We thank Reviewer 1 for the detailed and constructive evaluation of our manuscript. The review raises important concerns regarding the evidence for overfitting avoidance, the apparent contradiction between the two independent float tests, the relationship between parameter correlations and identifiability, and the clarity of the Methods section. These comments have led to substantial improvements in the revised manuscript.

The most significant change is the addition of a new section (Sect. 5) presenting a three-dimensional validation in which the optimised parameters are implemented in the high-resolution IBI regional model and evaluated against ~1,430 independent BGC-Argo profiles and a surface chlorophyll-a product over 2017–2019. This directly addresses the reviewer's central concern about the strength of the independent validation evidence. In addition, the former "Portability" section has been reframed as an assessment of ensemble calibration rather than a demonstration of portability, resolving the contradictory framing identified by the reviewer. A pedagogical paragraph linking parameter correlations to identifiability and overfitting has been added to broaden accessibility, and the Methods section has been restructured with a roadmap paragraph to improve readability. The explicit formula for the observation likelihood has been added to the Appendix (Eqs. A1–A3), clarifying how the 20 metrics jointly enter the iIS algorithm. Finally, claims about the "richness" of BGC-Argo data have been toned down throughout.

Below, we address the Major Comment and each Specific Comment in turn. All line numbers refer to the original manuscript unless stated otherwise. Reviewer comments are in normal text; our replies are in blue.

Major Comments:

Parametric uncertainties are a major issue in biogeochemical ocean modelling and the application of Iterative Importance Sampling while testing different estimation strategies is a nice approach. I specifically like that the parameter fits are tested on independent observations which were not used during the fitting process. This should be common, but isn’t in ocean biogeochemical modelling due to a general lack of independent observational data. However, in the presented study the analysis of the test data sets appears somewhat hidden and two somewhat contradictory conclusions are drawn from it. The authors use two data sets of seasonal cycles and argue, with a relatively good fit of the optimized model to the first test data set, that the fitted models are “portable” (Ln 944, 985, 988), while the rather poor fit to the second test data set is regarded as indication that parameters should be estimated depending on location (Ln. 929, 1019ff).

These aspects need some attention, because the authors draw very interesting and far reaching conclusions by implying that they could overcome the long-standing challenge of parameter overfitting (Matear,1995) and parameter (un)identifiability (Ln. 20ff, 995ff, 1005). The authors relate their solution to the "richness" of Biogeochemical Argo Floats (Ln.19, Ln 80ff, 970ff). As somewhat contradictory conclusion, however, it is stated

already in the Abstract that different optimal parameter sets lead to statistically indistinguishable model results (Ln. 18; cf. also Tab.4). This shows that solutions for an optimal parameter set are not unique which might point towards overfitting. Thus, the overfitting-statement needs in my eyes stronger evidence. E.g. it should be ruled out that the calibration may be over-confident due to too few noisy data during fitting (e.g. Hermans et al. 2022 and Yang & Zhu, 2018 illustrate the problem of too tight posteriors (small variance) and over-confident calibration for such cases; it seems straight-forward that also correlations will be impacted). This holds especially, because the presented results rely on a single seasonal cycle while the authors state that their 1D representation of the ocean has major flaws (e.g., Ln 400ff) and some of the observations were “pseudo” (e.g., Ln 165ff). This might be achieved by presenting the performances of all optimal models on independent test data more convincingly. Note, that the test data should preferably cover different nutrient/light regimes, seasons and mixing conditions in accordance to what may occur during model application. In case the test data do not suffice to cover this range, It might be interesting to explore if the different optimal model solutions diverge under changing external conditions (cf. Taucher & Oschlies, 2011; Löptien & Dietze, 2017). Also, it would be nice if the authors could briefly explain somewhere prominently how correlations among parameters relate to overfitting and parameter identifiability to reach a wider audience.

In summary, I think that the study has very nice aspects, but some conclusions need in my eyes more evidence or the authors could tone down a bit. Further, I found the manuscript partly very hard to follow and it could be much more concise to gain it’s full potential. This holds especially true for the Method and Data Sections, where large parts could be moved to an Appendix (e.g. extensive formula). Currently, this part covers 20 pages and I find it very hard to keep overview in the many details that partly occur in to me unexpected subsections. The authors could also be more explicit on the underlying ideas.

We thank the reviewer for this thorough and constructive comment, which raises several interconnected issues. We address each aspect below and describe the corresponding changes to the manuscript.

1. Apparent contradiction between the two independent float tests

We agree that the original manuscript created a misleading contrast by framing one test float as evidence for "portability" and the other as evidence against it. This framing has been removed. The section previously entitled "Portability of the Optimised Ensembles to Other Bioregions" (original Sect. 4.5) has been retitled "Ensemble Calibration Against Independent BGC-Argo Floats" and now follows the new 3D validation section. Its purpose is no longer to demonstrate portability but to assess the statistical calibration of the optimised ensemble — specifically, whether the posterior uncertainty envelope is

accurate — using the RCRV bias and dispersion diagnostics. The word "portability" has been removed throughout the manuscript.

The role of primary independent validation is now carried by the new 3D validation (Sect. 4.6), which compares the optimised simulation (OPTI) against the reference simulation (REF) using the parameter set from the All-parameters strategy implemented in the high-resolution IBI regional model. Crucially, whereas the parameters were trained on a single seasonal cycle from 2015 in a 1D configuration, the 3D validation covers a independent period (2017–2019). This validation is evaluated against approximately 1,430 independent BGC-Argo profiles across three sub-regions (Northern basin, Iberian Atlantic, Mediterranean) and the MOBTAC multi-observation surface chlorophyll-a product over five regional boxes. OPTI outperforms REF in 12 of 18 variable–basin combinations, with particularly large improvements for total alkalinity, nitrate, and phosphate in the North Atlantic, and better-reproduced surface chlorophyll-a across the entire domain. These results demonstrate that the 1D-optimised parameters genuinely improve the biogeochemical parameterisation in a fully three-dimensional, dynamically resolved configuration, directly addressing the concern raised by Löptien and Dietze (2019) and Taucher and Oschlies (2011) . The Conclusion now states that the generality of the method should be tested by applying it to additional BGC-Argo floats in other oceanic regions, and that the resulting parameter sets should be evaluated to determine whether they further improve model performance across biogeochemically distinct basins.

2. Risk of overfitting and over-confident calibration

The reviewer rightly highlights the risk, described by Hermans et al. (2022) and Yang and Zhu (2018), that Bayesian calibration with limited data may produce artificially tight posteriors and over-confident predictions. Two independent lines of evidence indicate that this scenario does not apply to our results.

First, the RCRV dispersion diagnostic directly tests whether the posterior ensemble is over-confident. Dispersion values above 1.0 would indicate under-dispersive, over-confident posteriors — the situation described by Hermans et al. (2022) and Yang and Zhu (2018). Our values are approximately 0.3 for the North Atlantic float and 0.4 for the Mediterranean float (Fig. 9c, d), well below unity. The ensemble is therefore conservative (over-dispersive): the predicted spread is wider than the observed model-data errors. This is the direct opposite of the over-confident calibration the reviewer is concerned about. We have added an explicit reference to Hermans et al. (2022) and Yang and Zhu (2018) in the revised Discussion to acknowledge this concern and point to the RCRV evidence.

Second, the 3D validation provides the most powerful evidence against overfitting. The parameters were trained on a single seasonal cycle from 2015 at a single North Atlantic location. If the optimisation had fitted noise from this specific period and configuration, those parameter values would not improve a fully three-dimensional model over a

completely different period (2017–2019) against ~1,430 independent BGC-Argo profiles spanning multiple regions and seasons, and against the MOBTAC multi-product reprocessed surface chlorophyll-a product over five regional boxes. The fact that OPTI outperforms REF across the majority of variable–basin combinations, and also produces better surface chlorophyll-a fields across the domain, demonstrates that the optimised simulation captured genuine biogeochemical signal rather than noise.

We acknowledge that calibrating on a single seasonal cycle at a single location is a genuine limitation, as the revised Discussion now states explicitly (Sect. 5). Extending the assimilation window to multiple years and multiple training sites would further strengthen confidence in the generality of the results, and we identify this as a priority for future work.

3. Clarification of "rich" in relation to long-term observing stations

The reviewer correctly identifies an ambiguity in our use of the word "rich." We did not intend to claim that one seasonal cycle is temporally richer than decades of data from long-term stations such as BATS or HOT. The advantage of the BGC-Argo dataset relative to the observations used in previous parameter-optimisation studies (e.g., Hunt et al. 1996; Schartau and Oschlies, 2003; Ward et al. 2013) lies in three complementary dimensions.

The first is multi-variable breadth. A single BGC-Argo float simultaneously provides 8 biogeochemical tracers (4 directly measured: Chl-a, POC, NO_3^-, O_2; plus 4 derived via CANYON-B/CONTENT: PO_4^{3-}, Si, DIC, TA), from which we construct 20 distinct metrics. Previous optimisation studies based on long-term stations typically assimilated far fewer tracers simultaneously — often chlorophyll-a alone, or chlorophyll-a plus one or two nutrients. Assimilating 20 metrics simultaneously provides substantially more independent, orthogonal constraints on the parameter space.

The second is temporal resolution. The BGC-Argo float provides vertical profiles approximately every 5 days, compared with the monthly sampling frequency of stations like BATS or HOT. This higher frequency resolves rapid biogeochemical transitions — such as the precise onset of the spring bloom, nutrient drawdown events, and deep-chlorophyll-maximum dynamics — that monthly sampling would alias or miss entirely.

The third is vertical resolution. BGC-Argo profiles are interpolated onto a 1 m vertical grid from the surface to 1000 m, whereas long-term stations rely on discrete bottle samples at typically 20–30 depths per cast. This near-continuous vertical coverage enables more accurate determination of depth-averaged quantities and more precise diagnosis of vertical-structure features such as the nitracline depth, the deep-chlorophyll-maximum depth, and the oxygen minimum.

In the revised abstract, "rich" has been replaced by "comprehensive, multi-variable" to remove the ambiguity. A clarifying sentence has been added in the Introduction (Sect. 1) to make the three dimensions of the dataset's advantage explicit. Regarding the pseudo-observations from CANYON-B/CONTENT, we note that these are not treated as equivalent to direct measurements: their uncertainties, which are larger, are propagated through the error budget (Sect. 3.6.4), and the negligible impact of their error covariances is demonstrated in Sect. 3.8.

4. Presenting performances of all optimal models on independent test data more convincingly

The new 3D validation directly addresses this request and goes beyond what the reviewer suggested. Whereas the parameters were trained on a single seasonal cycle from 2015, the 3D validation covers three full years (2017–2019), encompassing all seasons and mixing conditions; three sub-regions with fundamentally different nutrient and light regimes (Northern basin, Iberian Atlantic, oligotrophic Mediterranean); approximately 1,430 unique BGC-Argo profiles yielding 430,000–490,000 observation-model pairs per variable over 0–500 m depth; and multi-product reprocessed surface chlorophyll-a validated over five regional boxes. This provides the multi-regime, multi-season, independent test the reviewer was asking for.

The 3D validation uses the parameter set from the All-parameters strategy, because this is the strategy the paper recommends as the most practical and comprehensive approach. All three optimisation strategies produce statistically indistinguishable skill in the 1D framework (NRMSE reduction of 54–56%; Tab. 4) and indistinguishable RCRV scores (Kruskal-Wallis $p = 0.99$). Running all three strategies in the 3D model would be computationally expensive and, given their indistinguishable 1D performance, unlikely to reveal meaningful differences.

The alternative approach suggested by the reviewer — testing whether different optimal solutions diverge under changed external conditions (cf. Taucher and Oschlies, 2011; Löptien and Dietze, 2017) — was designed for situations where independent validation data are unavailable. Since we now provide extensive independent validation through the 3D experiment, this indirect diagnostic is no longer necessary.

5. Explanation of how parameter correlations relate to overfitting and identifiability

We agree that this link was not made sufficiently explicit for a general readership. A new paragraph has been added in the Discussion (Sect. 5) to provide this explanation. The key idea is the following.

Parameter correlations in the posterior distribution are a direct indicator of identifiability problems. When two parameters are strongly correlated, the data constrain only a combined effect of these parameters — for instance, the ratio of a growth rate to a loss rate — rather than their individual values. This means that the individual parameters are not identifiable from the available observations. Such correlated, under-determined solutions are prone to overfitting, because the compensating adjustments between parameters may match the specific noise structure of the training data but fail to generalise to independent conditions. Conversely, when posterior parameters are uncorrelated, each parameter is independently constrained by the data. This does not imply that each parameter is precisely determined — as shown by the HDI reductions of 16–41% (Table 5), substantial residual uncertainty remains for individual parameters. Rather, it means that the remaining uncertainty in one parameter does not depend on the value of another: each parameter carries its own, independent uncertainty range. This distinction is central to understanding the shift from correlated to uncorrelated equifinality reported in this study.

6. Manuscript structure and readability

We acknowledge that the Methods section was too long and difficult to navigate. The following structural changes have been made.

First, concise "roadmap" opening paragraphs have been added at the beginning of each key subsection (Sect. 2.1, 3.2, 3.3, and 3.6–3.9). Each opening paragraph states the purpose of the section and conveys the main idea in plain language before the technical detail begins.

Second, the most extensive formulae have been moved to a new Appendix A. This includes the RMSE and correlation coefficient definitions (former Eqs. 10–11), the full RCRV derivation (former Eqs. 13–16), and the detailed Sobol index equations. The main text now retains concise verbal descriptions of what each metric does and why it was chosen, with a reference to the Appendix for the mathematical formulation.

Third, Section 2.1 (BGC-Argo Data) has been restructured so that the essential information appears first: which float, where, when, which variables are measured and derived, and which independent datasets are used. The detailed float selection criteria and quality-control procedures follow as supporting material.

Fourth, the model description section has been renamed "Biogeochemical Model (PISCES)" to clarify that the section focuses on the biogeochemical model, not on the ocean circulation model that is not used in the parameter estimation.

Fifth, the technical detail of the 3D model–observation colocation procedure and the adaptation of the NRMSE metric for the 3D assessment (former Sect. 3.11–3.12) have been moved to Appendix A, with bridging sentences in the main text.

Sixth, Table 4, which lists the optimised values and HDI bounds for all 95 PISCES parameters, has been moved to the Supplementary Information. This table is provided as a concrete and reusable outcome for future studies, but its size (12 sub-tables) is not required to follow the main argument.

Together, these changes substantially reduce the length of the Methods and Results in the main text while preserving all technical detail for interested readers in the Appendix and Supplementary Information.

Specific Comments:

=> Abstract

Ln 12: Plead add: PISCES-model in a 1-dimensional framework

Done. The revised abstract now specifies "the PISCES biogeochemical model in a one-dimensional framework."

Ln 13: using seasonal observations of the year 2015

Done. The revised abstract now states "using seasonal observations from a single BGC-Argo float in the North Atlantic during the year 2015."

Ln 13: Better? Replace “metrics” by “observed biogeochemical tracers and related properties” or introduce what is meant by “metrics”.

We agree that "metrics" is ambiguous without context. In the revised abstract, we now write "twenty observational quantities derived from eight biogeochemical tracers" on first use. In the Methods (Sect. 3.1), we have added an explicit definition at the opening of the section: the term "metrics" refers to the set of twenty quantities — layer-mean concentrations and vertical-structure diagnostics — derived from the BGC-Argo observations and used as targets in the sensitivity analysis and optimisation.

Ln 14: Better? within a “1-dimensional representation of the ocean“ instead of “1D vertical configuration”

Agreed. Changed to "within a one-dimensional representation of the ocean."

Ln 18: add “all 95 poorly known model parameters”

Done. The revised abstract now reads "directly optimising all 95 poorly known model parameters."

Ln 19: reducing the NRSME relative to what?

The NRMSE reduction is relative to the reference simulation using the default PISCES parameter set.. The revised abstract now reads "reducing the normalised RMSE by over 50% relative to the reference simulation with default parameters."

Ln 19ff: These sentences need in my eyes more evidence. It’s especially confusing, since the authors stated just before that different fitting strategies lead to “statistically indistinguishable results” (cf. Tab.4 shows very different parameter sets).

We agree that this passage was confusing. The abstract has been substantially rewritten. The key clarification is the following: the three optimisation strategies achieve statistically indistinguishable skill improvements (Tab. 3), meaning that directly optimising all 95 parameters performs as well as targeting smaller, GSA-informed subsets. The different parameter values in Tab. 4 (now Table S1 in the Supplement) reflect uncorrelated equifinality — multiple parameter combinations produce equally good fits — but crucially, the parameters within each posterior ensemble are not structurally dependent on each other (median correlation < 0.04). See the response to Major Comment, subpoints 2 and 5, for a detailed discussion.

Ln 19: I don’t see why the one seasonal cycle considered here should be “richer” than using long-term observing stations (which has been done extensively since decades; e.g. Hunt et al. 1996; Schartau & Oschlies, 2003; Ward et al. 2013). I guess the authors refer to the number of observed prognostic tracers while some are "pseudo". Please clarify.

Addressed in the response to the Major Comment (subpoint 3). The word "rich" has been replaced by "comprehensive, multi-variable" in the abstract, and a clarifying passage has been added in the Introduction explaining the three dimensions of the dataset's advantage: multi-variable breadth (8 tracers, 20 metrics), temporal resolution (~5-day vs. monthly), and vertical resolution (1 m grid vs. discrete bottles).

Ln 22: Please “add of the seasonal cycle of the year 2015”.

Done. The revised abstract now specifies "of the seasonal cycle of the year 2015."

Ln 22: Please add what is the reference for parameter uncertainty which has been improved.

Done. The revised abstract now states "relative to the broad uniform prior distributions."

Ln 23: This statement contradicts somehow the conclusions drawn from the other data set: “the optimized ensembles demonstrate strong portability.”

Addressed in the response to the Major Comment (subpoint 1). All references to "portability" have been removed. The abstract now focuses on the 3D validation as primary evidence that the optimised parameters improve model performance, and the RCRV analysis of the two independent floats is presented as an ensemble calibration diagnostic rather than a portability test.

Ln 23: the “tightly constrained predictive spread” should refer to an independent test data. Otherwise reformulate “lead to similar fits to the observations”.

Agreed. The revised abstract now distinguishes clearly between (a) the goodness of fit on the training data (float #5904479, year 2015) and (b) the independent 3D validation (2017–2019). The phrase "tightly constrained predictive spread" has been replaced with language that accurately describes what each result demonstrates. The 3D validation sentence now reads: "the optimised parameter set from the All-parameters strategy improves the fully three-dimensional IBI model against approximately 1,430 independent BGC-Argo profiles and multi-product reprocessed surface chlorophyll-a over the period 2017–2019."

Ln 26ff: I am not sure where it has been shown that this fitting strategy is more "robust"

The term "robust" referred to the broader uncertainty quantification provided by the All-parameters strategy for unassimilated variables (Fig. 8), which is a consequence of allowing all parameters to vary rather than fixing a subset at their default values. We have clarified this in the revised abstract by replacing "more robust" with a more precise statement. The key reasoning is the following: when a parameter is fixed at its default value during optimisation, it is implicitly treated as perfectly known. In practice, however, default values are themselves poorly constrained — they originate from literature estimates or expert judgment and carry substantial uncertainty. Fixing such a parameter therefore artificially removes a genuine source of uncertainty from the predictive spread, making the model output appear more tightly constrained than it actually is. The narrower uncertainty envelopes produced by the Main Effects strategy (which fixes 29 of 95 parameters at their defaults) are thus not a sign of better knowledge but an underestimate of the true predictive uncertainty. The All-parameters strategy avoids this

by allowing all 95 parameters to vary, so the resulting predictive spread honestly reflects the full parametric uncertainty. The revised abstract now states: "the All-parameters strategy provides a more comprehensive quantification of predictive uncertainty for unassimilated variables, because it accounts for uncertainty in all parameters rather than artificially fixing poorly known parameters at their default values."

Ln 28: The amount of required model simulations would not be feasible in a full 3-dimensional ocean model – delete or reformulate “scalable”.

Agreed. The word "scalable" was misleading, as the method is feasible only through the computationally lightweight 1D column setup. The revised abstract now states: "The method is computationally tractable thanks to the use of a one-dimensional model configuration, requiring approximately 24 CPU-hours for the optimisation step."

Ln 29/30: This should be formulated as an outlook

Agreed. The final sentence of the abstract has been reformulated as an outlook: "The generality of the approach should be tested by applying it to additional BGC-Argo floats in other oceanic regions."

Ln 49: structural uncertainty is another problem which could be mentioned

Agreed. A sentence has been added in the Introduction acknowledging structural (model formulation) uncertainty alongside parametric uncertainty: "Beyond parametric uncertainty, structural uncertainty — arising from simplifications in the model's process formulations — is an additional source of error that is not addressed by parameter optimisation alone."

Ln 109/110: Please specify. It should be mentioned that the data refer to a single seasonal cycle and that the PISCES-model is initialized with these observations before starting the approx. 1-year simulation (or please correct me in case I misunderstood).

Done. The revised Introduction now states: "We assimilate observations from a single BGC-Argo float over one seasonal cycle (January 2015–January 2016) in the North Atlantic. The one-dimensional PISCES model is initialised using a hybrid approach (Table 1): variables directly observed or inferred from the float (NO_3^-, PO_4^{3-}, Si, O_2, TA, DIC) are taken from the BGC-Argo profiles; bulk Chl-a and POC observations are disaggregated into the model's plankton functional types using component ratios from the CMEMS-BGC analysis; and remaining unobserved variables (iron species, DOC, mesozooplankton, etc.) are prescribed from CMEMS-BGC. The start date is chosen to minimise the RMSE between float observations and CMEMS-BGC, ensuring the most consistent initial state across all variables."

Ln 131ff: The main idea of iIS is not conveyed.

Agreed. A plain-language summary of iIS has been added in the Introduction: "iIS is a Bayesian method that iteratively refines an ensemble of model simulations by reweighting parameter sets according to their agreement with the observations. At each iteration, the ensemble is concentrated towards the regions of parameter space that best reproduce the data, yielding full posterior distributions of parameters, including uncertainties and correlations." A similar roadmap opening has also been added to the Methods section (Sect. 3.2), as described in the response to the Major Comment (subpoint 6).

Ln 130ff BGC-Argo Data. These subsections could be much more concise. It should be stated right in the beginning which time period and location have been considered and how the model is initialized and which independent data have been used for testing. Then indicate which biogeochemical tracers were considered and list the data sources. All other details could be moved to an Appendix (such as the float numbers or properties of other floats and reasons why these were not considered).

Addressed in the response to the Major Comment (subpoint 6). Section 2.1 has been restructured so that the first paragraph provides the essential information: which float (#5904479), where (North Atlantic), when (January 2015–January 2016), which variables are measured (Chl-a, bbp/POC, NO_3^-, O_2) and derived (PO_4^{3-}, Si, DIC, TA via CANYON-B/CONTENT), and which independent datasets are used (two additional floats for ensemble calibration; ~1,430 BGC-Argo profiles and MOBTAC surface Chl-a for 3D validation over 2017–2019). Detailed float selection criteria and quality-control procedures follow as supporting material.

Ln 200ff: I find it very confusing to start with the 3-dimensional ocean model which is not used in the parameter estimation experiments. Otherwise, this subsection describes only PISCES. Rename to “Biogeochemical Model description”?

Addressed in the response to the Major Comment (subpoint 6). The section has been renamed "Biogeochemical Model (PISCES)" and reordered so that the PISCES model description comes first, followed by the 1D configuration, then the 3D configuration.

Ln 230: where does the vertical turbulent diffusion come from?

The vertical eddy diffusivity profile (Kz) used in the 1D configuration is extracted from the Copernicus Marine Service global ocean physics analysis (CMEMS-PHY,

https://doi.org/10.48670/moi-00016). The original manuscript stated this (Sect. 2.4.1: "forced by [...] daily profiles of temperature, salinity, and vertical diffusivity extracted from [...] CMEMS-PHY"), but the formulation was insufficiently clear about the distinct roles of the two sets of forcing dataIn the revised text, we now distinguish explicitly: the 1D configuration does not solve any ocean physics — no turbulence closure scheme is computed. Instead, the physical environment is entirely prescribed by daily vertical profiles of temperature (T), salinity (S), and vertical eddy diffusivity (Kz) extracted from the CMEMS-PHY analysis along the float trajectory. These profiles impose the thermal, haline, and mixing conditions on the biogeochemical model. Separately, surface atmospheric fluxes, also extracted from CMEMS-PHY along the float trajectory, drive the biogeochemical surface processes in the 1D model.

Benefits and challenges of the 1-dimensional setup do not become clear and I did not understand how ocean mixing was derived. Some important information is scattered; e.g. that some metrics were not usable because the ocean model is too simple. Ultimately, all biogeochemical prognostic variables depend heavily on ocean mixing such that this point needs more careful discussion (this might impose huge uncertainties and has the potential to lead to the situation described by Hermans et al. 2022 and Yang & Zhu, 2018 with too tight posteriors and over-confident calibration).

The revised 1D configuration section (Sect. 2.4.1) now opens with a clear statement of benefits and limitations. The benefit is computational: the 1D setup enables thousands of ensemble simulations required by iIS at a fraction of the cost of a 3D model. The main limitation is the absence of horizontal advection, which means that laterally driven biogeochemical variability is not represented. To account for this, a representativity error is included in the cost function (Sect. 3.7), quantifying the mismatch between the 1D and 3D representations. Ocean mixing (vertical eddy diffusivity profile, Kz) is prescribed from the CMEMS-PHY analysis, as clarified above. Regarding the potential for this simplified setup to produce over-confident calibration (Hermans et al. 2022; Yang and Zhu 2018), we refer to the response to the Major Comment (subpoint 2), where we show that the RCRV dispersion demonstrates the opposite: the ensemble is conservative, not over-confident.

Ln 217ff: I rather expected now a description on how the ocean is represented in 1D. This part rather refers to Section 3.1.

Agreed. In the revised manuscript, the 1D configuration section (Sect. 2.4.1) now focuses on how the ocean is represented: no ocean physics is computed; instead, the physical environment is entirely prescribed by daily vertical profiles of temperature, salinity, and vertical eddy diffusivity (Kz) extracted from the CMEMS-PHY analysis along the float trajectory, while surface atmospheric fluxes, also from CMEMS-PHY, drive the biogeochemical surface processes. Horizontal and vertical advection are neglected. The

section also describes the float selection criteria that ensure the 1D assumption is valid. The material about metric definitions has been moved to Section 3.1 where it belongs.

Ln 260ff: This subsection is a bit confusing since no 3d model has been involved in the fitting process. It might be deleted and the well-known referenced setup could be mentioned when it comes to error estimates.

In the original manuscript, the section described a single 3D configuration: the global NEMO-PISCES simulation (3D-Free), used solely to compute the representativity error for the 1D optimisation. In the revised manuscript, a second, distinct 3D configuration has been added: the high-resolution IBI regional model, which serves as the platform for the independent 3D validation (Sect. 4.6). These two configurations are now presented in two separate subsections: Sect. 2.4.2 describes the global 3D-Free simulation, which provides the representativity error, and Sect. 2.4.3 describes the regional IBI model, in which the optimised parameters are implemented and validated against independent observations. Each subsection explicitly states that the corresponding 3D model was not involved in the parameter fitting process.

=> Metrics for sensitivity analysis and parameter optimization

Again, the section should clearly and briefly explain in the beginning what has been done and come to all the details later.

Agreed. As part of the structural improvements described in the response to the Major Comment (subpoint 6), a roadmap opening paragraph has been added at the beginning of Section 3.1. This paragraph states upfront the purpose of the section (defining the observational targets used in the sensitivity analysis and optimisation), introduces the distinction between "metrics" (the 20 individual observational quantities) and the "objective function" (the NRMSE that aggregates them), and provides a brief overview before the detailed definitions follow.

Ln 269, 271ff: I find the use of “metrics” sometimes confusing and it should be clearly distinguished between “biogeochemical tracers and related properties” and objective functions (or whatever the authors want to call it). I am aware that many different name giving conventions exist and it might thus be helpful to clearly introduce what is meant in the beginning. I guess here the “20 metrics” were combined to one objective function (NRMSE)? Please clarify.

The reviewer is correct that the terminology needs clarification. The 20 metrics (layer-mean concentrations and vertical-structure diagnostics derived from 8 biogeochemical tracers) are the individual observational targets. Within the iIS framework, these metrics enter the observation likelihood function p(y|x), which quantifies the probability of the observations given the model output for each ensemble member. This likelihood, computed across all 20 metrics and their associated observation errors, determines the weight assigned to each ensemble member (Sect. 3.2). The NRMSE is not the objective function driving the optimisation; it is a post-hoc diagnostic used to evaluate the resulting model skill. A clarifying paragraph has been added at the opening of Section 3.1: "We define twenty observational metrics — layer-mean concentrations and vertical-structure diagnostics — derived from eight biogeochemical tracers measured or inferred from the BGC-Argo float. These metrics serve as individual targets for the sensitivity analysis and optimisation. Within the iIS framework, they enter the observation likelihood function p(y|x), which determines the weight of each ensemble member based on how well it reproduces the observations across all 20 metrics simultaneously, accounting for the observation errors associated with each metric (Sect. 3.5)."

Ln 313: It does not get clear which prior estimate for the parameter uncertainty has been chosen. What is the considered range and where does it come from?

The prior distributions are uniform, with broad perturbation intervals ranging from one-hundredth to twice the reference values. These ranges, as well as additional constraints on certain linked parameters to prevent unrealistic phenomena (e.g., artificial generation or loss of matter), were defined in close consultation with Olivier Aumont, the creator of the PISCES model and co-author of this study. The full list of all 95 parameters with their reference values, perturbation ranges, and inter-parameter constraints is provided in Table S1 in the Supplement. This information is described in Section 3.4, but was not introduced early enough. The revised Section 3.1 now includes a forward reference: "The prior parameter distributions are uniform, with bounds and inter-parameter constraints defined in consultation with the PISCES model developer (Sect. 3.4; Table S1)."

Ln 330ff: This section is really hard to follow for people not familiar with iIS. It would be important to convey the main idea while most other parts could be moved to the supplement.

Addressed in the response to the Major Comment (subpoint 6). A plain-language roadmap paragraph has been added at the opening of Section 3.2, and the most

technical formulae have been moved to Appendix A.

Ln 432ff Again this subsection is far too extensive and a bit unexpected since Sobol indices were already mentioned before. Again, it is key to clearly convey the main ideas.

Addressed in the response to the Major Comment (subpoint 6). A roadmap opening paragraph has been added, and the detailed Sobol equations have been moved to Appendix A.

Ln 634: I guess I overlooked something. What was the baseline?

The baseline (reference) is the simulation using the default, unoptimised PISCES parameter set. This is now stated explicitly wherever improvements are reported: "relative to the reference simulation with default PISCES parameters (hereafter REF)." The term REF is introduced formally at first use and used consistently throughout.

Ln 638ff: Could you briefly explain (here or elsewhere) how the correlations among parameters relate to overfitting and parameter identifiability to maintain a broader readership?

Addressed in the response to the Major Comment (subpoint 5). A new paragraph has been added in the Discussion explaining the chain: correlated posteriors → identifiability problem → overfitting risk, and the distinction with uncorrelated posteriors that retain individual uncertainty.

Ln 641: I would rather have the main ideas explained. The extensive formula could go to an Appendix. Most readers will we familiar with an RMSE and the concept of normalization. The thoughts behind using RCRV do not become clear and an easier metric that compares well to the fitting process might be considered.

Addressed in the response to the Major Comment (subpoint 6). The NRMSE and RCRV formulae have been moved to Appendix. The main text now explains the rationale concisely: the NRMSE quantifies the overall model-data misfit normalised by observation uncertainty (values below 1.0 indicate that errors fall within observational accuracy). The RCRV complements this by diagnosing ensemble calibration — specifically, whether the posterior predictive spread is consistent with the observed errors. RCRV bias measures systematic over- or under-prediction; RCRV dispersion measures whether the ensemble spread is too narrow (over-confident, >1) or too wide (conservative, <1). This diagnostic is essential because a good NRMSE score does not guarantee that the ensemble uncertainty is correctly calibrated.

Ln 780: I still don’t know which reference the improvements refer to and as such all the information has no value. Please clarify.

The reference is the simulation using the default PISCES parameter set (REF). This has been clarified throughout the Results section. Every statement of improvement now explicitly states "relative to REF" or "relative to the reference simulation with default parameters." The term REF is introduced at first use in the Results and used consistently thereafter.

Ln 898: Please check the wording. This paragraph cannot refer to the predictive skill if no independent data were considered (i.e. not used during fitting). I would say, it rather refers to the goodness of fit to the observational data used during the fitting procedure which can be obtained by the optimization.

The reviewer is correct. The term "predictive skill" was misleading for results on the training data. The revised text now distinguishes clearly between "goodness of fit" (performance on the training data, float #5904479) and "independent validation" (performance on data not used during fitting: the 3D validation against ~1,430 BGC-Argo profiles over 2017–2019, and the RCRV analysis on the two independent floats). The passage at Ln 898 now reads "goodness of fit to the assimilated observations."

Ln 916: Currently, I find the analysis of the test data not convincing to rule out overfitting and proofe what the authors term “portability” (this might become more clear when using a comparable metric as used during the fitting process for all optimal solutions and also some more discussion is needed why the Mediterranean float shows conflicting results). It does not become clear to me how much the "pseudo observations" contribute.

Addressed in the response to the Major Comment (subpoints 1, 2, and 4). The "portability" framing has been removed entirely. The primary evidence against overfitting is now the new 3D validation (Sect. 4.6), which demonstrates that OPTI outperforms REF across the majority of variable–basin combinations over 2017–2019. The RCRV analysis on the two independent floats has been repositioned as an ensemble calibration diagnostic (Sect. 4.7). Regarding pseudo-observations: the 3D validation uses only O_2 (directly measured) and CANYON-B/CONTENT-derived variables, each normalised by its own point-by-point uncertainty field (Sect. 3.12). The improvement of OPTI over REF is observed for both directly measured (O_2) and derived variables, indicating that the pseudo-observations are not driving artificial improvements.

Pros and Cons of the presented approach could be discussed more clearly. E.g., how does the chosen 1D setup with focus on seasonal processes and perfect initial conditions relate to 3D coupled biogeochemical ocean models used for projections and run for decades or more (where e.g. the global distribution of nutrients is key)? Also, it would be interesting to discuss the mentioned shortcomings in the 1D representation of the ocean and the planed implementation into a 3-dimensional model (Ln 1041ff) (in a 3-dimensional context it has been shown that biogeochemical model parameters can be tuned to compensate for ocean model differences – while this is not without problems for projections (Löptien & Dietze, 2019; Pasquier et al. 2023). From these studies one might conclude that differing parameters will be optimal in the 1D and 3D setups (unless the mixing matches fairly well) and I would be very interested on some thoughts on this.

This is an important point. A new paragraph has been added in the Discussion addressing the relationship between 1D-optimised parameters and their use in 3D models. The paragraph makes the following argument.

The concern raised by Löptien and Dietze (2019) and Pasquier et al. (2023) — that optimised biogeochemical parameters may compensate for biases in the ocean circulation model rather than correcting genuine biogeochemical errors — applies in principle to any parameter calibration framework. In our case, however, the 1D optimisation and the 3D validation differ substantially in their physical representation: the 1D model runs offline with prescribed T, S, and Kz profiles from the CMEMS-PHY analysis, while the IBI regional model runs fully online, computing its own ocean dynamics with CMEMS-PHY providing only boundary conditions. Furthermore, the initial conditions differ: the 1D model uses a hybrid approach combining float observations and CMEMS-BGC, whereas the 3D IBI model uses its own independent initialisation. The fact that parameters optimised in such a simplified, offline 1D configuration also improve a fully three-dimensional model with its own dynamical circulation and different initial conditions strengthens the conclusion that the parameter adjustments are primarily correcting biogeochemical process representations rather than compensating for circulation biases.. The extent to which 1D-optimised parameters remain beneficial in such long-term integrations is an open question that we identify as a priority for future investigation.

Ln 995ff: Not agreed (yet?).

This line refers to the claim about decorrelated posteriors indicating that the long-standing challenge of parameter equifinality has been addressed. We have toned down this claim in the revised Discussion, as described in the response to the Major Comment (subpoint 5). The revised text now states that equifinality has not been eliminated but has changed in character — from correlated to uncorrelated equifinality — and that each parameter retains its own independent uncertainty (HDI reductions of 16–41%). The 3D validation provides supporting evidence that these decorrelated parameter sets produce genuine improvements, not artefacts.

Ln 1001ff: I was not aware by the model description that several metrics were optimized simultaneously and assumed that all so called "metrics" were merged into a single NRMSE. To use more than one objective function is certainly possible, but not trivial (cf. Sauerland et al. 2019). Please clarify.

We clarify two points. First, regarding the optimisation mechanism: iIS does not use the NRMSE as its objective function, nor does it use multiple objective functions. Instead, all 20 metrics enter a single observation likelihood function p(y|x), which quantifies the probability of the observations given the model output. Concretely, for each ensemble member j, the time series from all metrics are concatenated into a single observation vector, and the likelihood is the product of individual terms over all metrics and all time steps. For the 17 metrics with Gaussian observation errors, each term is a Gaussian PDF evaluating the model-data misfit normalised by the total observation error. For the three depth-based metrics (H_{DCM}, $H_{nitracline}$, H_{O2min}), a uniform distribution is used instead, assigning zero weight to members outside the observed error bounds. The resulting scalar likelihood value determines the weight of each member in the iIS procedure (Eq. 4). The full explicit formula is now provided in the revised Appendix (Eqs. A1–A3; see also the response to Reviewer 2's comment on Lines 345/360). This is therefore neither a multi-objective optimisation in the sense of Sauerland et al. (2019), where multiple conflicting objectives are optimised with Pareto-front methods, nor a simple NRMSE minimisation. It is a Bayesian approach where the observations from all 20 metrics jointly constrain the posterior parameter distributions through a single product likelihood.

Second, regarding the reported NRMSE values: the NRMSE is computed individually for each metric M (Eq. 13, 16). The values reported in Table 3 and in the text (e.g., -55.6%) are the median of these per-metric NRMSE reductions, with the IQR/2 indicating the spread across metrics. This was stated in the Table 3 caption and in the Results text, but was missing from the Abstract, which simply read "reducing NRMSE by 54–56%" — making it sound as if all metrics had been merged into a single NRMSE. The revised Abstract now reads "reducing the median NRMSE across metrics by 54–56% relative to the reference simulation." This has been clarified in the revised Section 3.1 (see also the response to the comment on Ln 269/271).

Ln 1005ff: As outlined in my major comment, some conclusions need more evidence – especially given the fact that very different parameter sets (Abstract, Tab.4) lead to optimal statistically indistinguishable model results.

Addressed in the response to the Major Comment (subpoints 2 and 5). The conclusions about decorrelated posteriors and identifiability have been toned down and are now supported by the new paragraph (correlation → identifiability → overfitting), the RCRV dispersion evidence (conservative, not over-confident), and the 3D validation (improvement against independent data over 2017–2019). The distinction between equifinality (multiple equally good solutions) and overfitting (fitting noise) is now made explicitly.

=> Supplement:

Please add some introductory text what is shown.

Done. A brief introductory text has been added to the Supplement explaining its contents and the relationship of the supplementary figures and tables to the main text.

References

We thank the reviewer for these references. We have incorporated citations to Hermans et al. (2022) and Yang and Zhu (2018) in the Discussion when discussing over-confident calibration risk. We have cited Löptien and Dietze (2017, 2019) and Pasquier et al. (2023) in the new Discussion paragraph on the relationship between 1D and 3D parameter optimisation. Sauerland et al. (2019) is referenced to clarify that our approach is single-objective, not multi-objective. Hurtt and Armstrong (1996) is now cited alongside Schartau and Oschlies (2003) and Ward et al. (2013) in the discussion of long-term observing station studies.