.

The SPLASH model was described in detail and all parameters defined in (Davis et al., 2017). The only difference to their model version is that we include a soil texture-dependent water holding capacity instead of a globally uniform value. This is now stated more explicitly. Modified text reads:

Soil moisture (θ), AET, and PET are simulated using the SPLASH model (Davis et al., 2017), which treats soil water storage as a single bucket and calculates potential evapotranspiration based on Priestley and Taylor (1972). ***The only difference to the model version described by (Davis et al., 2017) is that*** *we account here for a variable water holding capacity calculated based on soil texture and depth data from SoilGrids (Hengl et al., 2014). A detailed description of the applied empirical functions for calculating plant-available water holding capacity from texture data is given in Appendix D.*

Appendix D then provides a detailed description of how we derived water holding capacity values around the FLUXNET sites. We decided not to include further descriptions of the SPLASH model here in order to save space and to focus on what is implemented also in the *rpmodel* R package.

> It seems that the evaluation set included the calibration set plus additional sites. A more robust evaluation would use an independent set of sites for evaluation. I recommend presenting results for the set of independent sites to help understand how the model works out-of-sample.

We thank the reviewer for this important point. Due to the small number of parameters that we calibrated simultaneously (1 for ORG and BRC, 3 for FULL setup) and due to the large amount of data from a wide variety of sites, the risk of over-fitting is small. This is confirmed by the additional calibration and evaluation we performed. Results show that the calibrated parameter values vary within around 1% across the entire set of out-of-bag calibrations. To clarify this point, we added content as follows.

We added text in the methods description in Section 3.3 Model calibration:

To test the robustness of the calibration and evaluation metrics, we additionally performed out-of-sample calibrations for the FULL setup where the training set included data from all but one site. The test dataset, used to calculate R2 and RMSE, contained only data from that single left-out site.

We added text and a new figure in a new results sub-section:

4.1 Calibration results

The calibration of model parameters, done with data from all calibration sites simultaneously, yielded values that closely matched the means across parameter values derived from the out-of-sample calibrations (Fig. 2). This confirms the robustness of the calibration and a negligible degree of overfitting. Similarly for the evaluation metrics, the R2 and RMSE values reported from evaluations against data from all evaluation sites pooled yielded values that closely match the means across the out-of-sample evaluation metrics (each calculated with data from the single left-out site). This analysis also shows that the distribution of the evaluation metrics is skewed, with evaluations against a few sites indicating particularly relatively performance (R2 below 0.5 for ZM-Mon, AR-Vir, and FR-Pue), while the most frequent values indicate very good model performance (evaluations at 21 sites giving R2 values of above 0.8). Because the out-of-bag calibrations are computationally very demanding, we performed this analysis only for one setup (FULL) and below report evaluation metrics done with pooled data from all evaluation sites.

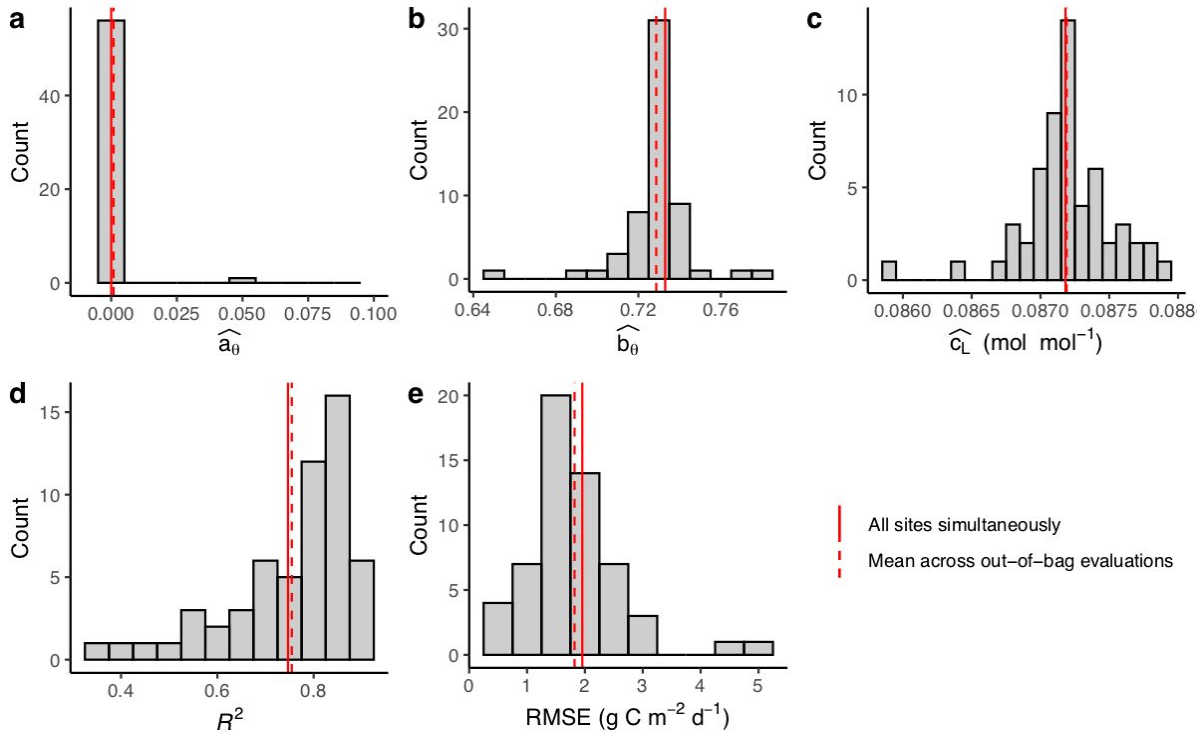

Figure 2*.Out-of-sample calibration and evaluation results. (a-c) Distribution of parameter values from calibrations where data from one site was left out for each individual calibration. Parameterŝaθand ˆbθare unitless. (d, e) Distribution of evaluation metrics calculated on data from the left-out site based on simulations with model parameters calibrated on all other sites' data. Solid red vertical lines represent the parameter values calibrated with data from all calibration pooled. These are the values reported in Tabs. 3 and 4. Dashed red lines represent the mean across values from out-of-bag calibrations and evaluations*

The variable I_{abs} is not defined in the text (only in equations)

We added on line 118:

I_{abs} is the amount of absorbed light and φ_0 is the intrinsic quantum yield efficiency.

I recommend having the main text or the SI present the full model used predict GPP in its entirety. While I can piece it together across equations, a combined equation would help me understand the model in its complete form

Unfortunately, to some degree, piecing it together is inevitable in view of the monstrous algebraic expression this would yield. To facilitate this point, though, we added a summary of the theory for GPP in Appendix F2:

To sum up, the P-model calculates GPP as

$$\mathrm{GPP} = I_{\mathrm{abs}}\, \varphi_0(T)\, \beta(\theta)\, m'\, M_C\ ,$$

where

$$m' = m \sqrt{1 - \left(\frac{c^*}{m}\right)^{2/3}}\ .$$

and

$$m = \frac{c_a - \Gamma^*}{c_a + 2\Gamma^* + 3\Gamma^* \sqrt{\frac{1.6\eta^* D}{\beta\,(K+\Gamma^*)}}}\ .$$

I_{abs} is the absorbed light (taken as fAPAR x PPFD, mol m-2), φ_0 is the temperature-dependent intrinsic quantum yield, $\beta(\theta)$ is the soil moisture stress factor, and M_C is the molar mass of carbon (g mol-1).

The * in Table 1 is not defined (I think it is the footnote).

We added the asterisk in the footnote.

More detail is needed about the comparison to other GPP products described in the first paragraph of the discussion. Were they using the same evaluation dataset? Where their evaluations on out-of-sample sites or the same sites used in calibration? Also, it seems the P-model is only marginally better than the VPM and slightly worse that the annual MODIS GPP. Please provide more information and context for interpreting these comparisons.

We write on line 457:

Unfortunately, we cannot present a direct comparison between these models, based on data from identical dates and sites. A targeted model intercomparison may address this.

In other words, we cannot state why the P-model performs better or worse than other comparable models. (Stocker et al., 2019) found that all the remote sensing-based models they investigated (most of which are referred to here as well), exhibited a systematic bias under drought conditions. The evaluation we provide here (Fig. 6) indicates that the P-model in its FULL setup largely resolves this issue. We further show that this resolution also leads to better predictions across all other scales investigated here (spatial, annual, seasonal, daily anomalies, see Tabs. 3 and 4). Due to missing published information on the other models' performance across all scales, we cannot provide a comparison at this level of detail.

References

Davis, T. W., Prentice, I. C., Stocker, B. D., Thomas, R. T., Whitley, R. J., Wang, H., Evans, B. J., Gallego-Sala, A. V., Sykes, M. T. and Cramer, W.: Simple process-led algorithms for simulating habitats (SPLASH v.1.0): robust indices of radiation, evapotranspiration and plant-available moisture, Geoscientific Model Development, 10(2), 689–708, 2017.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G. B., Walsh, M. G. and Gonzalez, M. R.: SoilGrids1km--global soil information based on automated mapping, PLoS One, 9(8), e105992, 2014.

Stocker, B. D., Zscheischler, J., Keenan, T. F., Colin Prentice, I., Seneviratne, S. I. and Peñuelas, J.: Drought impacts on terrestrial primary production underestimated by satellite monitoring, Nature Geoscience, 12(4), 264–270, doi:10.1038/s41561-019-0318-6, 2019.