We thank Reviewer 2 for the careful and insightful evaluation of our manuscript. The review highlights the need for greater transparency in how the 20 observational metrics enter the iIS likelihood function, raises important questions about the limited improvement in the deep water column and the convergence of iIS in a high-dimensional parameter space, and suggests promising extensions including multi-location assimilation and the incorporation of zooplankton observations.

In response, the most substantial methodological addition is the explicit likelihood formula, now provided in the revised Appendix (Eqs. A1–A3), which shows how the individual metric terms combine into the single scalar likelihood that determines each ensemble member's weight. This bridges the general form of Eq. (12) to the weight computation in Eq. (4) and clarifies that the approach is neither a sum of cost terms nor a multi-stage setup. The revised manuscript also includes a new three-dimensional validation (Sect. 5) 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, providing stronger transferability evidence — including improvements in the Mediterranean sub-region. The deep water column discussion has been clarified, and multi-location assimilation and the use of zooplankton observations (including UVP6-derived data; Picheral et al., 2022) are now identified as future priorities. Below, we address each comment in turn. Reviewer comments are in normal text; our replies are in blue.

The approach to leverage the upcoming enrichment of BGC data by BGC-Argo sounds promising, since a thorough calibration of coupled marine BGC models was so far limited due to a lack of comprehensive observational data for all state variables addressed by BGC (cf. Kriest et al., 2023; Petric et al., 2022; Rohr et al., 2023).

We thank the reviewer for this positive assessment and for recognising the potential of BGC-Argo data for model calibration.

According to the results, the applied iIS works quite well for the productive layer of the 1D setup of PISCES, where an ensemble misfit reduction of 50% is reached and model state trajectories are closer to the observational time-series from BGC-Argo.

In the deeper water column, however, both the misfit and most tracer trajectories (except POC) stay close to the reference run, a fact that is attributed to impacts of the ocean circulation that dominate and act on longer time scales than the assimilated time-series in the 1D model setup.

This is an accurate summary. The limited NRMSE improvement in the deeper water column (Table 3b) requires a nuanced interpretation. Over the timescale of a single seasonal cycle,

most mesopelagic tracer concentrations (NO_3^-, PO_4^{3-}, Si, O_2, DIC, TA) exhibit relatively little variability: the processes that drive change at those depths operate on timescales much longer than one year. Consequently, both the reference simulation and the observations show similar, nearly steady values at depth, leaving little room for the optimisation to demonstrate measurable NRMSE improvement. This interpretation is consistent with the fact that the reference simulation already reproduces the mesopelagic observations reasonably well — confirming that, at seasonal timescales, the 1D framework captures the dominant dynamics at depth. POC is the exception, because the sinking particle flux provides a direct vertical link between productive-layer dynamics and mesopelagic concentrations on seasonal timescales.

However, this does not mean that the mesopelagic metrics are uninformative. Within the likelihood framework, they serve as a filter: ensemble members that produce unrealistic mesopelagic dynamics — for instance, excessive nutrient depletion or oxygen drawdown at depth — are penalised and receive low weights. The mesopelagic metrics therefore contribute to discarding implausible parameter vectors, even though their NRMSE does not improve substantially. We have clarified this dual role (improvement vs. filtering) in the revised Discussion.

Another emphasis of the paper is the use of a preliminary global sensitivity analysis using Sobol' indices to define three different subsets of the parameters used for optimisation: (1) all parameters, (2) parameters filtered by first order Sobol' indices (parameters that are most sensitive alone), (3) parameters filtered by all-order Sobol' indices (parameters that are most sensitive in combination with other parameters).

Here, the corresponding iIS results are found to be statistically similar for all considered subsets and --- as calculating the Sobol' indices is computationally more expensive than the application of iIS --- optimising all parameters at once is stressed to be the best strategy.

This is quite amazing since I would expect the incorporation of many different parameters to impede convergence of optimisation.

We agree this is a striking result. The key explanation lies in the nature of the iIS algorithm: it does not require convergence in the traditional optimisation sense. Instead, it reweights an ensemble of simulations drawn from the prior. Each ensemble member is a complete model run with a different parameter vector, and the observation likelihood p(y|x) — computed across all 20 metrics — assigns a weight to each member. Members whose output is more consistent with the observations receive higher weights. Because the algorithm does not iterate toward a single optimum but rather reshapes the full prior

distribution into a posterior, the dimensionality of the parameter space affects the required ensemble size (to adequately sample the space) but does not impede convergence per se. The broad perturbation ranges (1/100 to 2× reference values) and the large ensemble size (8,192 for the first iteration, then 2,048 members per iteration) help explore the high-dimensional space, although full coverage of a 95-dimensional parameter space remains inherently challenging and the iterative resampling is designed to progressively focus on the most plausible regions. We have added a brief clarification of this point in the revised Discussion.

Related to my amazement, as a non-statistician I would like to see a little bit more about the details how the measures (mean concentrations of DIC, TA, O_2, NO_3^-, PO_4^{3-}, Si, and POC) enter the quantities of interest (QoI) and the iIS algorithm.
Are these measures actually replacing the large vectors y and s (observed and simulated tracer states) in the iIS?
My first and second guesses at first glance where that you consider a sum of cost terms over all tracers, and that you probably apply iIS in a multi-stage setup w.r.t. disjoint subsets of the parameters.
Therefore, I would like to see more details about it in an Appendix section.

We appreciate this request and understand the confusion. The iIS framework is already described in Equations (2)–(5): Bayes' theorem (Eq. 2), the prior as an ensemble of equally-weighted particles (Eq. 3), the weight of each member as the normalised likelihood $\omega_j = p(y|x_j) / \Sigma\, p(y|x_j)$ (Eq. 4), and the posterior as a weighted combination of particles (Eq. 5). Equation (12) then gives the general form of the likelihood: $p(y|x) = p_\varepsilon(y - H(x))$.

What is missing is the explicit bridge between Eq. (12) and Eq. (4): how $p(y|x_j)$ is actually computed from the 20 metrics. To clarify: the 20 metrics do indeed replace the observation vector y and the corresponding simulated vector s. The optimisation is neither a sum of cost terms nor a multi-stage setup with disjoint parameter subsets. For each ensemble member j, all time series from all metrics are concatenated into a single observation vector. The likelihood is then the product of individual terms over all metrics and all time steps. The full formula, now provided in the revised Appendix (Eqs. A1–A3), can be summarised as follows.

For the 17 metrics with Gaussian observation errors, the likelihood raised to the power α is:

$$p(\mathbf{y}|\mathbf{x}_j)^\alpha = \prod_{M \in \mathcal{M}_{\text{Gauss}}} \prod_{t \in \mathcal{T}_M} \left[\frac{1}{\sqrt{2\pi}\,\varepsilon_t^M} \exp\left(-\frac{1}{2} \left(\frac{y_t^M - s_{j,t}^M}{\varepsilon_t^M} \right)^2 \right) \right]^\alpha$$

where y_t^M is the observed value of metric M at time t, $s_{j,t}^M$ is the corresponding simulated value for ensemble member j, and ε_t^M is the total observation error (quadratic sum of measurement and representativity errors). The set $\mathcal{M}_{\text{Gauss}}$ contains the 17 metrics for which the observation error is assumed Gaussian, and $\mathcal{T}_M$ is the set of time steps for metric M. The exponent α is the inflation parameter described in Section 3.2, which is dynamically adjusted at each iteration to maintain a target effective sample size (Eq. 6).

For the three depth-based metrics (H_{DCM}, $H_{\text{nitracline}}$, $H_{\text{O}_2\text{min}}$), the observation error follows a uniform distribution. Their contribution to the likelihood is:

$$p_{\text{uniform}}^M(t) = \begin{cases} \dfrac{1}{\varepsilon_{\text{Sup},t}^M + \varepsilon_{\text{Inf},t}^M} & \text{if } \; y_t^M - \varepsilon_{\text{Inf},t}^M \le s_{j,t}^M \le y_t^M + \varepsilon_{\text{Sup},t}^M \\ 0 & \text{otherwise} \end{cases}$$

where $\varepsilon_{\text{Inf},t}^M$ and $\varepsilon_{\text{Sup},t}^M$ are the lower and upper error bounds, respectively. Members whose simulated depth falls outside the observed error interval receive zero weight and are effectively discarded.

The full likelihood combining both types of metrics is therefore:

$$p(\mathbf{y}|\mathbf{x}_j)^\alpha = \prod_{M \in \mathcal{M}_{\text{Gauss}}} \prod_{t \in \mathcal{T}_M} \left[\mathcal{N}\left(s_{j,t}^M \mid y_t^M, \varepsilon_t^M\right)\right]^\alpha \times \prod_{M \in \mathcal{M}_{\text{Uniform}}} \prod_{t \in \mathcal{T}_M} p_{\text{uniform}}^M(t)$$

This scalar value is then used in Equation (4) to compute the normalised weight ω_j of each ensemble member.

There is no multi-stage setup: all parameters are perturbed simultaneously in each ensemble member, and all metrics constrain all parameters jointly through this single product likelihood. The resulting scalar value is what appears in Eq. (4) to determine each member's weight.

Despite having good statistical properties, the posterior parameter distribution does not represent well a validation BGC-Argo data set taken in a different ocean basin. The authors conclude that parameter distributions should be location-dependent. It would also be interesting to see whether good data assimilation of multiple locations simultaneously is possible.

The reviewer accurately identifies a key result and its implication. The Mediterranean float exhibits two large RCRV biases, but they have distinct origins. The nitracline depth bias is physically meaningful: the optimisation, constrained by a North Atlantic float with a consistently shallow nitracline, yields biological rate parameters that incorrectly maintain this

shallow structure in the oligotrophic Mediterranean, where the nitracline is much deeper. This reflects fundamentally different biogeochemical regimes and supports the conclusion that parameter distributions should be location-dependent. The productive-layer nitrate bias, by contrast, is largely a normalisation artefact: in the oligotrophic Mediterranean, observed concentrations and their associated errors are extremely small, causing the RCRV normalisation to amplify a physically negligible model–data mismatch into a large score — the optimised model does not actually perform poorly for this variable. To better reflect these nuances, the revised manuscript renames the former "Portability of the Optimized Ensembles to Other Bioregions" section to "Ensemble Performance Against Two Independent BGC-Argo Floats" (Sect. 4.6), toning down the portability claim.

Regarding multi-location assimilation: we agree this is a natural and promising extension. The iIS framework can in principle accommodate multiple floats by extending the observation vector y to include metrics from several locations simultaneously; the likelihood would then jointly constrain the parameters across all sites. The main challenge is computational, as each ensemble member would need to be run at multiple locations. We have added a paragraph in the revised Conclusions identifying this as a priority for future work.

Finally, to provide stronger evidence for the transferability of the optimised parameters, the revised manuscript now includes a new 3D validation (Sect. 5) in which the optimised parameter set is implemented in the high-resolution IBI regional model and validated against ~1,430 independent BGC-Argo profiles and a surface chlorophyll-a product (MOBTAC) over 2017–2019. Notably, the 3D validation shows that the optimised parameters improve the simulation not only in the North Atlantic but also in the Mediterranean sub-region of the IBI domain, suggesting that the single-location optimisation captures improvements in the biogeochemical parameterisation that extend beyond the training basin

Since the PISCES model has an elaborate representation of zooplankton processes, it would be worthwhile to consider derived zooplankton observations (cf. Petrik et al., 2023; Rohr et al., 2022) in conjunction with BGC-Argo data and other data sources.

We agree that incorporating zooplankton observations would strengthen the constraint on the model, particularly given that our GSA identifies zooplankton-related parameters as the dominant source of sensitivity. Derived zooplankton products such as those proposed by Petrik et al. (2022) and Rohr et al. (2023) could complement the BGC-Argo data by constraining the model's trophic transfer and grazing dynamics more directly. Looking further ahead, the recent development of the Underwater Vision Profiler 6 (UVP6; Picheral et al., 2022), a miniaturised imaging sensor that can be integrated on BGC-Argo floats, may in the future provide in-situ particle size spectra and zooplankton abundance profiles from

the same autonomous platforms. Although such data are not yet routinely available, their eventual integration into frameworks like iIS could help constrain the zooplankton grazing parameters that our GSA identifies as the dominant source of model sensitivity. We have added a sentence in the revised Conclusions to this effect.

SPECIAL COMMENTS:

Lines 345, 360:

Could you provide an explicit formula for p(y|x) as used in equations (2) and (4) in this study, perhaps in the Appendix? This will make the methodology clearer to readers, together with the code availability section.

Done. The manuscript already provides the general form of the likelihood in Equation (12): $p(y|x) = p_\varepsilon(y - H(x))$, and shows in Equation (4) that the weight of each member is the normalised likelihood. What was missing is the explicit expansion showing how $p(y|x_j)$ is computed from the 20 metrics. The revised Appendix (Eqs. A1–A3) now provides the full formula: a product of Gaussian PDF terms over all 17 Gaussian metrics and their time steps (Eq. A1), a piecewise uniform distribution for the three depth-based metrics (Eq. A2), and the combined likelihood (Eq. A3). This bridges Eq. (12) to the weight computation in Eq. (4), making the full likelihood transparent. The code implementing this computation is available in the associated repository referenced in the Code Availability section.

Line 356:

The phrase "which has as the main objective restraint the sampling to regions of the state space of high probability" does not sound natural to me. Do you mean "The main objective is to restrict sampling to regions of the state space with high probability"?

We agree, and we thank the reviewer for the suggested rewording. The revised text now reads: "whose main objective is to restrict sampling to regions of the state space with high probability."

Lines 643 ff: The variables y^M_t and ε^M_t already appear in equation (13) but are only explained after equation (16). In contrast, p_j and $s^M_{j,t}$ are explained both after equation (16) and before their first use. I think redundancy is beneficial here, as readers need to

understand the meaning of symbols immediately after each formula. Therefore, y^{M}_{t} and ε^{M}_{t} should also be defined after equation (13).

We agree. The revised manuscript now defines all variables (y_t^M, $s_{Best,t}^M$, ε_t^M, and T) immediately after Equation (13), where they first appear. The definitions are retained after Equation (16) as well for completeness.