1 General Comment

I thank both reviewers for their thoughtful and constructive feedback on my manuscript. I would like to explain the delay in submitting my revision. To fully address the comments, I needed to run additional simulations. During the revision period, the original HPC used for these simulations (Beluga) was decommissioned, and I migrated to a new system (Rorqual). On November 6, a major hardware failure in one of Rorqual's storage components prevented access to files, and the system remained offline for an extended and uncertain period. After about two weeks, I attempted to migrate to alternative HPC systems, first Trillium and then Nibi. However, neither system was suitable due to limitations on computing time and the computational environment. On November 21, Rorqual came back online, and I resumed the simulations there. I appreciate your patience and understanding regarding the resulting delay.

The main focus of my revision is to address the reviewers' concerns about the robustness of the optimized parameter values. The exact parameter values may depend on the selection of grid cells during the optimization and may also be affected by equifinality. Following Reviewer 1's suggestion, I first created an ensemble of optimized runs by repeating the optimization with different random selections of 160 grid cells. The resulting optimized parameter sets varied considerably across runs. Several factors may explain this variability. One possibility is that a sample size of 160 grid cells was insufficient. Another is that model performance is controlled by only a small subset of parameters, which therefore converged consistently, while the remaining parameters varied largely by chance with little effect on performance. A third possibility is that the model exhibits strong equifinality, such that small differences in the optimization set-up lead to substantially different parameter values.

As a first step, I increased the sample size from 160 to 360 grid cells, a number that is still computationally feasible. Prior to re-running the optimization, I modified the code to save the full set of parameter values and the corresponding objective-function score from each individual optimization, rather than only summary statistics. I then used these outputs to construct a Gaussian Process emulator. The emulator statistically relates model performance to the parameter sets and can replace the original model to estimate performance for new parameter combinations sampled within the training range. A key advantage of the emulator is its negligible computational cost, which allows rapid exploration of a large number of parameter sets. As a second step, I used the emulator to assess whether the model exhibits equifinality and to determine whether this equifinality is driven by a small or a large number of parameters. As a third step, I used the emulator to identify alternative parameter sets with similarly good performance and employed these emulated parameter sets to generate an ensemble of model simulations. This ensemble was then used to assess the sensitivity of model results to different parameter values that yield very similar performance.

My new results are similar to those obtained using the previous set of 160 grid cells. While the optimization improves GPP, this improvement does not automatically translate into an improvement in globally accumulated NBP. This is not surprising, as NBP is not included in the optimization due to the lack of reliable globally gridded NBP datasets. Because the globally accumulated NBP obtained with the optimized parameter set is not improved relative to the default simulation, I have removed the analysis of future NBP projections and instead focus on the robustness of the results during the historical period.

Using a Gaussian Process emulator, I show that the model exhibits equifinality and that this equifinality is driven by many parameters rather than just a few. By using the emulator to generate alternative parameter sets with comparable performance, I find that these sets produce similar, but not identical, results. Global GPP values from these alternative sets can be either lower or higher than those of the optimized simulation. The resulting uncertainty range is approximately ± 2 PgC yr^{-1} around a total flux of about 128 PgC yr^{-1}. Furthermore, the ensemble of simulations derived from the emulated parameter sets perform similarily well when assessed against all other reference data sets. I conclude that the results are robust despite the pres-

ence of high-dimensional equifinality. The following sections provide a detailed response to the reviewer's comments, with my answers written in blue.

2 First Reviewer

I thank the reviewer for their thoughtful and constructive feedback on my manuscript. Please find my point-by-point responses below.

REVIEWER: This study applies a machine learning-based Genetic Algorithm (GA) and multiple global Earth observation datasets to systematically optimize poorly constrained parameters in the CLASSIC land surface model. The optimization is conducted over a long historical period (1701-2020), simultaneously targeting multiple variables and using multiple observational data streams, aiming to improve historical simulation performance and assess future terrestrial carbon fluxes under the SSP5-8.5 scenario. Despite these strengths, several issues may limit the scientific impact and clarity of the manuscript. My detailed comments are as follows:

REVIEWER: L233: The global representativeness of the randomly selected 160 grid cells should be evaluated. These cells may not capture regional differences or small-scale processes, and if the selected grids differ substantially from the target regions, the optimized parameters may not be suitable for local applications. While the 160 grids were randomly selected, it is not stated whether multiple random samplings were performed to test the stability of results. Different random seeds could lead to different optimal parameter sets.

ANSWER: I have now performed multiple optimizations using different combinations of grid cells and found that the optimized parameter values differed considerably, confirming the concern you raised. I then increased the number of grid cells from 160 to 360. Owing to the substantial computational cost, I did not test multiple combinations of these 360 grid cells. Even if different grid cell selections were to yield different optimized parameters, this would not necessarily indicate undersampling but could instead reflect equifinality. I therefore conducted a single optimization using 360 grid cells and subsequently assessed equifinality using a Gaussian Process emulator trained on the data generated during the optimization. The results show that different parameter combinations with similarly good performance do not substantially affect global carbon fluxes. I therefore conclude that the uncertainty associated with equifinality is relatively small. Note that I did not document the multiple 160-grid-cell optimizations in order to maintain the focus of the manuscript.

REVIEWER: Using the same set of observational data for both fitness evaluation and parameter optimization lacks an independent validation set or cross-validation. This may result in good performance on the training data but poor generalization capability.

ANSWER: The new optimization performed for this revision is based on 360 out of 2,444 grid cells (15%). Thus, 85% of the grid cells used in the global evaluation were not included in the optimization. Moreover, the optimization is based on a spinup using the default parameter set, whereas the optimization itself is performed only over the transient period (1701-2022). For the global evaluation, the optimized parameter sets are applied during both the spinup and the transient period. As a result, the values of the grid cells used in the optimization differ from those in the global simulation. I therefore need to include all grid cells in the global evaluation. I now included this argument in section 2.4 Model Evaluation: "*The same data streams used during optimization are also used to assess model performance across all 2,444 land grid cells. This global evaluation includes the 360 grid cells used during optimization because the optimization applies default parameters during the spinup and optimized parameters only during the transient period (1901-2022). In contrast, the global evaluation uses the optimized parameter set consistently during both spinup and transient periods. Therefore, it is necessary to assess model performance for all grid cells, including*

those used in the optimization."

REVIEWER: L235: The computation time of two weeks is substantial, yet the manuscript does not specify the convergence criteria, number of iterations, or early stopping strategy, raising concerns about potential waste of computational resources. If the solution space is large, GA may still remain trapped in suboptimal solutions.

ANSWER: This information is shown in Figure 4 and is now described in the text in section 3.1 Optimization: "*The number of grid cells was selected such that the optimization could be completed within two weeks of wall-clock time on the Digital Research Alliance of Canada's high-performance computing system (Rorqual). The optimization was run for 25 generations, with 100 individuals per generation, resulting in a total of 2,500 simulations.*" The improvement in performance becomes increasingly gradual from generation to generation, and Figure 4 illustrates that very little gain can be expected after generation 25. One might argue that computational time could have been saved by stopping the optimization earlier. However, this is not evident unless additional simulations are conducted that demonstrate diminishing progress. While I am confident that the solution could be improved by adding more iterations, I believe that the cost-benefit ratio would become too large. It is possible that the solution represents a local rather than a global optimum, particularly given the model's high-dimensional equifinality. I therefore explore this issue further using a Gaussian Process emulator. Even if the result reflects only a local optimum, it still represents a clear improvement over the default parameter set. Finally, in the absence of systematic parameter optimization, parameter values must be hand-tuned, which is a cumbersome approach that is far more likely to yield a suboptimal solution.

REVIEWER: L253-258: Are the six land surface variables (ALBS, GPP, HFLS, HFSS, LAI, LST) weighted equally in the cost function? Different variables may differ greatly in importance (e.g., GPP is more critical for the carbon cycle), but the manuscript does not explain how weights were assigned.

ANSWER: I assign all variables equal weight. I have considered weighting them differently, but that immediately raises the question of how to determine the weights. One could argue that GPP is more critical for the carbon cycle, but the carbon, energy, and water cycles are all coupled and must remain consistent. It could also be argued that larger weights should be assigned to variables with lower observational uncertainty, but such uncertainties are difficult to quantify. In my view, defining weights opens the door to very subjective discussions that I would prefer to avoid. From my perspective, all aspects of the carbon, water, and energy fluxes should be considered equally important. I have now added this argument in section 2.4 Model evaluation: "*Although this research focuses on carbon fluxes, all datasets are weighted equally because the carbon, energy, and water cycles are coupled and must be represented consistently.*"

REVIEWER: L270-272: The robustness analysis was conducted with fewer grid cells, a shorter time period, and fewer generations. The representativeness of these reduced settings should be discussed in the manuscript.

ANSWER: I have now replaced this analysis with the emulator-based approach, which provides a more rigorous and consistent framework.

REVIEWER: L299: The finding that model performance stops improving after 25 generations may be due to GA parameter settings. This should be considered and discussed.

ANSWER: Optimizing the optimization process is challenging given the large number of different possible combinations of selection, crossover, and mutation functions and corresponding hyperparameters. I now discuss this topic in the Discussion section: "*One aspect not addressed in this study is the sensitivity of the results to the choice of optimization strategy, including the selection, crossover, and mutation operators, as well as their associated hyperparameters. Different combinations of these components may yield more efficient convergence or lead to different optimized parameter sets. However, systematically tuning the optimization itself is challenging due to the high computational cost. Although alternative settings could*

be tested using shorter simulations or fewer grid cells, it is unclear whether such results would generalize to the full optimization setup. Future research could therefore explore more efficient ways of identifying effective optimization strategies, for example by using an emulator to guide the design and tuning of the optimization procedure."

REVIEWER: L315: The statement that "some variables did not improve" is made without analyzing the possible causes. This could be due to structural model errors rather than parameter settings, or uncertainties in the observational datasets. The discussion should include potential reasons and possible future improvements.

ANSWER: The parameters included in the optimization were identified in a global sensitivity analysis by S.N. et al. (2025) as the most influential for carbon and turbulent heat fluxes. It is therefore expected that the optimization did not substantially improve variables such as albedo and land surface temperature. Nevertheless, these variables were included in the observational dataset as safeguards. I have now added the following text to section 3.3 Global Model Performance: "*The limited impact of the optimization on variables such as albedo and surface temperature was expected, as the optimized parameters predominantly influence carbon and turbulent heat fluxes, as discussed further in the Discussion section.*".

REVIEWER: L338: Although the optimized simulation is slightly better than the default in some statistical metrics, the differences are described as "too minor to be considered meaningful." The manuscript should discuss why optimizing 28 parameters results in only limited improvement in NBP, which may be related to observation errors, insufficient parameter representativeness, or model structural deficiencies.

ANSWER: Please note that the optimization leads to a substantial improvement in model performance for leaf area index and gross primary productivity (GPP). The model was not optimized for net biome productivity (NBP), as no reliable globally gridded observational datasets for NBP are available. The rationale was that improvements in GPP might translate into improved NBP estimates, given that NBP is defined as GPP minus ecosystem respiration and emissions from disturbances such as wildfires. However, the results show that NBP does not automatically improve when GPP improves. This outcome is likely because the optimization does not constrain ecosystem respiration, for which no reliable globally gridded datasets exist. As NBP did not improve in response to the optimization, I have removed the previous analysis of future NBP projections. The Discussion section now includes the following text outlining how NBP could be optimized more effectively in future work: "*While the optimization clearly improved global GPP when evaluated against GOSIF reference data, this did not yield an obvious improvement in global NBP when evaluated against GCB2024 NBP (Friedlingstein et al 2024). In particular, neither the default nor the optimized simulations reproduces the GCB2024 NBP trend. One way to improve simulated global NBP would be to optimize CLASSIC parameters using the GCB2024 NBP estimates as a target. However, this would require running the model for all grid cells during the optimization process, which is computationally prohibitive. A feasible alternative would be to conduct the global optimization using an emulator. This would require constructing a training dataset based on global CLASSIC simulations, with parameter values sampled randomly across their respective uncertainty ranges. Once trained, the emulator could first be used to perform a global sensitivity analysis to identify the most influential parameters and then to optimize those parameters.*"

REVIEWER: L385: While two GA configurations were found to perform better than the default, the manuscript does not analyze their characteristics (e.g., differences in selection/crossover/mutation strategies) or why they perform better. Such analysis would help in better understanding the influence of GA settings on optimization results.

ANSWER: This section was based on test runs using fewer grid cells and shorter time periods, and it remains unclear whether the results can be generalized to the full optimization setup. I have therefore removed this section, particularly as the manuscript now places a stronger emphasis on uncertainty arising from equifinality. However, as noted above, I provide suggestions in the Discussion section on how different

optimization settings could be explored using an emulator.
REVIEWER: In the main text, some figures and tables could be moved to the supplementary materials to improve readability, such as Figures 1, 2, 7 and Tables 1, 2.
ANSWER: After careful consideration, I have reduced the number of figures to ten and the number of tables to three, which falls well within the range typically expected for a research paper.

3 Second Reviewer

I thank the reviewer for their thoughtful and constructive feedback on my manuscript. Please find my point-by-point responses below.
REVIEWER: The paper proposes a Genetic Algorithm-based framework for optimizing parameters in the CLASSIC land surface model, using multiple global Earth observation datasets. It finds that the optimized parameters significantly improve key variables including GPP, LAI, and sensible heat fluxes. The paper is generally well-written and is suitable for publication after addressing the following comments.

3.1 Major comments

REVIEWER: 1. The author notes that multiple datasets are used per variable "to reduce the risk of overfitting" and "help account for observational uncertainty". However, it seems like the paper does not rigorously incorporate observational uncertainties into the optimization. A more rigorous treatment, or discussion on this, of observational uncertainty would strengthen the robustness of the conclusions.
ANSWER: I now address this topic in the Discussion section as follows: "*Global reference data used for land surface model evaluation are subject to considerable uncertainties (Seiler et al, 2022). Many data assimilation frameworks treat observational uncertainty in a more formal manner than is presented here, commonly by specifying error covariance matrices for both model and observational errors (Tarantola, 2005). These covariance matrices determine how strongly observations constrain the model and help avoid overfitting in regions or periods of large uncertainty. However, estimating error covariance matrices reliably remains challenging, particularly in the absence of comprehensive uncertainty documentation for many Earth observation products (MacBean, 2022). The approach presented in this study is intentionally simpler. All observational data products are given equal weight, and no explicit assumptions are made about their error magnitudes or spatiotemporal error correlations. For example, consider two satellite-based GPP products. If both products indicate lower GPP than the model, the optimization will adjust parameters such that the modelled GPP is reduced. If, however, one product indicates higher and the other lower GPP, the performance score remains similar as long as the modelled GPP lies within the range spanned by the two products. In this sense, the approach implicitly downweights conflicting information and avoids tuning the model toward any single potentially biased product. While this procedure does not explicitly model observational error covariances, it achieves a similar practical outcome by reducing sensitivity to individual datasets and by preventing overfitting in the presence of observational disagreement. Importantly, it avoids introducing additional assumptions about poorly constrained error structures. Moreover, whereas covariance-based data assimilation primarily focuses on minimizing instantaneous model-data mismatch, the approach adopted here explicitly constrains multiple statistical properties of the system, including bias, centralized RMSE, seasonal timing, interannual variability, and spatial correlation (the five AMBER scores). Improvements in these higher-order diagnostics are not guaranteed through bias correction alone and therefore benefit from being constrained explicitly.*"
REVIEWER: 2. I am particularly concerned about the generalizability of the optimized parameters, which the paper does not fully address. Since the optimization uses Earth observations from the modern cli-

mate, it remains unclear whether these parameter values will remain valid under future climate conditions, potentially limiting the robustness of the projections. A discussion of this limitation can strengthen the manuscript.

ANSWER: Whenever a new parameterization is introduced in a model, developers typically select parameter values within an uncertainty range so that the model output matches observations from the modern climate. This kind of ad hoc tuning is common practice, and your criticism applies equally to it. Replacing ad hoc tuning with a more systematic approach is not different in principle; it is simply far more effective. Furthermore, I now dropped the analysis on future projections for the reasons explained above. Your concern about future climate conditions, therefore, no longer applies. Also, my new analysis on equifinality addresses your concern about generalizability, and shows that different parameter sets that perform similarly well can yield similar results.

REVIEWER: 3. The author acknowledges that the optimization is evaluated only in offline mode, with prescribed CO2 and meteorological forcing, and notes that a fully coupled setup would alter NBP feedbacks. It would strengthen the paper if this limitation can be emphasized more clearly in the conclusions, with a brief discussion of how coupled feedbacks might influence the results.

ANSWER: The discussion section elaborates on the potential use and impacts of optimized parameter sets: *"The model configuration used here is an offline simulation with prescribed [CO_2] and meteorological forcing. Therefore, the impacts of optimization do not alter atmospheric [CO_2] or the climate. In a fully coupled simulation, lower NBP would imply a faster increase in [CO_2] and temperature, affecting NBP in return. This feedback could be evaluated in an emissions-driven simulation where CLASSIC runs within CanESM and the carbon cycle is fully coupled. Such simulations would not only be scientifically relevant but also important for global climate change mitigation policy, as explained next."*. The text then continues discussing how exactly such experiments would affect climate change policies. I now briefly reiterate the potential application in fully coupled runs in the conclusion section: "*Although parameter tuning was conducted in an offline mode, the optimized parameter sets could be applied in online, emission-driven simulations in which the carbon cycle is fully coupled.*"

3.2 Minor comments

REVIEWER: L300: Figure 5a

ANSWER: Yes, thank you. I have changed 4a to 5a.

REVIEWER: L304: Figure 5b

ANSWER: Yes, thank you. I have changed 4b to 5b.

REVIEWER: Figure 10: caption does not mention (g) and (h)

ANSWER: I have deleted this figure from the analysis for the reasons described above.

REVIEWER: Maybe Figures 2 and 7 can be moved to supplementary materials.

ANSWER: After careful consideration, I have reduced the number of figures to ten and the number of tables to three, which falls well within the range typically expected for a research paper.