Report on the use of 3D or 4D-Var for the ocean component in the coupled data assimilation context

Arthur Vidard

Work Package 2: Future coupling methods

Deliverable 2.5:

Report on the use of 3D or 4D-Var for the ocean component in the coupled data assimilation context

Type: Report

Author: Arthur Vidard

Reviewer(s): Matt Martin, Roberto Buizza

Delivered: 22/03/2017

Summary

This report describes the Inria contribution to WP2 regarding the use of 4D-Var in the ocean component of the reanalysis system. This task was twofold: 1. To study the relevance of the use of 4D-Var respect to 3D-Var. 2. To propose ways to reduce the computing cost, without altering the results. For the first part it turns out that 4D-Var is not really relevant for the ocean of the CERA reanalysis. Indeed at coarse resolution and with the use of observations of subsurface temperature and salinity only, the 3D-Var approximation is efficient enough. Getting a noticeable impact would require a data assimilation longer than 5 days, which is not feasible in coupled mode. For the CERA-SAT reanalysis however, mostly due to its use of SSH observations, 4D-Var does become a better option. The second part of the study shows that multi-incremental implementation of 4D-Var, even though challenging in the ocean, is a promising solution to significantly reduce the cost of 4D-Var, and could be used in future coupled reanalysis In addition to the scientific study of the above-mentioned points, it has required some significant technical developments. The first one was to include the ocean 4D-Var capabilities in the CERA system and was performed successfully in collaboration with P. Laloyaux at an early stage of the ERACLIM2 project. The second one was to develop the alternatives to full 4D-Var proposed in section 3. This was done in the framework of the NEMOVAR collaborative environment and is therefore available to the partners through, for instance, the ECMWF git repository more details about software transfer are given in appendix.

1 Introduction - CERA and Nemovar formulation

In the context of operational meteorology and oceanography, forecast skills heavily rely on proper combination of model prediction and available observations via data assimilation techniques. Historically, numerical weather prediction is made separately for the ocean and the atmosphere in an uncoupled way. However, in recent years, fully coupled ocean-atmosphere models are increasingly used in operational centres to improve the reliability of seasonal forecasts and tropical cyclones predictions. For coupled problems, the use of separated data assimilation schemes in each medium is not satisfactory since the result of such assimilation process is generally inconsistent across the interface, thus leading to unacceptable artefacts Mulholland et al. (2015). Hence, there is a strong need for adapting existing data assimilation techniques to the coupled framework, as initiated in Smith et al. (2015).

The CERA system (Laloyaux et al. 2015) aims at finding the coupled initial condition $\mathbf{x}_0 = \begin{pmatrix} \mathbf{x}_{a,0} \\ \mathbf{x}_{o,0} \end{pmatrix}$ that minimises

$$J(\mathbf{x}_0) = \left(\mathbf{x}_0 - \mathbf{x}^b\right)^T \mathbf{B}^{-1} \left(\mathbf{x}_0 - \mathbf{x}^b\right) + \sum_{i=0}^{N} \left(\mathcal{H}_{t_i}(\mathcal{M}_{t_i}(\mathbf{x}_0)) - \mathbf{y}^o_{t_i}\right)^T \mathbf{R}^{-1}_{t_i} \left(\mathcal{H}_{t_i}(\mathcal{M}_{t_i}(\mathbf{x}_0)) - \mathbf{y}^o_{t_i}\right) \quad (1)$$

where the first term is called the background term (denoted J_b) and the second term is commonly referred to as observation term (J_o)

- $\mathbf{B} = \begin{pmatrix} \mathbf{B}_a & 0 \\ 0 & \mathbf{B}_o \end{pmatrix}$ is the coupled system background error covariance and $\mathbf{B}_a$ and $\mathbf{B}_o$ are the atmosphere and ocean background error covariances matrices respectively (no cross system error covariances are considered)

- $\mathcal{H}(.) = \left(\begin{array}{cc} \mathcal{H}(.)_a & 0 \\ 0 & \mathcal{H}(.)_o \end{array}\right)$ is the observation operator whose components are $\mathcal{H}(.)_a$ for atmospheric and $\mathcal{H}(.)_o$ for oceanic observations.
- The numerical coupled model can be written with a slight abuse of notation as
$$\mathcal{M}(.) = \left(\begin{array}{cc} \mathcal{M}(.)_a & \mathcal{M}(.)_{ao} \\ \mathcal{M}(.)_{oa} & \mathcal{M}(.)_o \end{array}\right)$$
It includes both atmosphere ($\mathcal{M}(.)_a$) and ocean ($\mathcal{M}(.)_o$) as well as coupling ($\mathcal{M}(.)_{ao}$, $\mathcal{M}(.)_{oa}$) components

Due to non linearities in $\mathcal{M}$ and $\mathcal{H}$, minimising efficiently (1) is not straightforward. Common practice is to use the so called incremental 4D-Var approach (a.k.a. Gauss-Newton in the optimisation community) where the original problem is solved through successive minimisations of the quadratic cost functions (inner loops)

$$J^k(\delta\mathbf{x}^k) = \left(\delta\mathbf{x}^k\right)^T \mathbf{B}^{-1}\left(\delta\mathbf{x}^k\right) + \sum_{i=0}^{N}\left(\mathbf{H}_{t_i}^{k-1}\mathbf{M}_{t_i}^{k-1}\delta\mathbf{x}^k - \mathbf{d}_{t_i}^{k-1}\right)^T \mathbf{R}_{t_i}^{-1}\left(\mathbf{H}_{t_i}^{k-1}\mathbf{M}_{t_i}^{k-1}\delta\mathbf{x}^k - \mathbf{d}_{t_i}^{k-1}\right) \quad (2)$$

with $\mathbf{d}_{t_i}^0 = \mathbf{y}_{t_i}^o - \mathcal{H}_{t_i}(\mathcal{M}_{t_i}(\mathbf{x}_0))$ and $\mathbf{d}_{t_i}^k = \mathbf{y}_{t_i}^o - \mathcal{H}_{t_i}\left(\mathcal{M}_{t_i}(\mathbf{x}_0 + \sum_{i=1}^{k}\delta\mathbf{x}^i)\right)$

and $\mathbf{M}_{t_i}^k$ (resp. $\mathbf{H}_{t_i}^k$) being the tangent linear operator of $\mathcal{M}_{t_i}$ (resp $\mathcal{H}_{t_i}$) differentiated around $\mathbf{x}_0 + \sum_{i=1}^{k}\delta\mathbf{x}^i$

Non linearities are therefore accounted for through the re-linearisation of the $\mathbf{M}^k$ and $\mathbf{H}_{t_i}^k$ operators and the computation of the innovation vectors $\mathbf{d}^k$. Under some locality hypotheses, such algorithm is known to converge toward the solution of the original problem (see Gratton et al. (2007) for more details)

Constructing such system in a couple framework is challenging since individual models are generally developed separately and simulate physics with very different space and time scales. In addition, technical difficulties are important since the existing non-coupled data assimilation systems can use different algorithms to solve the problem presented above. Moreover an existing assimilation system is a very complex piece of software, that has been designed and tuned carefully. An additional hurdle is the differentiation of the coupling algorithm between models, which is often implemented in a non-differentiable manner.

For these reasons, in CERA an additional hypothesis has been made: both components of $\delta\mathbf{x} = \left(\begin{array}{c} \delta\mathbf{x}_a \\ \delta\mathbf{x}_o \end{array}\right)$ can be estimated separately thanks to the minimisation of two different cost functions. Additionnally, since it is based on the ECMWF's ocean reanalysis system Balmaseda et al. (2013) the ocean makes use of a further approximation, so-called 3D-Fgat (First Guess at Appropriate Time) or incremental 3D-Var, where the linear evolution of the ocean increment is assumed to be stationary over the assimilation window (i.e. $\mathbf{M}_{o,t_i} = \mathbf{I}$). Note that in the following, for the sake of readability, incremental 4D-Var and incremental 3D-Var will simply be called 4D-Var and 3D-Var respectively.

In summary, the quadratic functions to be minimised by the inner loops are those of IFS for the atmosphere and NEMOVAR for the ocean (see appendix A for more details on the actual NEMOVAR formulation).

$$J_a^k(\delta\mathbf{x}_a^k) = \left(\delta\mathbf{x}_a^k\right)^T \mathbf{B}^{-1}\left(\delta\mathbf{x}_a^k\right) + \sum_{i=0}^{N}\left(\mathbf{H}_{a,t_i}\mathbf{M}_{a,t_i}\delta\mathbf{x}_a^k - \mathbf{d}_{a,t_i}^k\right)^T \mathbf{R}_{t_i}^{-1}\left(\mathbf{H}_{a,t_i}\mathbf{M}_{a,t_i}\delta\mathbf{x}_a^k - \mathbf{d}_{a,t_i}^k\right) \quad (3)$$

and

$$J_o^k(\delta\mathbf{x}_o^k) = \left(\delta\mathbf{x}_o^k\right)^T \mathbf{B}^{-1} \left(\delta\mathbf{x}_o^k\right) + \sum_{i=0}^{N} \left(\mathbf{H}_{o,t_i}\delta\mathbf{x}_o^k - \mathbf{d}_{o,t_i}^k\right)^T \mathbf{R}_{t_i}^{-1} \left(\mathbf{H}_{o,t_i}\delta\mathbf{x}_o^k - \mathbf{d}_{o,t_i}^k\right) \tag{4}$$

As for the non-linearities, the coupling between the two systems is accounted for through the computations of the $\mathbf{d}^k$. Influence from data from the other component through the coupling in the outer iteration have also been illustrated in Laloyaux et al. (2015).

There is no properties of convergence toward the original problem (Equation 1) proven yet, but it is under consideration in the forthcoming deliverable D2.11

In any case such convergence is likely not achievable when approximation of the tangent model is used in the inner loop, one can only hope that the sought minimum is close to that of the original problem (Gratton et al. 2007). This closeness is obviously linked to the quality of that approximation, and said quality is linked to the configuration of the system (resolution, data sets, assimilation window). Studying the quality of possible approximations is the topic of this study. The next section will study both end of the spectrum from 3D-Var approximation (stationary tangent) to full tangent model, the third one will look into intermediate solutions such as simplified tangent model and multi-grid approaches.

2 4D-Var for the ocean

2.1 CERA-like settings

During the first year of the ERACLIM 2 project, 4D-Var capabilities for the ocean has been implemented in the CERA system, by importing NEMO tangent and adjoint models (see Vidard et al. (2015) for description of the models)). A few month of coupled reanalysis with CERA-like settings have been run with 4D-Var in both ocean and atmosphere, and compared to the same period with the original 4D-Var in the atmosphere / 3D-Var in the ocean. As for CERA settings, no additional ocean/atmosphere cross-correlation was added to the background term. Results were quite disappointing with no noticeable impact on the system. This suggest that, for this particular configuration (1° resolution, 1 day assimilation window, T and S profiles observations), the 3D-Var approximation is sufficient. Figure 1 shows the heat content (averaged top 300m temperature) of a typical assimilation increment in this configuration. It shows that there are barely no differences between the two schemes. At this resolution the ocean is not very active and therefore over one day considering the increment to be stationary is a sound hypothesis.

This would obviously not be true anymore were the assimilation window be longer, since the ocean would have time to evolve more significantly. Going beyond one day assimilation window is difficult in coupled mode because of the atmosphere, but it can be done in ocean-only mode. Figure 2 shows the mean ratio of both observation (J_o) and background (J_b) terms of the optimised non quadratic (outer) cost function respect to the assimilation window length (from 1 to 30 days). From this plot it appears that at this resolution and with this dataset, there is no clear improvement coming from 4D-Var for assimilation windows shorter than 5 days. Beyond that 4D-Var better fit the observations with a smaller increment, suggesting this increment to be more consistent with the ocean physics.

Mimicking Figure 1, Figure 3 shows the heat content increment for both 4D-Var and 3D-Var in the same configuration as before, but with a 30 days assimilation window. While large scale are

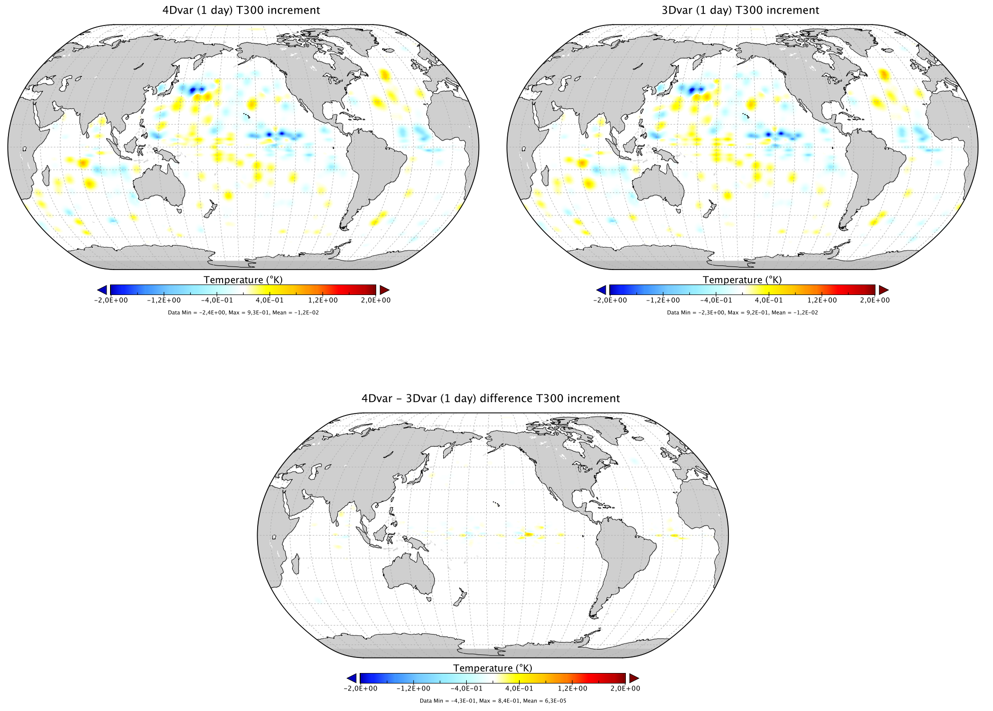
4Dvar (1 day) T300 increment
Temperature (°K)
-2,0E+00
-1,2E+00
-4,0E-01
4,0E-01
1,2E+00
2,0E+00
Data Min = -2,4E+00, Max = 9,3E-01, Mean = -1,2E-02
3Dvar (1 day) T300 increment
Temperature (°K)
-2,0E+00
-1,2E+00
-4,0E-01
4,0E-01
1,2E+00
2,0E+00
Data Min = -2,3E+00, Max = 9,2E-01, Mean = -1,2E-02
4Dvar - 3Dvar (1 day) difference T300 increment
Temperature (°K)
-2,0E+00
-1,2E+00
-4,0E-01
4,0E-01
1,2E+00
2,0E+00
Data Min = -4,3E-01, Max = 8,4E-01, Mean = 6,3E-05

Figure 1: ORCA 1 4D-Var and 3D-Var increments and differences of the averaged top 300m temperature, from a typical one day assimilation window, assimilating T and S profiles

quite similar albeit of smaller amplitude for 4D-Var, some small scale differences appears due to the tangent physics and dynamics.

As hinted in the text above, the impact of 4D-Var is also dependent on the model configuration and on assimilated data. Indeed, higher resolution models tend to be more active, limiting the validity of the 3D-Var approximation. Likewise assimilation of fast evolving quantities, such as sea surface anomaly and sea surface temperature, is likely to impact this validity as well.

Instead of reproducing the above sensitivity study for each combination of model configuration and observation dataset, one can try to predict the system behaviour using analytical tools. Indeed this is actually uniquely linked to the quality of the 'model' used to describe the evolution of the observed quantities in the inner loop for which robust quality estimators exist, as presented in the next section.

2.1.1 Estimation of the approximations error

Assessing the quality of a tangent model is tricky since there is generally no affordable exact tests available. A classical method of testing a numerical tangent linear model $\mathbf{M}$ is to compare the evolution of a perturbation by $\mathbf{M}$ with the difference of two evolutions, with and without the perturbation, by the full non-linear model $\mathcal{M}$.

Considering a fixed small perturbation vector $\delta\mathbf{x}_0$, and γ a scale parameter, the Taylor expansion of $\mathcal{M}$ reads:

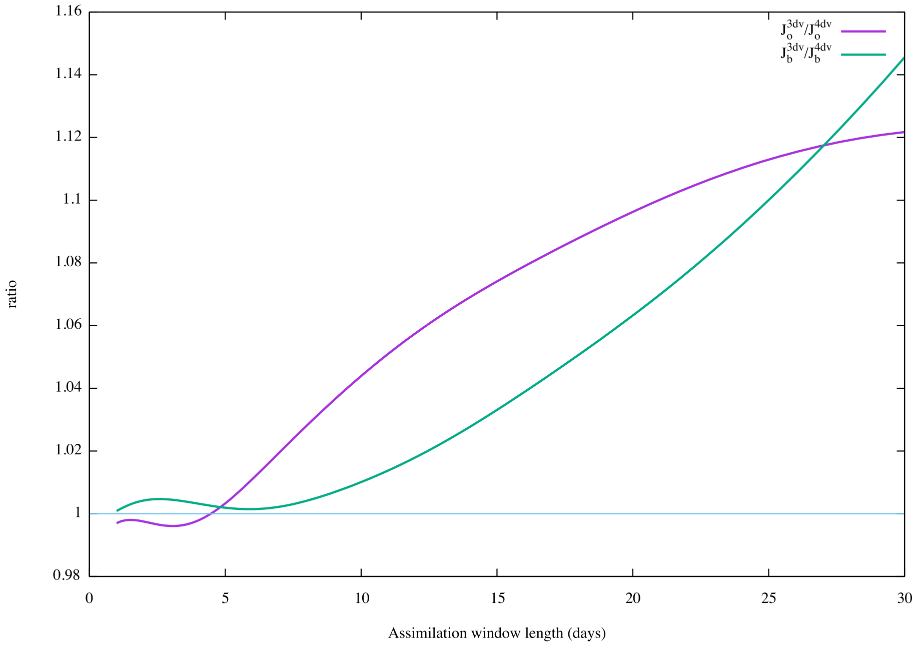

Figure 2: 3D-Var vs 4D-Var cost function ratio for J_o and J_b terms. 4D-Var is better when above 1 (thin line)

$$\mathcal{M}(\mathbf{x}_0 + \gamma\,\delta\mathbf{x}_0, t) = \mathcal{M}(\mathbf{x}_0, t) + \gamma\,\mathbf{M}(\mathbf{x}, t)\,\delta\mathbf{x}_0 + O(\gamma^2) \tag{5}$$

If $\mathcal{N}(\gamma\delta\mathbf{x}_0, t)$ denotes the non-linear evolution of a perturbation,

$$\mathcal{N}(\gamma\delta\mathbf{x}_0, t) = \mathcal{M}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0, t) - \mathcal{M}(\mathbf{x}_0, t) \tag{6}$$

The linearisation error $\mathcal{E}(\gamma\delta\mathbf{x}_0, t)$ is defined by:

$$\mathcal{E}(\gamma\delta\mathbf{x}_0, t) \quad = \quad \mathcal{N}(\gamma\delta\mathbf{x}_0, t) - \gamma\mathbf{M}(\mathbf{x}, t)\delta\mathbf{x}_0 \tag{7}$$

From (5), $\mathcal{E}(\gamma\delta\mathbf{x}_0, t)$ behaves like $O(\gamma^2)$.

By looking at $\mathcal{E}(\gamma\delta\mathbf{x}_0, t)$'s behaviour when γ tends to 0, one can validate an exact tangent model, however if $\mathbf{M}$ is not exact this test will fail without giving a precise information about the quality of the approximation.

To estimate the effect of approximations on the numerical tangent linear model $\mathbf{M}$, one must first estimate the truncated part of the Taylor expansion of equation (5). In order to do this, following Lawless et al. (2003), Vidard et al. (2015), one can write the Taylor expansion of $\mathcal{E}(\delta\mathbf{x}, t)$ whose individual components l follow:

$$\mathcal{E}_l(\delta\mathbf{x}_0, t) = \frac{1}{2}\partial^2\mathcal{M}_l\delta\mathbf{x}_0^2 + \frac{1}{6}\partial^3\mathcal{M}_l\delta\mathbf{x}_0^3 + \ldots \tag{8}$$

On the other hand, from two non-linear perturbations:

$$\begin{array}{rcl} \mathcal{N}(\delta\mathbf{x}_0, t) & = & \mathcal{M}(\mathbf{x}_0 + \delta\mathbf{x}_0, t) - \mathcal{M}(\mathbf{x}_0) \\ \mathcal{N}(\gamma\delta\mathbf{x}_0, t) & = & \mathcal{M}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0, t) - \mathcal{M}(\mathbf{x}_0) \end{array} \tag{9}$$

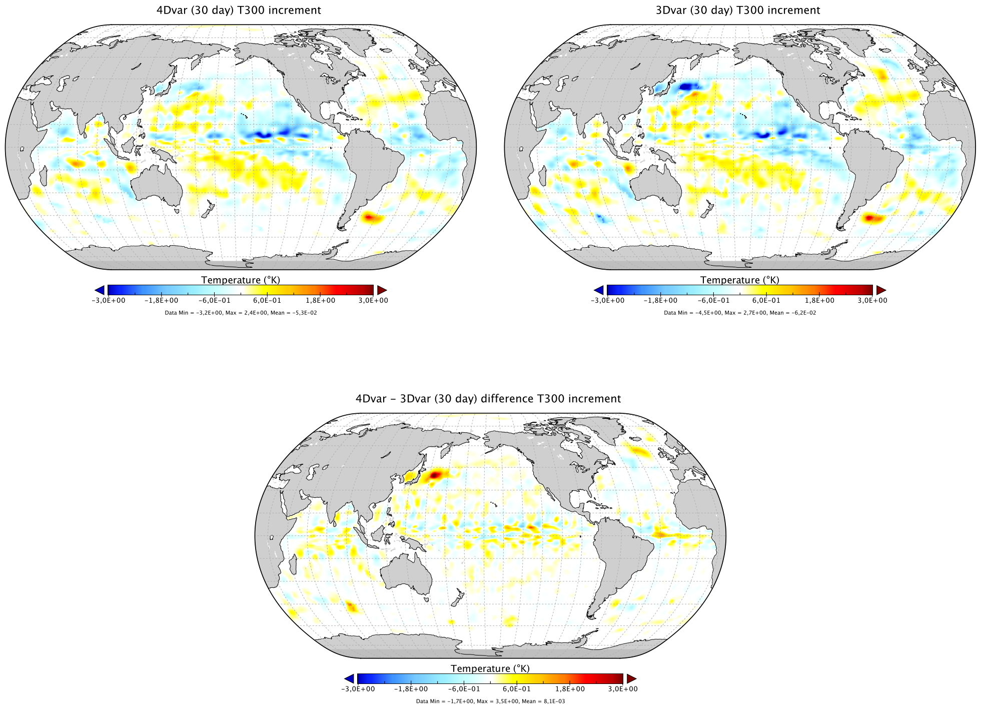
4Dvar (30 day) T300 increment
Temperature (°K)
-3,0E+00
-1,8E+00
-6,0E-01
6,0E-01
1,8E+00
3,0E+00
Data Min = -3,2E+00, Max = 2,4E+00, Mean = -5,3E-02
3Dvar (30 day) T300 increment
Temperature (°K)
-3,0E+00
-1,8E+00
-6,0E-01
6,0E-01
1,8E+00
3,0E+00
Data Min = -4,5E+00, Max = 2,7E+00, Mean = -6,2E-02
4Dvar - 3Dvar (30 day) difference T300 increment
Temperature (°K)
-3,0E+00
-1,8E+00
-6,0E-01
6,0E-01
1,8E+00
3,0E+00
Data Min = -1,7E+00, Max = 3,5E+00, Mean = 8,1E-03

Figure 3: ORCA 1° 4D-Var and 3D-Var increments and differences of the averaged top 300m temperature, from a typical thirty days assimilation window, assimilating T and S profiles

one can compute

$$\mathrm{E}(\gamma\delta\mathbf{x}_0, t) \quad = \quad \tfrac{\mathcal{N}(\gamma\delta\mathbf{x}_0,t)-\gamma\mathcal{N}(\delta\mathbf{x}_0,t)}{\gamma^2-\gamma} \tag{10}$$

whose Taylor expansion reads (for each individual component l)

$$\mathrm{E}_l(\gamma\delta\mathbf{x}_0, t) = \frac{1}{2}\partial^2\mathcal{M}_l\delta\mathbf{x}_0^2 + \frac{1+\gamma}{6}\partial^3\mathcal{M}_l\delta\mathbf{x}_0^3 + O(\gamma^4) \tag{11}$$

For small values of γ and $\delta\mathbf{x}_0$, one can compare E and $\mathcal{E}$. That way one builds-up an estimator of the numerical tangent linear model error:

$$\hat{\mathcal{E}} \quad = \quad 100\left(1 - \|\tfrac{\mathrm{E}}{\mathcal{E}}\|\right) \tag{12}$$

Moreover, in NEMO, the vast majority of the non-linearities are quadratic ones, meaning the third order and above derivatives vanish from the Taylor expansion and one gets $\mathrm{E} = \mathcal{E}$.

This diagnostic is very valuable when comparing different simplifications made to the tangent linear model. Here it can be used to assess the quality of the 3D-Var approximation. Figure 4 shows the evolution of $\hat{\mathcal{E}}$ with respect to the assimilation window length for the different component of the control vector *i.e.* , tracers (T,S), velocities (u, v) , and sea surface elevation (ssh).

This diagnostics confirm that the 3D-Var approximation is of reasonable quality for representing tracers' evolution at shorter time scale, indeed it is as good as the 4D-Var one for 1 day assimilation window and drift slowly and steadily away as time increases. For the other components of the control vector it is another matter, 3DFGat approximation being quite poor from day one for

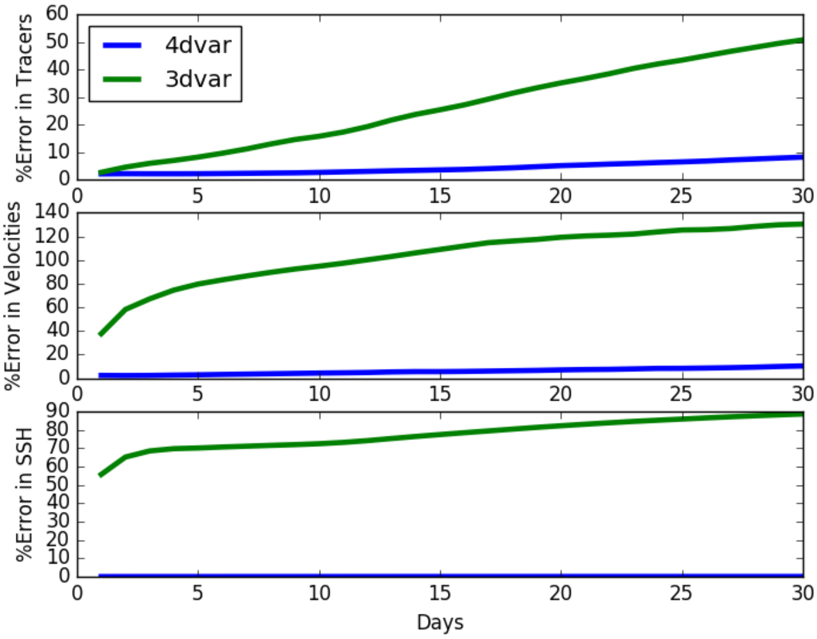

Figure 4: Inner loop approximation error (in percent) for 4D-Var (blue) and 3D-Var (green) according to window length for tracers (top), velocities(middle) and SSH (bottom), on ORCA 1° configuration.

velocities and sea surface elevation (40% and 55% errors respectively). This steep increase during the first day is mostly due to small scales perturbations that are evolving quickly. After a couple of days of fast adjustment, the approximation error in a similar way to that of tracers, but at a significantly higher level.

When assimilating only tracers, the 3D-Var inner loop approximation only requires to represent properly the tracers behaviour, the balance transform $\mathcal{U}$ (see Appendix A) being in charge of producing increments in velocities and sea surface elevation, explaining its good performance respect to 4D-Var.

Likewise, for assimilating Sea Level Anomalies (SLA), one needs the inner loop to model its evolution. Here the 3D-Var is likely to be less effective. Note that this test is quite stringent, so it does not mean that the 3D-Var will behave terribly, but instead that the potential benefit from switching to 4D-Var is important. This is indeed the case, while Figure 2 showed that, in case of assimilation of T and S profiles, no improvement came from 4D-Var for less than 5 days assimilation windows, assimilating SLA data only leads to an improvement of 20% on the fit to data (J_o) with 4D-Var respect to 3D-Var, with an increment norm (J_b) that is 3.5 times smaller!

In addition the linear evolution of a SSH perturbation is a good approximation of the nonlinear one, which makes the 4D-Var a good tool to assimilate such data. This is illustrated by Figure 5 that shows a typical SSH increment (left) and the corresponding linearisation error ($\mathcal{E}$ from equation 7)

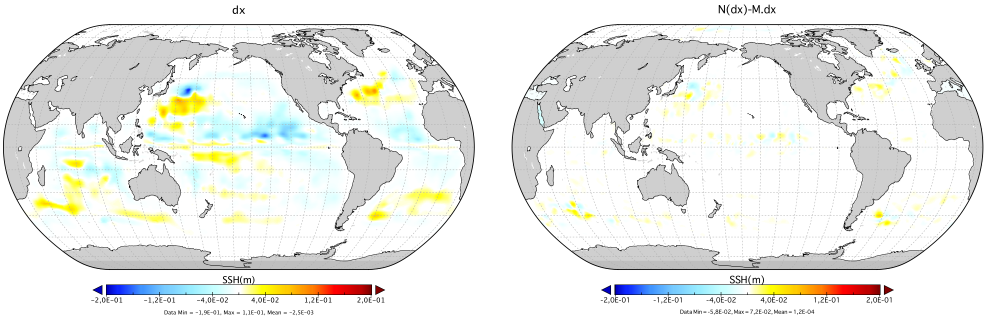

Figure 5: Typical 4D-Var SSH increment (left) and linearisation error after 30 days (right)

2.2 CERA-SAT like settings

CERA-SAT settings may be more prone to see improvements from 4D-Var in the ocean than original CERA. Indeed the ocean resolution has been increased to 1/4° and SLA observation are assimilated alongside T and S profiles. Performing the same diagnostic of inner loops approximation as for 1° resolution leads to a similar conclusion. Figure 6 shows the approximation error of the different control vector components. Once again for T and S, 3D-Var offers a good approximation for short time scales, but it remains below 10% error only for the first day, while it takes 5 days at 1° resolution. Note that 4D-Var as well reach 10% error after 5 days, while it takes 30 days at 1°, suggesting that some of the approximations made in the tangent model are a bit too strong at this resolution.

Approximations on SSH behaves similarly as the lower resolution as well, starting at 55% error from day 1 for 3D-Var, while remaining very small for 4D-Var. The importance of such difference can be illustrated with the ratio of cost function terms summarised in table 1. Here one notices that after minimisation the total (in situ + SSH) J_o is 50% to 80% larger for 3D-Var than for 4D-Var this is vastly due to the SSH contribution. At one day, 4D-Var halves the observation mismatch with SSH compared to 3D-Var, while they are equal for T and S profile observations. This is even amplified when going to 5 days assimilation window.

Assimilation window length	J_b^{3D}/J_b^{4D}	$J_o^{3D}/J_o^{4D}(all)$	$J_o^{3D}/J_o^{4D}(T/S)$	$J_o^{3D}/J_o^{4D}(SSH)$
1 day	1.2	1.5	1.	2.
5 days	1.3	1.8	1.1	2.6

Table 1: Final outer J_b and J_o ratio between 3D-Var and 4D-Var for an ocean configuration mimicking CERA-SAT, according to the length of the assimilation window.

Improving SSH does have an impact on other quantities. Figures 7 and 8 show the heat content of the 3D-Var and 4D-Var increments for typical 1 day and 5 day assimilation windows at 1/4° assimilating profiles and SLA data. It can be noted that most of the differences in this temperature increment are located around altimeter tracks (measuring SLA), which is well in accordance with results of table 1. The smaller increment norm (J_b) of 4D-Var suggest that by small remote adjustments (possibly in time and space, and across variables) this approach is able to better fit the observations,

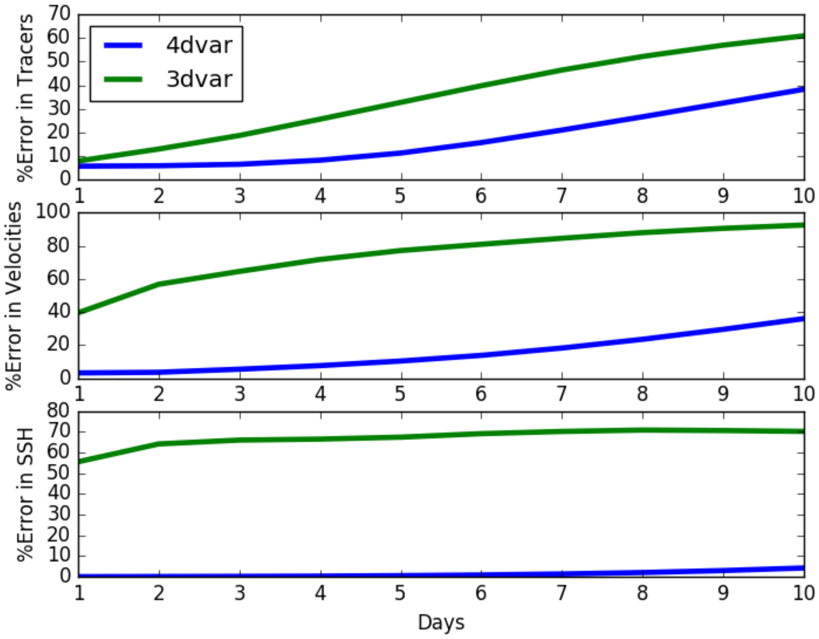

Figure 6: Inner loop approximation error (in percent) for 4D-Var (blue) and 3D-Var (red) according to window length for tracers (top), velocities(middle) and SSH (bottom), on ORCA 1/4° configuration.

4D-Var in the ocean could have been a preferred choice for CERA-SAT, however by increasing the resolution of the ocean, the computing cost of this component has become dominant. Going to 4D-Var would have made it impossible for the project to perform enough years of this stream of reanalysis, so it was decided not to. Next section discusses options to reduce the cost of 4D-Var, that could be used in future ocean or coupled reanalyses.

3 4D-Var cost reduction

The larger cost of 4D-Var compared to 3D-Var lies in the integration of the tangent model to compute the inner cost function and of the adjoint model to compute its gradient.

Currently tangent and adjoint models includes some simplifications, mainly on the computation of the vertical diffusion coefficients. The motivation for this simplification was the strong non-differentiabilities of the scheme used in Nemo. Since this part is relatively expensive, a side effect was to reduce noticeably the cost of the tangent and adjoint models. Some further approximation have been done in the handling of the non linear trajectory (see Vidard et al. (2015) for more details) so that the additional computing burden has been contained to around thrice the non-linear integration cost per inner iteration. This is a good ratio compared to standard 4D-Var implementations, but it can represents a huge over-cost compared to 3D-Var and some further optimisations / approximations have to be made.

Two main routes can be explored : further simplification of the tangent model, removing some of the processes it simulates and multi-grid approaches where part of the problem is solved at lower

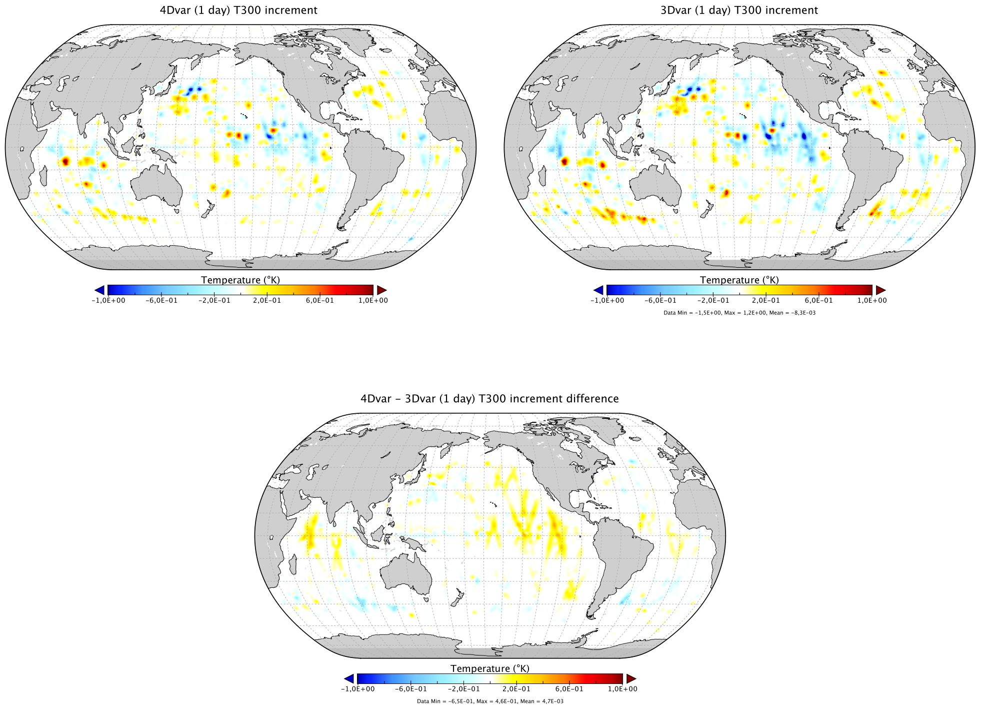
4Dvar (1 day) T300 increment
Temperature (°K)
-1,0E+00
-6,0E-01
-2,0E-01
2,0E-01
6,0E-01
1,0E+00
3Dvar (1 day) T300 increment
Temperature (°K)
-1,0E+00
-6,0E-01
-2,0E-01
2,0E-01
6,0E-01
1,0E+00
Data Min = -1,5E+00, Max = 1,2E+00, Mean = -8,3E-03
4Dvar - 3Dvar (1 day) T300 increment difference
Temperature (°K)
-1,0E+00
-6,0E-01
-2,0E-01
2,0E-01
6,0E-01
1,0E+00
Data Min = -6,5E-01, Max = 4,6E-01, Mean = 4,7E-03

Figure 7: ORCA 1/4° 4D-Var and 3D-Var increments and differences of the averaged top 300m temperature, from a typical one day assimilation window, assimilating T and S profiles and SLA

resolution. The former is only mentioned here and illustrated through a small example. The latter is studied in more details in the following sections.

3.1 Further simplification of the tangent model

One possibility for drastic simplification, is to focus on the part of the model that most influences the observed quantities. For example if one assimilates tracers data, one can assume that the most important part is the advection of such tracers. In NEMO it reads

$$\frac{\partial T}{\partial t} = -\nabla.(T\mathbf{U}) + D^{vT} + F^T \tag{13}$$

$$D^{vT} = \frac{\partial}{\partial z}\left(A^{vT}\frac{\partial T}{\partial z}\right) \tag{14}$$

where T is the temperature, $\mathbf{U}$ is the vector of 3D velocities, ∇ is the generalised derivative vector operator in (i, j, k) directions, t is the time, z is the vertical coordinate, D^T is the parameterisation of small-scale physics for temperature and F^T surface forcing terms. A^{vT} being the aforementioned tensor of vertical diffusion coefficients. A similar equation is describing transport of salinity.

Making the strong assumption that $\mathbf{U}$ does not depend on variation of T and S, the tangent

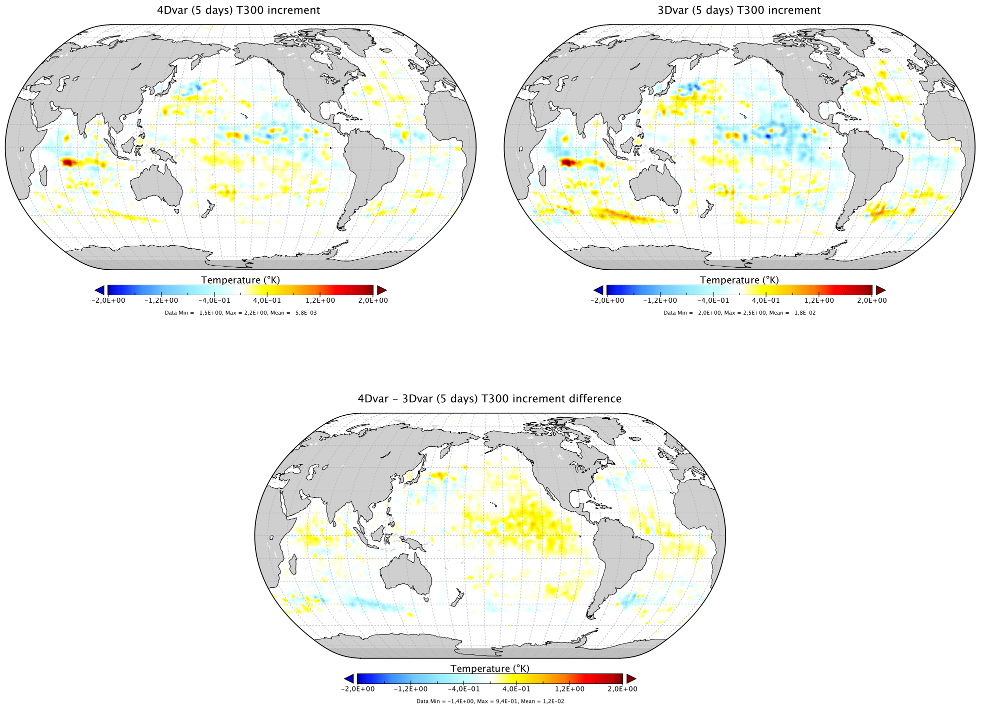
4Dvar (5 days) T300 increment
Temperature (°K)
-2,0E+00
-1,2E+00
-4,0E-01
4,0E-01
1,2E+00
2,0E+00
Data Min = -1,5E+00, Max = 2,2E+00, Mean = -5,8E-03
3Dvar (5 days) T300 increment
Temperature (°K)
-2,0E+00
-1,2E+00
-4,0E-01
4,0E-01
1,2E+00
2,0E+00
Data Min = -2,0E+00, Max = 2,5E+00, Mean = -1,8E-02
4Dvar - 3Dvar (5 days) T300 increment difference
Temperature (°K)
-2,0E+00
-1,2E+00
-4,0E-01
4,0E-01
1,2E+00
2,0E+00
Data Min = -1,4E+00, Max = 9,4E-01, Mean = 1,2E-02

Figure 8: ORCA 1/4° 4D-Var and 3D-Var increments and differences of the averaged top 300m temperature, from a typical five days assimilation window, assimilating T and S profiles and SLA

linear of the above system is

$$\frac{\partial \delta T}{\partial t} = -\nabla.(\delta T \mathbf{U}) + \delta D^{vT} \tag{15}$$

$$\delta D^{vT} = \frac{\partial}{\partial z}\left(A^{vT}\frac{\partial \delta T}{\partial z}\right) \tag{16}$$

Discarding all the other processes from the tangent (and adjoint) models is a strong approximation, but is weaker than that of 3D-Var for a significant reduction in cost compared to the standard 4D-Var. Figure 9 shows the approximation error for 4D-Var, 3D-Var, and simplified 4D-Var. Even though it is not as good as 4D-Var, it allows to remain below the 10% error for couple of days more than 3D-Var. As comes out of the first part of this report, it would be more interesting to focus the simplified tangent model on the evolution of SSH, but this quantity is the result of intricate contributions from velocities and tracers, therefore finding an appropriate simple system is tricky. A potential candidate could be a simple barotropic free surface equation, but it is not readily available within NEMO, and therefor would and therefore would require its implementation. It is anyway beyond the scope of this report.

3.2 Multi-grid techniques

In order to deal with the increase of computing cost and non linearities when going toward higher resolution applications, Numerical Weather Prediction centres (ECMWF, Météo-France,

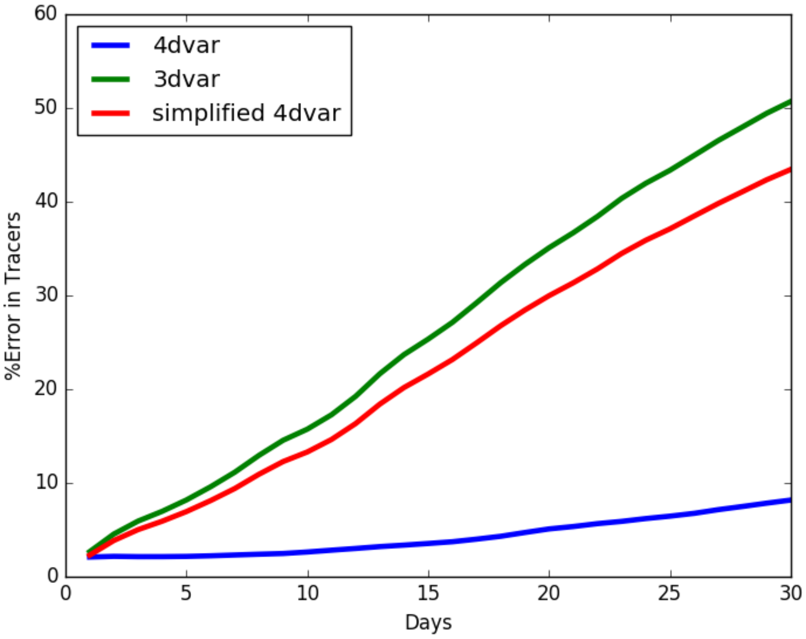

Figure 9: Inner loop approximation error (in percent) for 4D-Var (blue), 3D-Var (green), and simplified 4D-Var (red) according to window length for tracers (top), velocities(middle) and SSH (bottom), on ORCA 1° configuration.

UK-MetOffice, ...) usually use the so-called Multi-incremental approach: the models used in the successive minimisation of the inner loop are approximation of the tangent model (lower resolution and simplified physics). The first outer loop is performed using a very coarse grid for the inner loops models, and the resolution (and the physics) is improve for each subsequent outer loops. The inner loop approximation becomes, reusing notations from appendix A:

$$\mathcal{G}(\mathcal{U}(\mathbf{v}^{k-1} + \mathbf{T}_c^f \delta\mathbf{v}^k)) \approx \mathcal{G}(\mathcal{U}(\mathbf{v}^{k-1})) + \mathbf{T}_c^f \mathbf{G}_c^{k-1} \mathbf{U}_c^{k-1} \delta\mathbf{v}^k \tag{17}$$

where $\mathcal{G}$ is the generalised observation operator $\mathcal{G}(.) = \mathcal{M}(\mathcal{H}(.))$ and $\mathbf{G}$ its tangent, the index $_c$ meaning lower resolution, $\mathbf{T}_f^c$ is the simplification operator that translate high resolution fields into coarse ones and $\mathbf{T}_c^f$ the interpolation operator that goes the other way around.

Contrary to spectral models, transfer operators ($\mathbf{T}_f^c$ and $\mathbf{T}_c^f$) are not easy to define and implement on quadrilateral grids. This is particularly true for NEMO and its somewhat peculiar grid. The ORCA grid family denotes the preferred global ocean grids for NEMO. They are tripolar, meaning there are two north poles located on land (see figure 10) in order to avoid dealing with the north pole singularity. Doing so it requires the so-called northfold algorithm to handle the north periodicity, a somewhat complicated piece of code that requires special care when switching from one grid resolution to another. Due to this tripolar configuration, in the North hemisphere the orca grid is not aligned with meridian and parallels (and this misalignment can be different from one resolution to another). As a consequence vectors are not represented in a cartesian coordinate system. Moreover NEMO uses an Arakawa C-grid where horizontal velocities components are not located at the same place in the grid cell (see figure 11). So transferring velocity vectors requires to

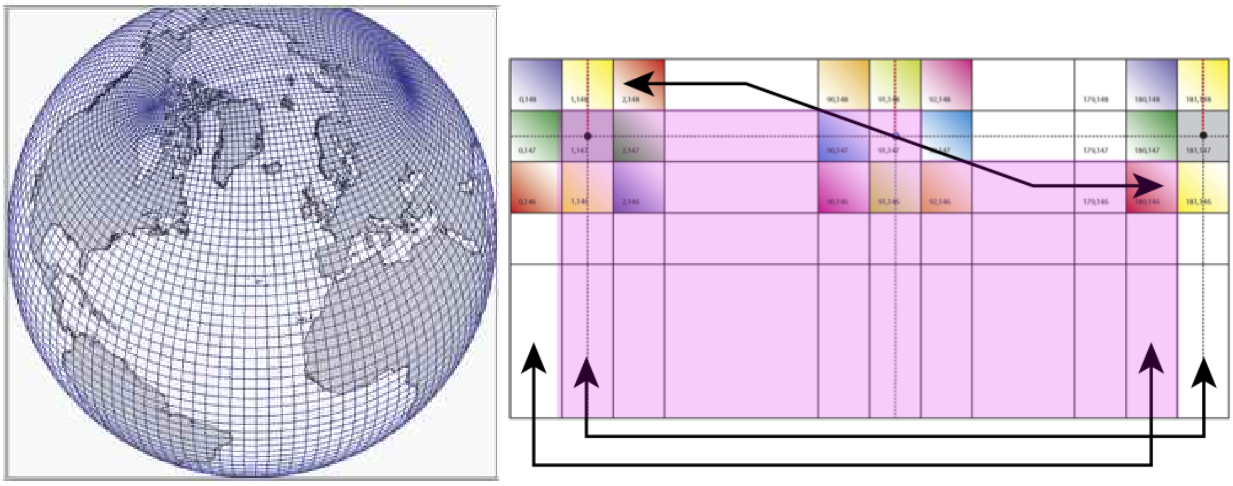

Figure 10: ORCA tripolar grid and Northfold exchanges (from NEMO documentation)

interpolate velocity components at T point, to rotate the vector on the cartesian grid, to transfer it to the coarse grid T point(s), to rotate it back to the orca grid, and finally interpolate back to u- and v-points.

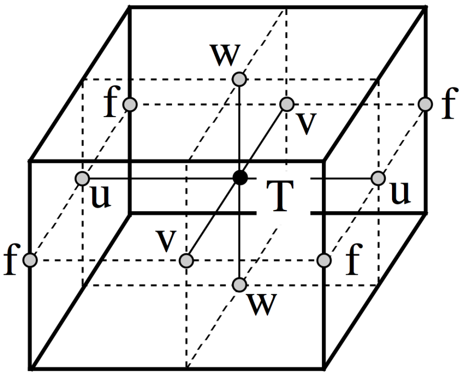

Figure 11: Arakawa c-grid in use in NEMO (from NEMO documentation)

Another difficulty is due to the complex boundaries of the discrete grid (coasts) that may lead to what can be called orphan points, *i.e* sea cells in the higher resolution grid that do not correspond to a sea cell in the coarser one. Depending on the application, dealing with orphan points may be an important matter. More details on this specific difficulty are given in the following

Next section discusses the current Nemovar transfer operators. It acts off-line, between outer iteration and inner loops which are 2 different executables. The following section presents the future on-line transfer operator currently under development.

3.2.1 Off-line transfer operator

This section describes the implementation of the Simplification Operator and its generalised inverse within the NEMOVAR framework. It first presents the context and highlights the difficulties associated to such operators. Section 3.2.1.b will quickly describe the interpolation operator (*i.e.* the generalised inverse of the simplification operator) with illustration an a practical test case. Section 3.2.1.c will describe more extensively the actual simplification operator and use the same test case as previously as illustration. Finally the inner loop approximation will be evaluated when using a coarse grid tangent model.

3.2.1.a Formalism

Multi-incremental versions of variational assimilation require an operator to transform ("simplify") the state vector from high resolution (HR) to low resolution (LR) and a generalised inverse of that operator to transform ("interpolate") the increment from low resolution to high resolution. The simplification operator, $\mathbf{T}_f^c$, is needed to define the basic state at low resolution of the linearised operators. It is easier to define the interpolation operator first ($\mathbf{T}_c^f$) and to derive the simplification operator from the solution that minimises the objective function

$$F(\mathbf{x}_c) = \frac{1}{2}[\mathbf{T}_c^f \mathbf{x}_c - \mathbf{x}]^T \, \mathbf{W} \, [\mathbf{T}_c^f \mathbf{x}_c - \mathbf{x}] \tag{18}$$

where $\mathbf{x}$ is the (known) $N \times 1$ state vector at high resolution, $\mathbf{x}_c$ is the (unknown) $L \times 1$ state vector at low resolution ($L < N$), and $\mathbf{W}$ is a $N \times N$ symmetric, positive definite weighting matrix. The minimising solution of (18) is

$$\mathbf{x}_c \;=\; \left[(\mathbf{T}_c^f)^T \, \mathbf{W} \, \mathbf{T}_c^f \right]^{-1} \, (\mathbf{T}_c^f)^T \, \mathbf{W} \, \mathbf{x} \;\equiv\; \mathbf{T}_f^c \, \mathbf{x}. \tag{19}$$

The operators $\mathbf{T}_f^c$ and $\mathbf{T}_c^f$ satisfy the mathematical properties $\mathbf{T}_f^c \mathbf{T}_c^f = \mathbf{I}$ and $\mathbf{T}_c^f \mathbf{T}_f^c = \mathbf{P} = \mathbf{P}^2$; i.e., transforming from LR → HR → LR does not change the solution, whereas transforming from HR → LR → HR is a projection. The filtering properties of $\mathbf{T}_f^c$ are controlled by the weighting matrix $\mathbf{W}$. Equation (19) requires the inversion of a large matrix. This could be achieved, for example, by iteratively minimising (18) using conjugate gradient method (available in NEMOVAR). Since the simplified state is required only for defining the basic state of linearised operators, it may be acceptable to replace (19) by an approximate solution,

$$\mathbf{x}_c \;\approx\; \mathbf{W}_c^{-1} \, (\mathbf{T}_c^f)^T \, \mathbf{W} \, \mathbf{x} \tag{20}$$

where $\mathbf{W}_c$ is a $L \times L$ symmetric, positive definite weighting matrix with simpler structure than the matrix $\left[(\mathbf{T}_c^f)^T \, \mathbf{W} \, \mathbf{T}_c^f \right]^{-1}$ in (19). Equation (20) may be interpreted as the adjoint of the interpolation operator with respect to the inner products $\mathbf{x}^T \mathbf{W} \mathbf{x}$ on HR-space and $\mathbf{x}_c^T \mathbf{W}_c \mathbf{x}_c$ on LR-space. The choice of $\mathbf{W}_c$ is a key point and will be discussed later on.

3.2.1.b Interpolation operator

The general interpolation operators (and their adjoints) used in the profile and altimeter observation operators have been exploited here in defining $\mathbf{T}_c^f$ (and $(\mathbf{T}_c^f)^T$) considering that the location of

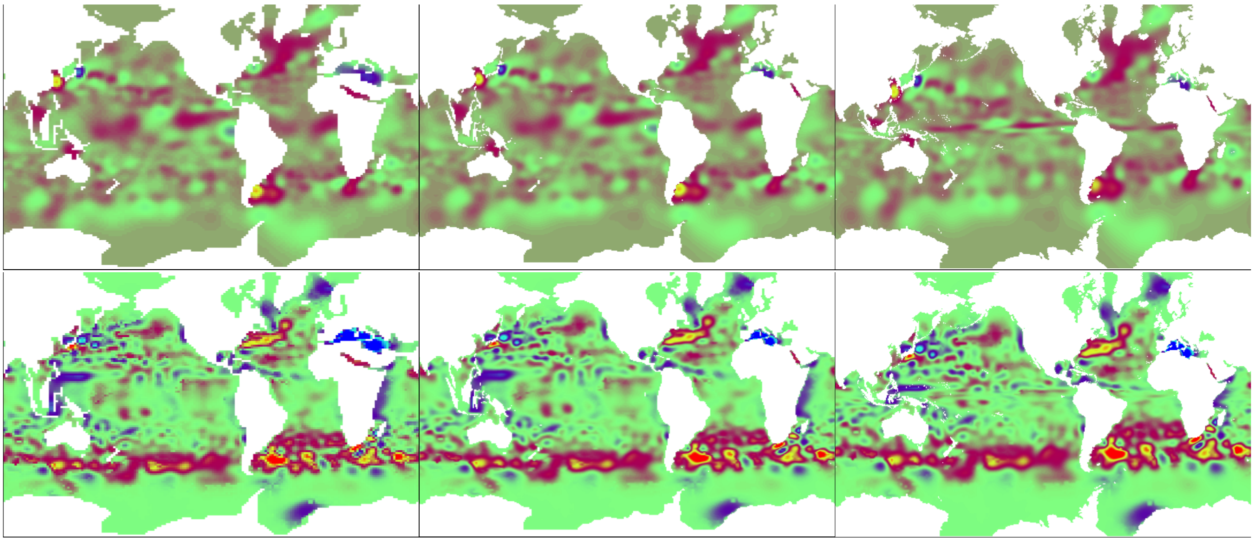

Figure 12: Interpolation of a temperature (top) and SSH (bottom) increments from Orca2° (left) to Orca1°(middle) and Orca$\frac{1}{4}^{\circ}$ (right)

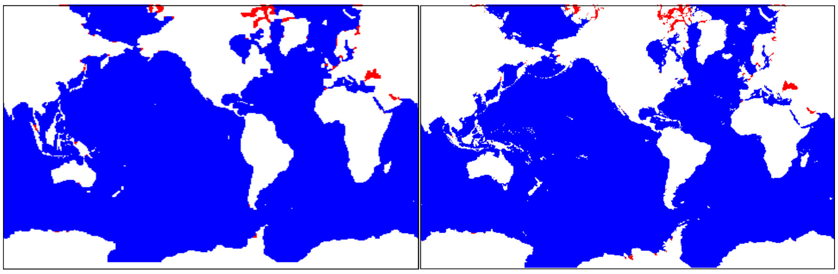

Figure 13: In red: Orphan high-resolution Sea Points where the interpolation failed to find a suitable value

each High-Resolution grid-point as an observation location. The grid search and the horizontal and vertical interpolation routine are exactly the same, the interface that calls the interpolation routines has been adapted to NEMO gridded data.

The fields that requires to be interpolated from low resolution to high resolution are restricted to assimilation increments. The Interpolation operator behave reasonably well with no obvious problems along the coast and at singular points as shown in figure 12. A special treatment has been done for sea point at higher resolution that are in the interior of land areas in the low resolution, such as closed, sea, channels or fjord (coastal point are generally OK, see Fig 13). In these areas, there is no obvious way to estimate the increment. It has therefore been set to 0, which is fine since we are only treating increments. If one needs to use the interpolator for the full field, this particular problem will need to be addressed.

Regarding the CPU cost, the interpolation of an entire assimilation increment (four 3D fields and one 2D field) from Orca2° to Orca1° (resp Orca$\frac{1}{4}^{\circ}$) takes 35s (resp 8mn) on a common workstation

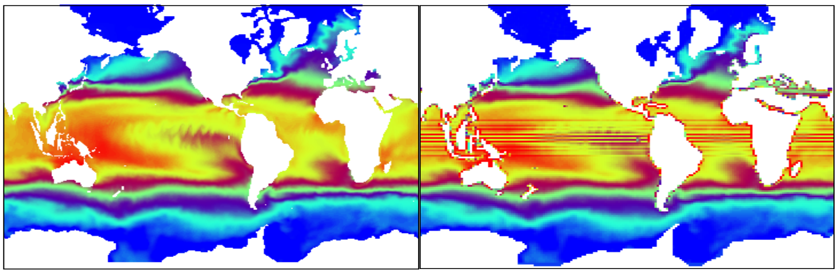

Figure 14: Original Orca1° background temperature field and the simplified Orca2° field obtained with the approximate simplification operator and using the LR volume elements as weighting matrix

3.2.1.c Simplification operator

Approximate Simplification operator - the choice of $\mathbf{W}_c$ A simple and natural choice for $\mathbf{W}$ is the diagonal matrix of volume elements (scale factors) that define the HR model grid. Therefore, the obvious idea for $\mathbf{W}_c$, is to approximate it by a diagonal matrix of volume elements for the LR model grid. However this leads to an unsatisfactory result (to say the least) as shown in Fig. 14.

However surprising at first, this bad behaviour has a clear explanation. The bad results are located mainly along the coasts and in the equatorial belt. These are actually coming from 2 separate problems:

- Along the coasts and around the islands: This kind of weighting does not account for the land point of the HR grid that are sea point of the LR-grid. Indeed the weight is the same whatever the number of sea-point present in the HR cells used to compute the value of the corresponding LR cell

- Equatorial belt: this is a bit more complicated and is case dependent. This is actually due to the increase of meridional resolution in both ORCA1° and ORCA2°. Looking, for instance, at what happens along the 180°th meridian (see Fig. 15) In both extra equatorial cases, everything is fine, with the adjoint of the interpolation operator, the contributions to the lower resolution grid come, in the south, from the two HR points before and the two after and, in the north, from the one before, the one after and the one right spot on. The weighting by the volume elements do the trick. In the third case (Fig. 15 bottom), and that is what is happening in the equatorial belt, some problems arise. For instance, the LR point at the equator (0,180) gets contributions from the HR point at the equator (with weight 1) and the two surrounding HR grid points (both with weight 1/3) (total weight = 5/3) ; while the point at (0.5,180) gets contributions from the two surrounding HR grid point only, both with weight 2/3 (total weight = 4/3). and there is a cycling of this 5/3 - 4/3. The volume elements being roughly constant for each grid in that area, its use as a weight cannot correct this problem and we get the oscillation seen on Fig. 14.

If we assume that the weighting matrix $\mathbf{W}_c$ is diagonal, which is a reasonable assumption, the computation of $\mathbf{W}_c$ is actually straightforward. Indeed, remember that we want to find a weighting matrix such as $\mathbf{T}_f^c\mathbf{T}_c^f = \mathbf{I}$, so combining this with equations (19) and (20) we get

$$\begin{aligned} \mathbf{x} &\approx \mathbf{W}_c^{-1}\,(\mathbf{T}_c^f)^T\,\mathbf{W}\,\mathbf{T}_c^f\mathbf{x} \\ \mathbf{W}_c\mathbf{x} &\approx (\mathbf{T}_c^f)^T\,\mathbf{W}_f\,\mathbf{T}_c^f\mathbf{x} \end{aligned} \tag{21}$$

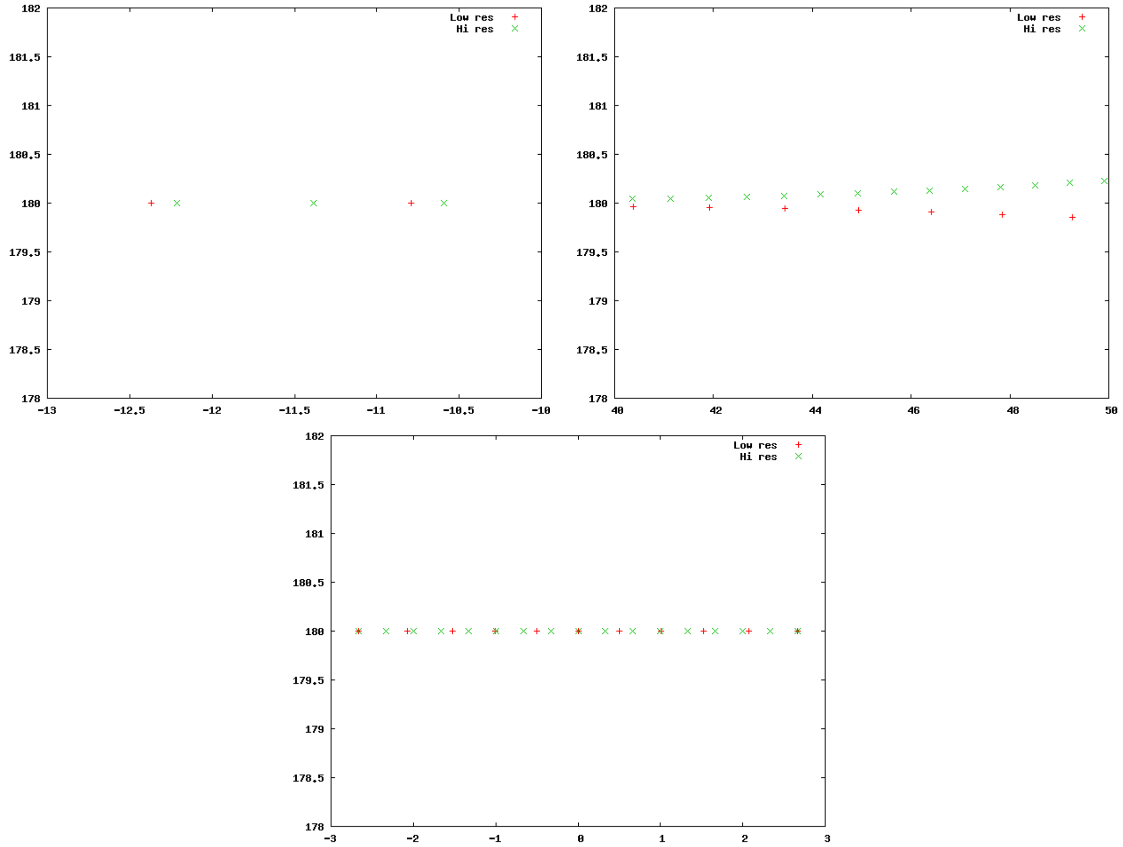

Figure 15: Grid points locations along the 180°th meridian for some part of south (top left) and north hemisphere and for the equatorial belt (bottom). Red pluses are the locations of the ORCA2° T-grid points and the green crosses are the ORCA1° T-grid points.

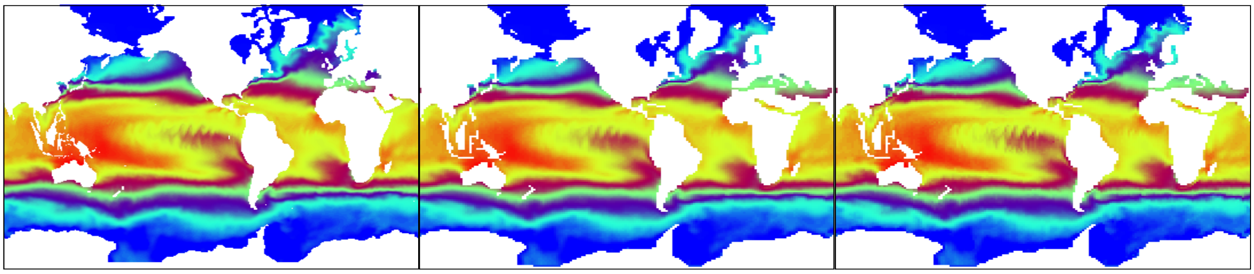

Figure 16: Original Orca1° background temperature field (left) and the simplified Orca2° field obtained with the Approximate Simplification Operator (middle) and the Full Simplification Operator (right)

For all $\mathbf{x}$ from the high resolution grid. This is in particular true for $\mathbf{x} = 1_{HR}$ the vector with 1 everywhere. Thus we get, since we assumed $\mathbf{W}_c$ diagonal,

$$\mathbf{W}_c \approx (\mathbf{T}_c^f)^T \, \mathbf{W}_f \, \mathbf{T}_c^f 1_{HR} \tag{22}$$

that can be computed thanks to the interpolation operator and its adjoint.

Full Simplification operator (FSO) versus Approximate Simplification Operator (ASO) Alternatively the FSO can be applied through the minimisation of the function described in equation 18. This leads to a different result with slightly sharper details as shown in Fig. 16. The fact that the ASO tends to smooth out details may not be detrimental, indeed small scales features may not be well represented by the coarse grid model. For that reason, it is not obvious to assess whether it is better to use the output from the ASO or from the FSO as background state of the inner loop, and a more in depth study would be required.

Obviously, the result of the FSO is closer to the actual $\mathbf{S}$ than the one from ASO, but it comes at a cost. Figure 17 shows the norm $\parallel \mathbf{S}\mathbf{S}^{-I} - I \parallel^2$ (that should be equal to zero for a perfect $\mathbf{T}_f^c$) along with the CPU cost on a standard workstation respect to the number of minimisation iteration for the FSO using the result of the ASO as a first guess (therefore 0 iterations means ASO). A good compromise may be to do only a few iterations of FSO (less than 5).

Coarse grid inner loop approximation error Once again, one can evaluate the inner loop approximation error using the coarse grid approximation. Figure 18 shows the usual approximation error for 3D-Var and coarse grid 4D-Var. Note that here, the perturbation is generated on the coarse grid and interpolated on the fine grid. Doing so may be favourable to the coarse grid approximation, since it does not include sub-grid scale perturbations. As a result the level of error for 3D-Var is equivalent to that of ORCA 1°, even though it is computed on the ORCA 1/4° configuration. In this framework the reduced grid approximation shows a significant improvement over 3D-Var, being almost equivalent to full grid for velocities and SSH.

As an illustration of the multi incremental capabilities introduced in Nemovar thanks to the transfer operator described in this section, figure 19 shows the increments in SSH computed by a 5 day assimilation of T, S and SSH on the ORCA 1/4° configuration. The top left panel is the result of a full grid assimilation, the top right panel is the result of a first inner loop at coarse resolution interpolated at the outer resolution and the bottom panel is the result of a subsequent second inner loop at full resolution but with only a couple of iterations. It shows that the coarse

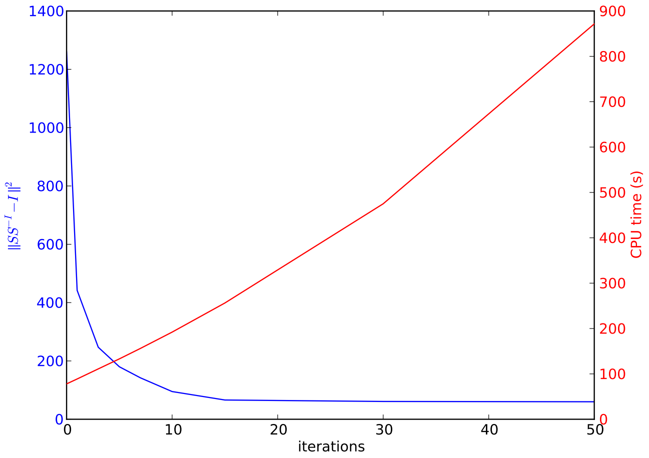

Figure 17: Error on the simplification operator (blue) and CPU time (red) regarding the number of FSO iterations (0=ASO)

grid inner loop is able to reproduce the large scale signal, but misses some of the smaller scales that probably cannot be reproduced on ORCA 1°. A second outer iteration with only a couple of high resolution inner iterations is able to retrieve the smaller scale signal.

3.2.2 On-line transfer operator

Performing the transfer off-line as presented before is easier technically, since it does not require NEMOVAR to handle two different grids at the same time. However it restricts its use to multi-incremental applications. Indeed the representation of the background error statistics requires the inversion of large linear system. This part of the code becomes really expensive as the resolution increases. Using proper multi-grid techniques for this specific part of NEMOVAR could be very beneficial. Also other kind of multi-grid 4DVar algorithms such as FAS or multi-grid preconditioning as advocated by Debreu et al. (2015), require multi-grid operations within the inner loop.

It has been decided to restrict the possible grids to grids being multiple of each other, meaning coarser grids can be obtained by subsampling the finer grid. The main reasons for that being it has the advantage to allow the same domain decomposition on all grids and to render grid searches trivial. It may appear restrictive, but this is actually not a very strong constraint. Even though it prevents from using some standard NEMO grids such as ORCA2 and ORCA1 as coarse grids, from ORCA1/2° fineward standard ORCA configurations fits this requirement (G. Madec, personal communication).

A grid generating tool is anyway implemented as part of the on-line transfer operator in case no standard NEMO configuration at required resolution is available. It uses the same strategy as AGRIF nesting tool package. For an even coarsening factor, in order to avoid ignoring some area at the boundaries, it is not a simple subsampling, indeed T-points become F-points and vice versa and U (V) points are combination of U (V) and F coordinates (see figure 20). Ideally, in order to avoid orphan point, a coarse grid point should be a sea point if and only if at least one of the corresponding fine points is a sea point. However for most case it would create undesirable connection between basins in particular it would open wide the Panama isthmus.

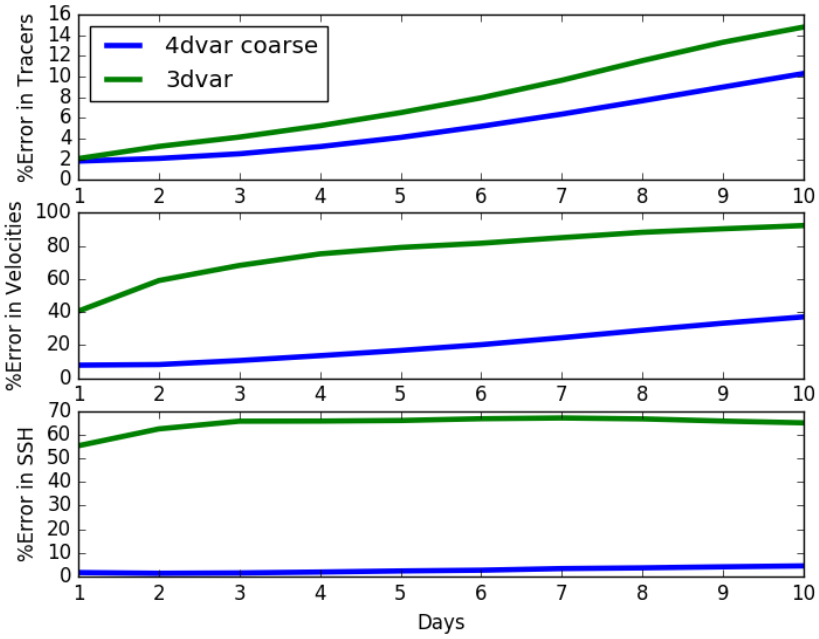

Figure 18: Inner loop approximation error (in percent) for 3D-Var (green), and coarse-grid 4D-Var (blue) according to window length for tracers (top), velocities(middle) and SSH (bottom), on ORCA 1/4° configuration (coarse grid at 1°).

In the off-line transfer operators there is no explicit smoothing that is generated by the restriction, some can be achieved using the approximate formulation, but it is limited. This is a possible drawback, smoothing can indeed be important in some multi-grid applications (see Debreu et al. (2015) and bibliography for more details). For this reason, the on-line transfer operator includes a smoothing filter.

More precisely, in current implementation the interpolation operator $\mathbf{T}_c^f$ uses a combination of a simple filter, here a Shapiro filter ($\mathbf{F}$), and a basic extension operator ($\mathbf{E}$).

We seek the total filtering effect to be that of a "2-iterations" Shapiro filter ($\mathbf{F}_2$), built from the product of a single-iteration Shapiro filter (F) and its adjoint respect to the fine grid metric ($\mathbf{F}^* = \mathbf{W}_f^{-1}\mathbf{F}^T\mathbf{W}_f$:

$$\mathbf{F}_2 = \mathbf{F}\mathbf{F}^* = \mathbf{F}\mathbf{W}_f^{-1}\mathbf{F}^T\mathbf{W}_f$$

where $\mathbf{W}_f$ contains the area/volume elements for the fine grid. Such formulation ensure that $\mathbf{F}_2$ is self-adjoint respect to the fine grid metric.

The role of the Shapiro filter, is to remove only the scales from the fine grid which are not represented on the coarse grid. The diffusion operator, on the other hand, accounts for the correlation length scales.

Assuming the coarse grid is sampled from the fine grid, the extension operator $\mathbf{E}$ will be a rectangular matrix of 1s and 0s. This would be true only for odd refinement factor, extension operator for an even refinement factor would include some local averaging or interpolation (see figure 20) .

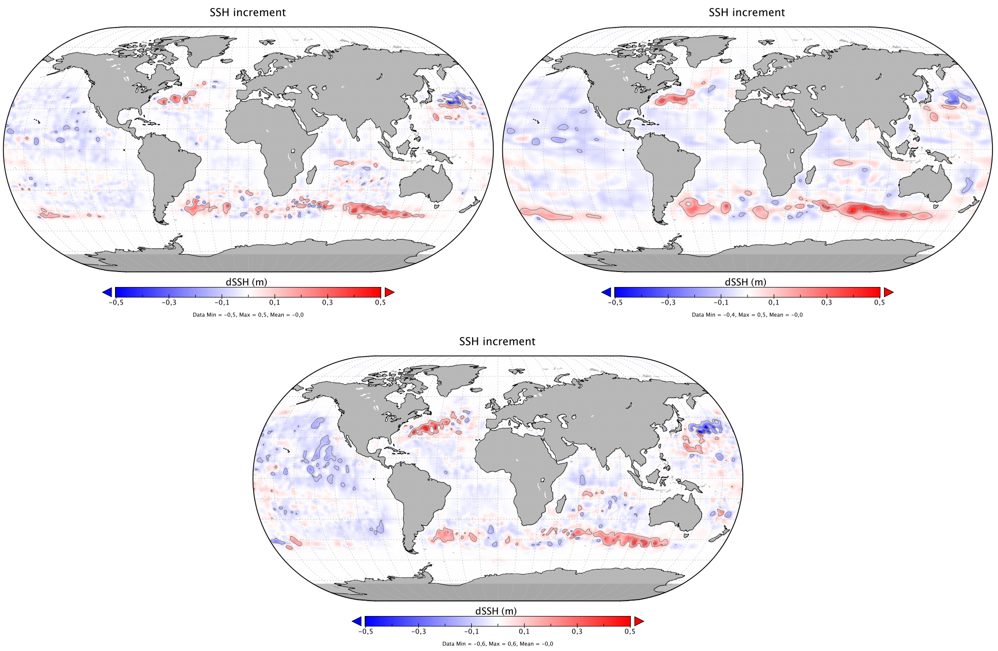

Figure 19: SSH increment from ORCA 1/4° full grid and ORCA 1/4°/ORCA 1° coarse resolution first inner loop and fine resolution second inner loop assimilations

The interpolation operator is then

$$\mathbf{S}^{-\mathbf{I}} = \mathbf{FE}$$

The restriction operator ($\mathbf{R}$) (fine to coarse) is defined as the adjoint of $\mathbf{E}$:

$$\mathbf{R} = \mathbf{W}_c^{-1}\mathbf{E}^T\mathbf{W}_f$$

and the simplification operator $\mathbf{T}_f^c$ is defined as the combination of the restriction operator and the adjoint of the filtering.

$$\mathbf{T}_f^c = \mathbf{RF}^* = \mathbf{W}_c^{-1}\mathbf{E}^T\mathbf{W}_f W_f^{-1}\mathbf{F}^T\mathbf{W}_f = \mathbf{W}_c^{-1}\mathbf{E}^T\mathbf{F}^T\mathbf{W}_f$$

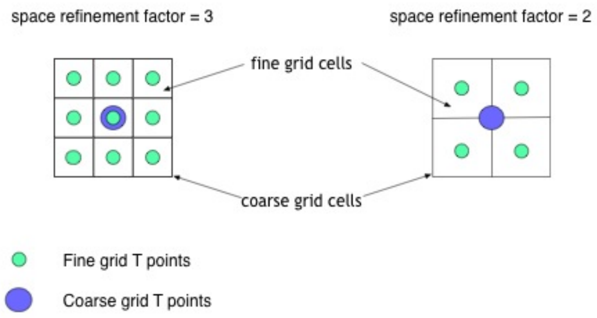

Figure 20: Refinement strategy (from Agrif's nesting tool documentation)

At time of writing this report, the implementation of this transfer operator is being finalised within the new NEMOVAR version that is in particular hosted at ECMWF's repository, and should be available, tested and validated before the end of ERACLIM2 project.

4 Conclusion

To the original question "is it worth using 4D-Var in the ocean for the CERA-20C reanalysis?" the answer is clearly "no" for short assimilation window (shorter than 5d) as is used in the CERA system. Indeed at low resolution, assimilating only T and S subsurface data on a short time window, the 3D-Var approximation holds well and there is no need for more complexity. However when assimilating observation of faster evolving quantities such as Sea Level Anomalies and probably Sea Surface Temperature, accounting for the dynamics of the system can be significantly beneficial, even at short time scales.

It could then have been an option for the CERA-SAT reanalysis which assimilate such observations. However, due to the significant extra cost of 4D-Var with respect to 3D-Var it was decided not to use 4D-Var in the ocean component in order to avoid further reduction of the length of this reanalysis. Two options for reducing this extra cost have been investigated here: drastic simplifications of the modelled processes used in the inner loop of the ocean analysis system, and the use of multi-grid techniques. The latter being more promising, but also more complex to implement. Existing multi-grid tools have given encouraging results and a more flexible new version is being implemented as a final contribution from this task to the ERA-CLIM2 project.

The table below summarises the computing cost of some of the options discussed in this document. In particular one can notice that a 3D-Var at Orca025 full resolution is similar in cost as a multi incremental 4Dvar using ORCA1 in the inner loop.

Configuration	3D-Var	4D-Var	Multi-inc 3D-Var	Multi-inc 4D-Var
ORCA1, 10 iterations, 1 day (1 node)	6mn (11mn)	12mn (17mn)	–	–
ORCA1, 10 iterations, 10 day (1 node)	6mn (16mn)	48mn (1h)	–	–
ORCA025, 5 iterations, 5 days (6 nodes)	45mn (2h45)	7h (9h)	2mn (2h)	45mn (2h45)

Table 2: Comparative computing time on our local cluster for selected options and configurations for the inner loop and for the total assimilation cycle (in parenthesis). Multi-incremental is done using ORCA1 in the inner loop.

A Change of Variables in NEMOVAR

The variational assimilation problem solved by NEMOVAR is defined very generally as the minimisation of a cost function of the form

$$J[\mathbf{v}] \;=\; \frac{1}{2}\left[\mathbf{v}-\mathbf{v}^{b}\right]^{T}\left[\mathbf{v}-\mathbf{v}^{b}\right] \;+\; \frac{1}{2}\left[\mathcal{G}\left(\mathcal{U}(\mathbf{v})\right)-\mathbf{y}^{o}\right]^{T}\mathbf{R}^{-1}\left[\mathcal{G}\left(\mathcal{U}(\mathbf{v})\right)-\mathbf{y}^{o}\right] \tag{23}$$

where $\mathbf{v}$ is the control (analysis) vector, $\mathbf{v}^b$ is the background estimate of the control vector, $\mathbf{y}^o$ is the vector of observations, $\mathbf{R}$ is an estimate of the observation error covariance matrix, $\mathcal{U}$ is a (possibly) nonlinear operator that maps the control vector onto the model state (initial condition) space ($\mathbf{x} = \mathcal{U}(\mathbf{v})$)[1] , and $\mathcal{G}$ is a nonlinear operator that maps from model state space onto the space of the observation vector (this includes the integration of the model from initial time to the observation times, as well as the interpolation onto the observation points). The background-error covariance matrix of the control vector is assumed to be the identity matrix ($\mathbf{B}_{(\mathbf{v})} = \mathbf{I}$) as evident by the use of the canonical inner product for the background term in (23). In other words, background errors for $\mathbf{v}^b$ are assumed to be uncorrelated and to have unit variance. There are two advantages that result from this formulation where the background term takes on a very simple form. First, it generally improves the convergence properties of the minimisation when the problem is solved with a conjugate gradient algorithm. For quadratic cost functions, this is often explained by a reduction in the condition number of the Hessian Second, all constraints in the assimilation problem are now imposed through the nonlinear observation operators $\mathcal{G}$ and $\mathcal{U}$, including multivariate and smoothness constraints that are used in conventional model-space (matrix) formulations of the background-error covariance model. In particular, this opens the way for incorporating potentially more realistic (nonlinear) multivariate balance relationships in the analysis problem. Details on techniques for constructing the transformation $\mathcal{U}$ (and its inverse $\mathcal{U}^{-1}$) can be found in Weaver et al. (2005)

Currently NEMOVAR employs a variant of the incremental algorithm for approximately minimising the non-quadratic cost function (23). The algorithm is defined by the iterative minimisation of a sequence, $k = 1, ..., K_o$, of quadratic cost functions

$$\begin{aligned} J^k[\delta\mathbf{v}^k] \;&=\; \frac{1}{2}\left[\delta\mathbf{v}^k - \mathbf{d}^{b,k-1}\right]^T \left[\delta\mathbf{v}^k - \mathbf{d}^{b,k-1}\right] \\ &+\; \frac{1}{2}\left[\mathbf{G}^{k-1}\mathbf{U}^{k-1}\delta\mathbf{v}^k - \mathbf{d}^{o,k-1}\right]^T \mathbf{R}^{-1}\left[\mathbf{G}^{k-1}\mathbf{U}^{k-1}\delta\mathbf{v}^k - \mathbf{d}^{o,k-1}\right] \end{aligned} \tag{24}$$

where

$$\mathbf{d}^{b,k-1} \;=\; \mathbf{v}^b - \mathbf{v}^{k-1}, \tag{25}$$

$$\mathbf{d}^{o,k-1} \;=\; \mathbf{y}^o - \mathcal{G}(\mathcal{U}(\mathbf{v}^{k-1})) \;=\; \mathbf{y}^o - \mathcal{G}(\mathbf{x}^{k-1}), \tag{26}$$

$\mathbf{v}^{k-1}$ is a reference state, $\delta\mathbf{v}^k$ is an increment defined by $\mathbf{v}^k = \mathbf{v}^{k-1} + \delta\mathbf{v}^k$, and $\mathbf{G}^{k-1}$ and $\mathbf{U}^{k-1}$ are linearised operators defined such that $\mathcal{G}(\mathcal{U}(\mathbf{v}^{k-1} + \delta\mathbf{v}^k)) \approx \mathcal{G}(\mathcal{U}(\mathbf{v}^{k-1})) + \mathbf{G}^{k-1}\mathbf{U}^{k-1}\delta\mathbf{v}^k$ (when this equation is satisfied exactly, (25) is identical to (23)). The superscript $k-1$ indicates that $\mathbf{G}^{k-1}$ is the result of linearising $\mathcal{G}$ about $\mathbf{v}^{k-1}$. The sequence $k = 1, ..., K_o$ are called outer iterations

[1] By interpreting $\mathbf{x}$ to be the initial conditions, the model and external forcing fields are tacitly assumed to be perfect. This assumption can be relaxed in the above formulation by considering $\mathbf{x}$ to contain model-error or external forcing terms in addition to the initial conditions.

while the minimisation iterations performed within each outer loop are called inner iterations. Equations (25) and (26) are the effective "background" and "observation" vectors for the inner loop minimisation. In practice, it is customary to set $\mathbf{v}^0 = \mathbf{v}^b$ and to choose $\mathbf{v}^{k-1}$, for $k = 2, ..., K_o$, to be the solution obtained at the end of the previous outer loop. The minimum of (25) after the K_o-th outer iteration defines the analysis increment, $\delta\mathbf{v}^a = \delta\mathbf{v}^{K_o}$.

B Software developments summary

All developments of this task have been carried out within the NEMOVAR collaborative infrastructure, which consists mostly in *git* mirror repositories hosted by its developers' institutions (Cerfacs, ECMWF, Inria and Met Office). So in particular they are accessible to ERACLIM2 partners through the ECMWF repository. Developments described below are mostly in fortran, in addition to that PIANO (Python Interface for Assimilation with NemO) the set of python scripts shipped with NEMOVAR has been significantly rewritten to handle multiresolution. It is available in the common repository, but since it is not used in the CERA system, it is not described here.

B.1 4D-Var in NEMOVAR and CERA ocean

In early stage of the ERACLIM2 project, 4D-Var capabilities have been imported in the ocean part of the CERA system, in collaboration with P. Laloyaux (ECMWF). NEMOVAR was already 4D-Var ready, so this work consisted mostly in importing the latest version of the tangent and adjoint model for NEMO (NEMOTAM, Vidard et al. (2015)) and adapting the CERA SMS scripts to handle the non linear trajectory. Later on an updated version of NEMOTAM including bugfixes and additional capabilities was installed in the CERA system.

In parallel to our developments, NEMOVAR underwent two major updates, first to make feasible CERFACS and Met Office ERACLIM2 related developments and a second one to make NEMOVAR compatible with the OOPS framework. 4D-Var capabilities has been transferred to the first update. Since making NEMOTAM compatible with OOPS is a major challenge and that NEMOVAR OOPS compatibility is not completely achieved, it has been decided to postpone 4D-Var transfer to the latest version in order to avoid duplication of effort. In order to ensure future compatibilities the simplified tangent and adjoint models described in section 3.1 is being installed in the latest NEMOVAR version.

B.2 Transfer operators

The off-line transfer operator presented in section 3.2.1 is also included in the NEMOVAR repository. Being a stand-alone software, maintaining compatibility with recent updates of NEMOVAR is not really an issue. The on-line version is currently being developed within the NEMOVAR versioning system and in phase with latest NEMOVAR developments.

References

Balmaseda, M. A., Mogensen, K. S. & Weaver, A. T. (2013), 'Evaluation of the ECMWF ocean reanalysis system ORAS4', *Q.J.R. Meteorol. Soc.* **139**(674), 1132–1161.

Debreu, L., Neveu, É., Simon, E., Le Dimet, F.-X. & Vidard, A. (2015), ‘Multigrid solvers and multigrid preconditioners for the solution of variational data assimilation problems’, *Q.J.R. Meteorol. Soc.* pp. 1–19.

Gratton, S., Lawless, A. S. & Nichols, N. K. (2007), ‘Approximate Gauss–Newton Methods for Nonlinear Least Squares Problems’, *SIAM J. Optim.* **18**(1), 106–132.

Laloyaux, P., Balmaseda, M. A., Dee, D. P., Mogensen, K. S. & Janssen, P. (2015), ‘A coupled data assimilation system for climate reanalysis’, *Q.J.R. Meteorol. Soc.* **142**(694), 65–78.

Lawless, A. S., Nichols, N. K. & Ballard, S. P. (2003), ‘A comparison of two methods for developing the linearization of a shallow-water model’, *Q.J.R. Meteorol. Soc.* **129**(589), 1237–1254.

Mulholland, D. P., Laloyaux, P. & Haines, K. (2015), ‘Origin and Impact of Initialization Shocks in Coupled Atmosphere–Ocean Forecasts*’, *Monthly Weather Review* .

Smith, P. J., Fowler, A. M. & Lawless, A. S. (2015), ‘Exploring strategies for coupled 4D-Var data assimilation using an idealised atmosphere–ocean model’, *Tellus A* **67**(0), 217–25.

Vidard, A., Bouttier, P. A. & Vigilant, F. (2015), ‘NEMOTAM: tangent and adjoint models for the ocean modelling platform NEMO’, *Geosci. Model Dev.* **8**(4), 1245–1257.

Weaver, A. T., Deltel, C., Machu, E., Ricci, S. & Daget, N. (2005), ‘A multivariate balance operator for variational ocean data assimilation’, *Q.J.R. Meteorol. Soc.* **131**(613), 3605–3625.