Answers to the reviewer's comments

We thank the Reviewer for his/her careful reading of our paper, and for his/her remarks that will help improve the clarity of the manuscript. We did our best to take them into account as explained below. Our replies and comments are given in normal type, while the original comments from the Reviewer are in bold/italics.

This study proposes a methodology for combining numerical model output with observations. The framework is that of Bayesian analysis. In contrast to most data assimilation, however, the emphasis is on using offline pre-generated numerical model ensembles as the prior information, rather than embedding a numerical model in the analysis scheme. The application is for multiple biogeochemical variables (while observing only one). I played around with a similar approach years ago (using an enKF, but with explicit time evolution), but never did a real application nor proper assessment (we decided to take another approach for our analysis of ocean carbonate variable and so it ended there). Hence, I like the approach, and am pleased that someone has taken it forward to the community, as I always thought it would be useful. My opinion is that there is a real need for these types of space-time analysis procedures that don't require explicit running of numerical models within the estimation procedure (but still use numerical model information). I note that lots of statistics people are doing problems like this with sophisticated spatio-temporal approaches (e.g. integrated nested Laplace approximations, INLA), but these tend not to be very accessible the ocean community. The approach taken here is straightforward, builds on basic data assimilation principles, gets decent results, and should be accessible to most ocean data analysts. Hence I recommend publication with some minor revisions.

COMMENTS

There is a strong link with the foundational approach of optimal interpolation (which parallels the Kalman filter observation step). Since most people know about OI, it might be useful to make a quick note of it in order to make the approach more clear to the non-expert reader.

Yes, we agree that a link with OI was missing in the paper. This has been added in the introduction. Sea also our answer to reviewer 1 about this.

My main confusion in understanding the methodological development was how time was incorporated into the analysis. After a couple re-reads, I see this is made clear early on when you define the state vector (its dimension includes time). But this could be brought out more explicitly. When most people see that you have used the Kalman filter update machinery, they will wonder about evolution through time. Part of my confusion may also have arisen since when I did my version of this problem, I actually ran it sequentially in time with a daily time step, and did the (spatial) observation update whenever measurements were available. My time correlation model was an auto-regressive one, and I used an enKF/enKS methdology. Your time correlation is implicitly embedded in the space-time covariance matrix that defines the multivariate state.

Thank you. It is indeed very important that this point can be understood clearly. We have added the following sentences in the paper to clarify this.

In the introduction, while making the link with ensemble OI : « In practice, this can for instance be achieved by a direct application of just one analysis step of the ensemble optimal interpolation (EnOI) algorithm (e.g. Evensen, 2003 ; Oke et al., 2010), for a 4D estimation vector (thus embedding the time evolution of the system, as in Mattern and Edwards, 2023), but over an extended time window. »

In section 3.1, defining the 4D estimation vector : « As compared to classic sequential data assimilation (like the ensemble Kalman filter), the difference here is that the whole 3-month time sequence is packed together in the 4D estimation vector. This is makes the problem bigger, but also allows us to concentrate on a small subregion and a few selected variables. »

Lines 140-150. I found this discussion of uncertainties 2 and 3 confusing. I get that you are trying to account for unresolved scales. There is likely a better plain language way of saying what you are doing.

Yes, we have tried to clarify this by adding a short simple explanation on what we are doing, in one sentence at the beginning of each of these paragraphs, before going to the technical explanation.

Uncertainty 2 : « In the biogeochemical model, another important source of uncertainty is the effect of the unresolved scales on the large-scale component of the biogeochemical tracers. As a result of the nonlienar formulation of the model equations,... »

Uncertainty 3 : « Unresolved scales in the physical component of the model also produces large-scale effects that are difficult to parameterize, and thus produce uncertainties that are not easy to simulate. »

Do you think it is proper to equate a 4D inverse problem with a Bayesian estimation? I know there are links, but you have to hand-wave a lot to explain them. Why not just say you used a Bayesian method?

We just say that we use the Bayes theorem to solve the problem. We have a prior probability distribution describing the vector of variables to be estimated, and then conditions that we apply on this prior distribution (in our case, observations). Then, we make approximations on the shape of these distributions to solve the problem (which are different in the two methods that we have applied). How close we are to an exact Bayesian estimation will depend on the validity of these assumptions. For instance, there are much less assumptions in the MCMC sampler, since the observation condition can be applied using the native probability distribution for the observation uncertainties (i.e. without approximations on the likelihood function).

Nice job on highlighting the difficulties of using small ensembles, partial observation of the state, and the need to estimate a big multivariate state. The tricks to make this work (like localization) are appreciated by the reader. Similarly, nice job on trying out some "ecological indicators" which emphasize the multivariate state and how measurement on one variable can tell you about other variables (and project to depth). However, the ecological indicators are to me not so central, and if shortening the paper was required I would omit these.

Thank you. Concerning the estimation of indicators, we think that it is useful because it shows how the method can be used to obtain a direct estimation of variables of interest, and why a model-derived ensemble is needed (see answer to comments below).

The central quantity in such an estimation problem is the ratio of observation variance to the model (ensemble) variance. This will dictate how far the prior is moved by the observations in creating the posterior. This is captured in your Probabilistic Scores, I think. But with simple messaging, the point could be made clearer.

Yes, it is important to quantify how much information is brought by the observations, which depends on the ratio between the observation error variance and the background error variance. In our problem, however, the variables are non-Gaussian, so that we decided to use probabilistic scores rather than variances. We added the following sentence in the paper to clarify this point.

« In the case of Gaussian variables, the gain of information brought by the observations (i.e. the resolution component of the score) is often characterized by variance ratios, which could have been computed here for the transformed variables (by anamorphosis). But we have preferred providing an assessment of the original concentration variable using the CRPS score. »

Figures need more details. You don't label the axes in some. You don't define what variables is being plotted in others. Etc.

Yes, thanks. This has been fixed.

An alternative approach is to use a parametric space-time covariance matrix. For example, a common approach is to use a Matern covariances for space, and auto-regression in time (and generally assume space-time separability for simplicity). This contrasts to your sample covariance matrix with post-processing (localization). Thoughts? Pros and cons?

Yes, in the optimal interpolation algorithm, it is possible to use parametric space-time correlation structures. However, it is much more difficult to specify cross-variable correlation structures with these methods, and thus to provide estimates of non-observed variables. This is why model-derived ensemble covariance structures are used. This point is specifically illustrated by the direct estimation of ecosystem indicators from the observations using (space and time dependent) model-derived correlation structures.

The way you do MCMC would be likely be called approximate Bayesian computation. That is, you make use of a cost function, rather than a more exact likelihood ratio.

In the MCMC sampler, the observation cost function is just the log of the likelihood ratio (as explained in the text following equation 2). It can be evaluated exactly, without the usual approximation of a quadratic cost function. The resulting sample is thus a sample of the posterior distribution, without approximation in the observation condition. However, the prior distribution is still assumed Gaussian (for the transformed variables, after anamorphosis), as part of our approximate formulation of the Bayesian estimation problem.

Stationarity is, in general, likely the assumption that limits the forecast horizon. Training on one time period, and applying it to another time period is predicated on stationarity. However, your short term forecasts of a few days means this is not an issue since it is driven by de-correlation timescales.

Yes, this is a very important point. In our approach, there is no assumption of stationarity of the statistics. Stationarity of the statistics is indeed usually assumed in classic optimal interpolation or in learning methods to transpose statistics obtained from past data to the present situation. In our approach, the prior ensemble is obtained from an ensemble model simulation that is specifically performed for the time period where the estimation is requested. This is especially important when the model is constrained by a time-dependent forcing function (in our case, the atmospheric forcing), which makes the ensemble statistics very instationary (maybe mostly seasonal, but not only). The following sentences have been added in the introduction to insist on this point.

« However, it is important to emphasize that, unlike EnOI, we do not use historical data to prescribe the prior statistics, but an ensemble simulation that is specifically performed for the requested time period. This is necessary if we want to avoid assuming stationarity of the statistics. »