EAT v0.9.6: a 1D testbed for physical-biogeochemical data assimilation in natural waters

Jorn Bruggeman, Karsten Bolding, Lars Nerger, Anna Teruzzi, Simone Spada, Jozef Skákala, and Stefano Ciavatta

RC2: 'Comment on gmd-2023-238', Anonymous Referee #2, 31 Jan 2024

Citation: https://doi.org/10.5194/gmd-2023-238-RC2

The authors present a generalized DA framework for 1D ocean applications. The proposed system, EAT, uses GOTM as the physical model, FABM as the biogeochemistry platform and PDAF as the DA software. The authors examined the new system in 3 different locations assimilation various physical and biogeochemical data. They tested state estimation in addition to state and parameter estimation. I believe the system is highly beneficial for the community and quite attractive given its portability and flexibility. The paper is well-written and easy to read. I only have minor comments.

We would like to thank the referee for the careful review and positive comments on our work. Our point-by-point replies are provided hereafter in green.

- When generating the initial ensemble, how did the authors decide on the variable distribution (e.g., lognormal) and it's associated parameters? And for the perturbed parameters, I am wondering why did the authors choose k_min and scale_factor over other ones.

We will include this information by adding the following (new text in bold):

"These parameters were selected to showcase the ability to introduce uncertainty in meteorological forcing (u10, v10) as well as the parametrization of the physical and biogeochemical models. Incorporation of other sources of uncertainty in the ensemble will be discussed at the end of this section. All ensemble members start from the same initial conditions; spread in the ensemble state first seen by the DA filter is generated by simulating 12 hours to the time of the first SST observation.

For simplicity, the same probability distribution was used to scale all five parameters: scale factors were drawn from a log-normal distribution with a standard deviation of 0.2 in natural log units. This was done independently per ensemble member, for each of the five scale factors (i.e., they are independently distributed). **The assumption of log-normality is common in biogeochemistry (e.g., Campbell et al., 1995) and ensures that the affected variables remain positive definite; we note, however, that both the type of distribution and its parameters are easy to customize (Fig. 2)."**

As mentioned in the above, the application will be extended with a paragraph discussing the incorporation of additional uncertainty in the ensemble:

"This application can be extended in several ways. For example, additional sources of uncertainty can be introduced when constructing the ensemble. The current setup includes a primitive parametrization of meteorological uncertainty through scaling of the surface wind components; more realistic experiments might source an ensemble of different meteorological model realizations (i.e., separate meteorological forcing files) and distribute those over the EAT ensemble members. EAT facilitates this by allowing its ensemble generators to set YAML parameters (e.g., the location of meteorological forcing in gotm.yaml) to member-specific file paths, similar to how the biogeochemical configuration

(`fabm/yaml_file`) is treated in Fig. 2. Another option is to introduce uncertainty in biogeochemical parameters other than phytoplankton maximum growth rate. This is easy to realize: all biogeochemical parameters are set in fabm.yaml and any of these can be varied across the ensemble by adding a single line in the ensemble generation script as in Fig. 2. Finally, it is possible to vary physical and biogeochemical initial conditions across the ensemble, as shown in the last section of Fig. 2."

The updated Fig. 2 is included in our reply to referee comment 1 (RC1).

- Transforming the state and the data during the update is interesting. I do believe that having Gaussian distributed variables is better suited for Kalman-type correction. My only question is when you transform the actual data, how do you deal with the associated observation error variance. For instance, if I take the logarithm of the data what would be the corresponding error in transformed space? I think the manuscript would benefit from such information.

We appreciate the opportunity to elaborate on this and will add the following:

"All biogeochemical variables were log-transformed **to guarantee positivity, as common in biogeochemical data assimilation (e.g., Santana-Falcón et al., 2020; Skákala et al., 2022; Pradhan et al., 2020). Transformation was done with a standard plugin provided by EAT (the Log plugin in Fig. 3), which applies log-transformation to all model variables and by default also to any associated observations. For the latter, the mean and variance of each log-transformed observation is reconstructed from the untransformed mean and variance under the assumption that each observation is perfectly log-normally distributed."**

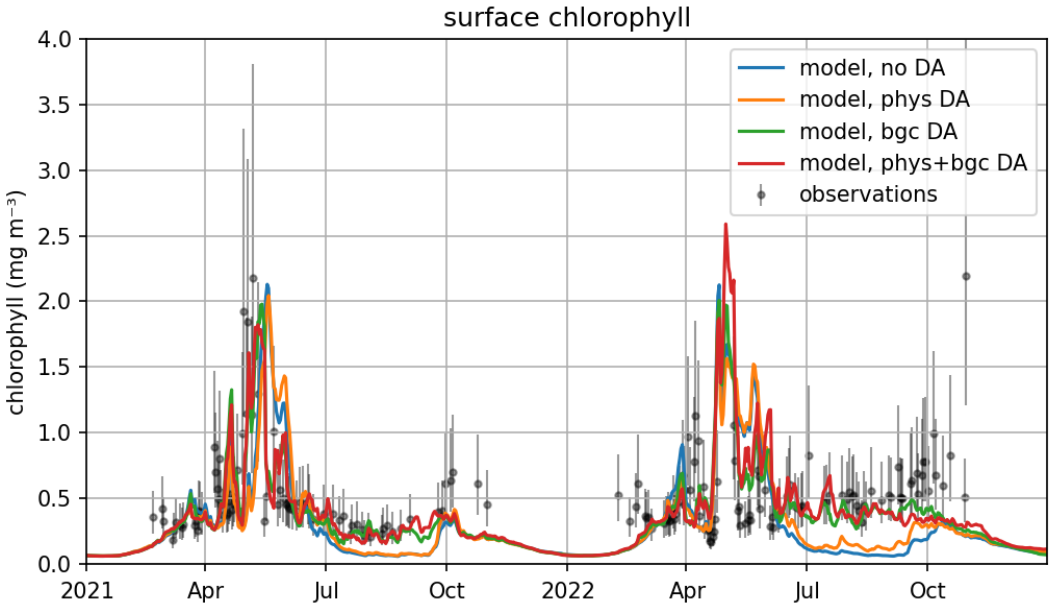
For further reference: the transformation logic described in the final sentence is implemented by https://github.com/BoldingBruggeman/eat/blob/797e3f3cad76bbeb2d929fe0b67601b344e164ff/eatpy/plugins/transform.py#L42-L53

- In section 3.1, why not add another experiment assimilating only bgc data? If physical data deteriorates chl then assimilating only chl and adding that to Fig. 6 should be very informative.

We agree that such an experiment could be informative and propose to add it to the application's Jupyter Notebook and to mention the experiment as possible extension at the end of the application section:

"Finally, EAT lends itself well for further experiments that investigate the impact of different types of observations. For instance, an experiment assimilating only surface chlorophyll (included in the application archive; see Data Availability section) could help ascertain whether coupled physical-biogeochemical assimilation performs better or worse than biogeochemistry-only assimilation."

We hope the referee will agree this is sufficient, as detailed treatment of the additional experiment and its results would excessively lengthen the application section (already the longest of the three) and complicate its figures. For example, including this additional time series in [already quite busy] Fig 6a makes it difficult to distinguish the time series of the individual experiments:



- It's very typical in similar 1D applications from the literature to assimilate nutrient profiles. I'm surprised the authors didn't consider that. Is it because the authors don't have access to such data in the tested locations? How about the reanalysis dataset? Addressing subsurface biogeochemical uncertainties can be crucial for adjusting PP across the entire water column (at least within euphotic zone).

EAT fully supports the assimilation of profiles, but to keep our examples both simple and representative of what can be easily achieved at any location with real observations (time series of remotely sensed temperature and chlorophyll are readily available everywhere), this is not done in the three included applications. We will now discuss the ability to add profiles in detail, as a possible extension of the first application:

"Another possible extension is to assimilate observations that describe not just the water surface, but also deeper layers. Notably, the inclusion of depth-explicit biogeochemical observations, e.g., from ship-based casts, automatic profilers or Argo floats, might help determine whether the decrease in subsurface chlorophyll in Fig. 6 is realistic or an artifact of surface-only chlorophyll assimilation. Inclusion of depth-explicit observations in EAT is straightforward, as it merely requires adding a column with depth information to the observation file and dropping the depth index (-1) in the linked model variable (Fig. 3)."

In the discussion, we will further note that the included applications demonstrate a subset of EAT's capabilities only:

"In general, EAT is flexible: user plugins are given full control over observations, forecasts, and analyses (with the ability to override proposed state updates). Ensemble members can differ in both state and configuration, and ensemble states can be saved and reused through support for restart files. These features can be combined in any number of ways to design new data assimilation experiments. Thus, the applications described here are representative of EAT functionality, but not exhaustive."

Related to my previous point, I believe assimilating data in the vertical may produce different parameter configurations at different levels beneath the surface.

This is correct in principle, but vertical variation in biogeochemical parameters is not supported by EAT, as it cannot be implemented without changing the underlying biogeochemical model codes (which explicitly treat parameters as space-invariant scalars). We will note this at the end of the parameter estimation example:

"Finally, it is worth noting that in EAT, parameters that are estimated during assimilation remain the same over the entire water column, even though they become variable in time. This is also common in 3D data assimilation (Doron et al., 2013). However, if parameter variability is assumed to stem from shifts in biological species composition, it is worth noting that the current approach cannot account for e.g. separate "light" and "shade" communities (Sournia, 1982), which would require parameter values to vary in depth. This is beyond the scope of EAT; the relevant parameters would need to be added to the depth-explicit model state within the biogeochemical model code."

Other comments:

- Line 20: that *are* sufficiently

We will correct this.

- Line 69: Water column models are *ideal testbeds*

We will correct this.

- Line 76: performing *state-parameter* estimation

We will correct this.

- Line 172: This allows *the* user

This section will be rewritten in response to RC1.

- Line 179: and that *runs* the data

The subject of this sentence is "scripts", therefore "run" is correct.

- Line 264: the model *serially*, without

We will rephrase this as:

"Finally, it is also possible to run the model stand-alone, without MPI; in this case, it behaves exactly as the original GOTM-FABM model would."

- Line 503: different *configuration* options

We will correct this.

Additional references

Campbell, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans*, *100*(C7), 13237–13254. https://doi.org/10.1029/95JC00458

Pradhan, H. K., Völker, C., Losa, S. N., Bracher, A., & Nerger, L. (2020). Global Assimilation of Ocean-Color Data of Phytoplankton Functional Types: Impact of Different Data Sets. *Journal of Geophysical Research: Oceans*, *125*(2). https://doi.org/10.1029/2019JC015586

Santana-Falcón, Y., Brasseur, P., Brankart, J. M., & Garnier, F. (2020). Assimilation of chlorophyll data into a stochastic ensemble simulation for the North Atlantic Ocean. *Ocean Science*, *16*(5), 1297–1315. https://doi.org/10.5194/os-16-1297-2020

Skákala, J., Bruggeman, J., Ford, D., Wakelin, S., Akpınar, A., Hull, T., Kaiser, J., Loveday, B. R., O'Dea, E., Williams, C. A. J., & Ciavatta, S. (2022). The impact of ocean biogeochemistry on physics and its consequences for modelling shelf seas. *Ocean Modelling*, *172*, 101976. https://doi.org/10.1016/j.ocemod.2022.101976

Sournia, A. (1982). Is There a Shade Flora in the Marine Plankton. *Journal of Plankton Research*, *4*(2), 391–399. https://doi.org/10.1093/plankt/4.2.391