Response to Reviewer 1

The comments of Reviewer 1 are accessible here:
https://doi.org/10.5194/egusphere-2025-1260-RC1

Below, we provide a point-by-point reply to the review. Text by the reviewer is indented and in blue font. Our reply is in black font and not indented. For ease of referencing our replies, we numbered them.

General comments:

> Manuscript by Paredes et al. titled "rsofun v5.0: A model-data integration framework for simulating ecosystem processes" describes an R-package that is mainly built around the P-model for efficient simulations and model-data synthesis workflows. Authors present a common use case where they calibrate the model with a single data-stream using a Bayesian approach and propagate the calibrated parameter uncertainty to model outputs.
>
> The objectives and the rationale of the study are clearly stated, and the manuscript is written in a clear and concise manner. I appreciate the efforts of the authors and developers for this development towards reprudicle modeling results and a lower bar of entry to vegetation modelling.
>
> While I would like to highlight that the reported development is in the interest of the GMD community, I found the writing style of the paper a bit different than typical GMD papers where many scientific and technical details are omitted with referrals to the vignettes. Below I commented on the manuscript parts that could be enriched, and I ultimately defer to the editor's decision but I suspect that it currently is not aligning quite well with categorization as a GMD "model description paper" which are expected to be comprehensive, detailed, complete and rigorous. Please consider supplementing the manuscript with deeper details and discussions.

[r1.1] We highly appreciate the useful comments and the reviewer's appreciation of our efforts to enable reproducible modelling and model-data integration. Our revisions will be made with a focus on providing a comprehensive documentation and demonstration of all functionalities implemented within the *rsofun* R package, and on

making the style, depth, completeness, and rigor of our study better aligned with the standards of GMD. Specifically (and in summary), we will additionally include:

- A demonstration of MCMC convergence and related diagnostics (correlation plots, Gelman-Brooks-Rubin diagnostic, …).
- A detailed description of the formulation of likelihood functions, uncertainty estimation based on the MCMC output, including the separation of model structural error and parameter error.
- A comparison of different calibration setups with different selections of simultaneously calibrated parameters and a detailed description of respective likelihood functions,
- A demonstration of the effect of simultaneously calibrating to flux time series and (static) traits data (ratios of the leaf-internal to ambient CO_2 concentration, and V_{max} over J_{max}).
- A larger set of model calibration data, representing flux measurements and traits from multiple sites, representing a wide diversity of environmental conditions.
- A more comprehensive and representative evaluation of model predictions, including an analysis of model prediction errors across sites for which data was not used for model calibration.
- A comparison of the results from a model calibration using BayesianTools vs. a non-Bayesian approach using GenSA (also implemented in *rsofun*).
- A brief overview of the architecture and implementation of the model code.
- All relevant details for reproducing simulations (in addition to code provided already along with our initial submission), and for reproducing model forcing and evaluation data.

The general feedback provided by the reviewer regarding the style and depth of model documentation and evaluation raises important points that we extensively discussed while designing this study. The following points motivated choices for our initial manuscript and will be considered also for the revisions of our manuscript:

- The technical details (all equations) of the P-model are comprehensively described and the model is evaluated against observations-derived GPP from eddy covariance ecosystem flux measurements from 126 sites in (Stocker et al., 2020). In contrast, the present manuscript presents the implementation of that model as an R package and how it can be used in combination with Bayesian methods for model-data integration.

- An evaluation of the Bayesian model calibration and the skill of predictions quickly gets into fundamental research questions about uncertainty and generalisability related to ecosystem characteristics and the environment. We wanted to avoid deviating into such questions. Instead, our intention here was to demonstrate a use case of P-model calibration, sensitivity analysis, and uncertainty estimation using Bayesian statistics.

Our revised manuscript will provide a comprehensive reference for how *rsofun* can be used and its results interpreted - thus serving as a *methodological basis* for addressing scientific questions related to modelling GPP in general. It will demonstrate the implementation and interpretation of the results from the Bayesian statistical methods in combination with observational data - widely used for analysis and modelling in ecosystem sciences - while limiting the scope by excluding investigations with wider ecological relevance.

We would like to note also that work that goes into publishing relatively complex code as an R package is substantial. The *rsofun* repository contains 2680 lines of R source code, and 11,360 lines of Fortran source code. The package is relatively complex as low-level code is in Fortran and R communicates with Fortran via a C wrapper. The package therefore needs to be compiled upon installation for standard test runs on a number of different platforms. Any error during these tests will prohibit publication of the package on the central repository for R packages (CRAN). Tests enable stable and reliable use of the package by any user on any machine and is an essential ingredient to making our scientific advance an Open Science resource. The work that goes into making this work remains somewhat "hidden" behind the paper, but its value should be taken into account when assessing the impact of the work.

Specific comments:

> P2.L36-39: The strength of Bayesian approaches in combining information from multiple sources and scales could also be emphasized here. Also, van Oijen (2017) could be a nice addition to the citations.

[r1.2] We will add a demonstration of the effect of simultaneously calibrating to flux time series (from multiple sites) and (static) traits data. For the latter, we will use two variables: the ratio of the leaf-internal to ambient CO_2 concentration, and the ratio of V_{max} over J_{max}. These two traits are directly informative for the specification of two key model parameters - the unit cost ratio (β) and the marginal cost of maintaining J_{max}, respectively. Data for these two traits were used in (Wang et al., 2017) for a direct

specification of these two parameters. Simultaneously calibrating to these two traits and GPP will enable a systematic assessment of the correlation structure among multiple fitted parameters.

> P2.L59: Would the authors consider rephrasing this sentence? Currently, it reads as if a novel solution is about to be presented whereas Bayesian calibration of a process-based vegetation model is done more times than I can count. Perhaps try presenting it as an application or an implementation of an existing solution.

[r1.3] We will rephrase this sentence to better reflect that calibration to data is also needed in parameter-sparse EEO-based models. However, please note that we wrote "*a* solution" (and not "*the* solution").

> P3.L67: It's a pity that the paper doesn't present this more sophisticated, and perhaps more realistic and valuable setup.

[r1.4] We will add a demonstration of the effect of simultaneously calibrating to flux time series (from multiple sites) and (static) traits data. See also [r1.2].

> Table 2: Could the parameters that were held fixed for the calibration be marked with a different character than asterisk (*) since soilm_thetastar and kc_jmax also have an asterisk in their symbols? Or the table caption could be modified to read "... marked with an asterisk in the last column"

[r1.5] We will revise the use of the asterisk symbol to avoid confusion.

> P6.L34: Could you elaborate as to which functions are included in this set and why/how they were chosen?

[r1.6] In the revised manuscript, we will provide a comparison of different calibration setups with (i) different selections of simultaneously calibrated parameters, and (ii) different calibration target variables. The different cost functions provided through the package will be used for the different setups, thereby demonstrating package functionalities more comprehensively.

> P7.L44: As GPP is not a measured but a derived quantity, could you please mention the approach used to derive it? FLUXNET2015 dataset is fairly well-documented but for the sake of being more complete, please also add how the GPP values were gapfilled and aggregated to daily values (which I assume is the model time step).

[r1.7] We will provide all relevant details for reproducing simulations (in addition to code provided already along with our initial submission), and for reproducing model forcing and evaluation data. This includes a description of the exact FLUXNET-standard variable, the filtering of data, and the selection of the multiple sites for which data will be used in the revised manuscript. For a description of the gapfilling technique, the reader will be referred to (Pastorello et al., 2020).

> P7.L61: Undermined how? Could you please elaborate on how the convergence was affected so that the readers can follow and apply the same logic in their applications when needed? Was there a strong correlation structure? Were the chains getting stuck? Was the result the same with different algorithms? Different chain lengths? The text also mentions before that kc_jmax and beta_unitcostratio (P3.L83) have previously been calibrated separately and fixed in this study, which was somewhat agreeable, but now saying that you decided to hold them constant because calibrating them with other model parameters undermined convergence is confusing. Please clarify and reconcile.

[r1.8] Answers to these questions will form a central pillar in our revised manuscript. Three calibration setups will be compared with a view to the questions mentioned:

- Setup 1: global, reduced parameter set (as in initial manuscript version), only GPP as target
- Setup 2: global, full parameter set, only GPP as target
- Setup 3: global, full parameter set, GPP and traits as target

We expect Setup 2 to yield wider posteriors than from Setup 1, and that posterior distributions will be narrowed again by Setup 3. This experimental design will allow us to demonstrate the robustness (or absence thereof) of the MCMC and the usefulness of using traits for simultaneously calibrating with fluxes.

Here, ‘global’ refers to using GPP data from 15 sites and traits data from other sites (around 16 sites for ratios of V_{max} over J_{max} and up to 400 sites for ratios of CO_2

concentrations), covering a wide environmental range. The GPP site selection is done by considering the following points:

- Good-quality GPP data available for >15 years.
- Stratified sampling with respect to the climate zone and vegetation type.

The selected GPP sites are:

Site	Long.	Lat.	Elevation	Climate zone	Vegetation type	Year start	Year end
US-MMS	-86.41	39.32	275	Cfa	DBF	1999	2020
DE-Hai	10.45	51.08	438.7	Cfb	DBF	2000	2019
US-Ha1	-72.17	42.54	340	Dfb	DBF	1992	2020
GF-Guy	-52.92	5.28	40	Af	EBF	2004	2019
DE-Tha	13.57	50.96	380	Cfb	ENF	1997	2019
CZ-BK1	18.54	49.5	875	Dfb	ENF	2004	2019
US-NR1	-105.55	40.03	3050	Dfc	ENF	2000	2015
US-Wkg	-109.94	31.74	1531	Bsk	GRA	2005	2021
US-Var	-120.95	38.41	129	Csa	GRA	2001	2020
BE-Vie	6	50.3	490	Cfb	MF	1997	2020
US-PFa	-90.27	45.95	470	Dfb	MF	1997	2014

> P7.L62: Also, on Table 2 (because they're fixed) no range to vary them was given for those parameters that were held constant in calibration. But I believe for them to appear in the sensitivity analysis (Fig 2) you varied them in some range. In fact, I can see that in the vignette these are specified, but please include those ranges also somewhere in the manuscript for clarity and reproducibility.

[r1.9] We will re-run the sensitivity analysis and will provide ranges for each parameter for which the sensitivity was quantified.

> Figure 2: Would it be possible to add the symbols to the figure? Because the paragraph right before (P7.L56-62), refers to these parameters with symbols while the figure refers to them with parameter names which requires the reader to go back to Table2 to do the mapping. Admittedly it's a small list, but I suspect it wouldn't be difficult to add symbols to the figure. Same goes for Figure 3.

[r1.10] Figures 2 and 3 will be revised to display symbols.

> P8.L75: 24K iterations sounds like an interesting choice. A more typical number would be, for example, 50K or 100K. Could you please explain in the text how you decided on this number (since the goal of the package and paper is to lower the bar and provide guidance to the audience)?

[r1.11] The length of the MCMC chains will be adjusted to 50 k.

> P8.L77-78: I really appreciate the vignettes but I feel like some of the results reported there should go together with the paper (e.g. in the supplement). I think vignettes are a lot more practical and ultimately more useful to the end users, and I wouldn't object if this was an open source scientific software journal but I expect GMD papers to be more complete. For example there is a kphio_par_a and kphio_par_b correlation discussion in the vignette which is completely missing from the paper.

[r1.12] The revised manuscript will cover more comprehensive contents (see also [r1.1]) - also those that are now exclusively given in vignettes.

> P9.L87: (from here onwards, including Figure 4) I believe there is a mix up in the way "model error" term is used. Perhaps the authors meant "error uncertainty" instead of "model error"? It is true that the credible interval is solely concerned about the uncertainty in the model parameters. Predictive interval here, however, is concerned with the overall residual error between the model and the data. In other words, the way the error term was jointly fitted in the calibration makes it intractable to decompose this term into data and model error. It is also known to dominate and cause overestimation of predictive uncertainty. Please refer to the relevant literature and revise (e.g. van Oijen, 2017).

[r1.13] These terms appear to be somewhat inconsistently used in the literature. In the revised manuscript, we will use terms ‘model structural error’ and ‘parameter error’ as used in (Dietze et al., 2013, 2018).

> Results - I was a bit surprised to see no quantitative reporting of the model improvement after the calibration. The only visual comparison (Figure 4) reports the posterior performance with no reference to prior performance, and only for one year. While showing a single year is useful for practical purposes, please consider providing results for all years (e.g. in the supplement):
>
> - It would be interesting for the readers to see how performance in different years compare, also with quantitative metrics.
>
> - Furthermore, even though measurements from years 2013 and 2014 were deemed problematic, it would be good to show what the calibrated model predicts for these out-of-sample years that the calibration did not see.
>
> - Last, but not least, there was no mention of further posterior predictive checks/diagnostics. One can for example plot residuals against predictors and employ formal tests that measure where the observed data falls on the distribution of simulated data. Is there a pattern in the residuals? Were the correct distributional assumptions made? Comparing credible and the predictive intervals is only a (very) limited part of the story.

[r1.14] The revised manuscript will provide a deeper evaluation of model predictions. This will include (see also [r1.1]):

- A more comprehensive and representative evaluation of model predictions, including an analysis of model prediction errors across sites for which data was not used for model calibration.
- A comparison of the results from a model calibration using BayesianTools vs. a non-Bayesian approach using GenSA (also implemented in *rsofun*).

The calibration and evaluation of model predictions will use data from multiple sites. Therefore, some of the reviewer’s points will be obsolete. As also noted in [r1.1], we will strike a balance between conciseness and depth, avoiding evaluations that address research questions with more specific ecological relevance.

> Discussion - Please consider enriching the discussion section with the following:

- Were the design choices adequate? I mean all the decisions from coupling the model with BayesianTools as a package to the prior and likelihood forms you selected?

- What does it take to transfer this calibration to another site/variable, multiple sites/variables?

- How do you recommend iteratively updating these results as new data becomes available (either more of the same type of observations or with new data sources)? Does your implementation allow re-reading its own outputs as inputs?

- What are the other limitations of this study? What are your outlooks?

- While I understand that it is not the main goal of the paper, the inexperienced readers could benefit from pointing to the literature that provides guidance on the peculiarities of model calibration, e.g. see MacBean et al. 2016, van Oijen 2017, Oberpriller et al. 2021, Cameron et al. 2022. Therefore, in addition to points above, please also consider adding some discussion along those lines since the section is rather thin at the moment. (MacBean et al. 2016 https://doi.org/10.5194/gmd-9-3569-2016, van Oijen 2017 https://doi.org/10.1007/s40725-017-0069-9, Oberpriller et al. 2021 https://doi.org/10.1111/ele.13728, Cameron et al. 2022 https://doi.org/10.1111/2041-210X.14002)

[r1.15] The revised manuscript will provide a deeper discussion of the results. The points suggested by the reviewer will help us to enrich contents and to better connect the Discussion section to the published literature (specifically the papers suggested here).

P11.L13: "complementary observational constraints" such as?

[r1.16] This refers to traits data, which will be additionally used for calibration in the revised manuscript.

P11.L16: Please elaborate some more. Is it a weakness of the study that FvCB parameters were kept constant? What was the reasoning? Is it recommended to do so? What are the expected challenges there? Are there future plans to address this?

[r1.17] The FvCB parameters were not kept constant but instead were predicted by the model. They are therefore not ‘model parameters’ as the term is used here. The revised manuscript will demonstrate how traits data informs the calibration of the model.

We thank the reviewer for the excellent input and the useful references. We hope that our revised manuscript will satisfy the points raised and the standards for publication in GMD.

References

Dietze, M. C., Lebauer, D. S., and Kooper, R.: On improving the communication between models and data, Plant, Cell & Environment, 36, 1575–1585, https://doi.org/10.1111/pce.12043, 2013.

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., and White, E. P.: Iterative near-term ecological forecasting: Needs, opportunities, and challenges, Proc. Natl. Acad. Sci. U.S.A., 115, 1424–1432, https://doi.org/10.1073/pnas.1710231115, 2018.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. de, Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. di, Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Sci Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Stocker, B. D., Wang, H., Smith, N. G., Harrison, S. P., Keenan, T. F., Sandoval, D., Davis, T., and Prentice, I. C.: P-model v1.0: an optimality-based light use efficiency model for simulating ecosystem gross primary production, Geoscientific Model Development, 13, 1545–1581, https://doi.org/10.5194/gmd-13-1545-2020, 2020.

Wang, H., Prentice, I. C., Keenan, T. F., Davis, T. W., Wright, I. J., Cornwell, W. K., Evans, B. J., and Peng, C.: Towards a universal model for carbon dioxide uptake by plants, Nature Plants, 3, 734–741, https://doi.org/10.1038/s41477-017-0006-8, 2017.