# Toward the S3DVAR data assimilation software for the Caspian Sea


Rossella Arcucci[1,2], Simone Celestino[1], Ralf Toumi[3] and Giuliano Laccetti[1]

[1]*University of Naples "Federico II", Italy.*
[2]*Euro Mediterranean Center on Climate Change (CMCC), Italy*
[3]*Imperial College London, UK.*



**Abstract.** Data Assimilation (DA) is an uncertainty quantification technique used to incorporate observed data into a prediction model in order to improve numerical forecasted results. The forecasting model used for producing oceanographic prediction into the Caspian Sea is the Regional Ocean Modeling System (ROMS). Here we propose the computational issues we are facing in a DA software we are developing (we named S3DVAR) which implements a Scalable Three Dimensional Variational Data Assimilation model for assimilating sea surface temperature (SST) values collected into the Caspian Sea with observations provided by the Group of High resolution sea surface temperature (GHRSST). We present the algorithmic strategies we employ and the numerical issues on data collected in two of the months which present the most significant variability in water temperature: August and March.



Keywords: Data Assimilation, oceanographic data, Sea Surface Temperature, Caspian sea, ROMS


## INTRODUCTION

Data Assimilation (DA) is an uncertainty quantification technique used to incorporate observed data into a prediction model in order to improve numerical forecasted results. Improvement in Caspian sea temperatures prediction is a crucial point for different climate phenomena simulation. An example is the study on the sea-ice coverage [1] or the prediction of the cyclonicity in winter and anticyclonicity in spring and summer as the water temperature influences the closed atmosphere [2]. This variability may be of interest in the long-term as it may act as an early indicator of large-scale climate change, as well as being an area of interest to industries and vulnerable species.
A suitable DA model must be identified taking into account both the users/applications requirements and the mathematical-numerical-algorithmic approaches. Following a problem-to-solve approach, the attention is devoted to:

1. the physical and mathematical assumptions concerning the definition/localization of the data (forecasting data and available observations);
2. the algorithmic strategies;
3. the computing environment in which the software is implemented.

The forecasting data which represent sea surface temperature (SST) values into the Caspian Sea produce are produced by using the Regional Ocean Modeling System (ROMS) [3]. The SST variabilities in the Caspian Sea have different characteristics in the different regions [4]. Caused their diversities, sometimes the studies focus on the North Caspian or South Caspian separately. This peculiarity suggests that a DA model able to opportunely assimilate data on different part of the domain indipendently could be recommended. The observations are satellite data provided by the Group of High resolution sea surface temperature (GHRSST) [5].
Due to the scale of the forecasting area used to describe the Caspian sea, DA is a large size problems then it is mandatory to develope a DA software in High Performance Computing (HPC) environment [6, 7]. Concerning the design of the algorithm to adapt to the evolutions of the node architectures foreseen at exascale, this paper looks at different algorithmic strategies, which can tackle issues related to available data (forecasted and observed data) produced by using supercomputers. As claimed in [8], problem partitioning (decomposability: to break the problem

into small enough independent less complex subproblems) is a universal source of scalable parallelism; the approach we use here meets the following demand: parallelization should be considered from the beginning [9, 10]. In this work, we employ the algorithm in [11] which splits the DA problem (let us say, the global problem) into several DA problems which reproduce the DA problem at smaller dimensions (let us say, the local problems). Finally, the testbed we consider is a distributed computing environment.

## THE S3DVAR COMPUTATIONAL KERNEL

Hereafter we provide a synthetic formalization of the DD-DA model we implemented in Algorithm 1 for assimilating the data collected into the Caspian sea, which is based on a Problem Decomposition approach [10, 9]
Let $t_k$, $k = 0, 1, \ldots, n$ be a sequence of observation times and, for each $k$, let be

$$x_k^{\mathcal{M}} \equiv x(t_k) \in \Re^N \tag{1}$$

the vector denoting the state of a sea system. At time $t_k$ it is $x_k = \mathcal{M}(x_{k-1})$ with $\mathcal{M} : \Re^N \mapsto \Re^N$ forecasting model. At each time step $t_k$, let be

$$y_k = \mathcal{H}_k(x_k) \in \Re^p \tag{2}$$

the observations vector where $\mathcal{H}_k : \Re^N \mapsto \Re^p$ is a non-linear interpolation operator collecting the observations at time $t_k$.
The aim of DA problem is to find an optimal tradeoff between the current estimate of the system state (background) defined in (1) and the available observations $y_k$ defined in (2).
Let (3) be an overlapping decomposition of the physical domain $\Omega$ such that $\Omega_i \cap \Omega_j = \Omega_{ij} \neq 0$ if $\Omega_i$ and $\Omega_j$ are adjacent and $\Omega_{ij}$ is called *overlapping region*.

$$\Omega = \bigcup_{i=1}^{N_{sub}} \Omega_i \tag{3}$$

For a fixed time $t_k = t_0$, according to this decomposition, the DD-DA computational model is a system of $N_{sub}$ non-linear least square problems described in (4)-(5) where $J_i$ in (5) is called cost-function.

$$x_0^{DA} = \sum_{i=1}^{N_{sub}} \tilde{x}_{0_i}^{DA}, \quad \text{with} \quad \tilde{x}_{0_i}^{DA} = \begin{cases} argmin_{x_0} J_i(x_{0_i}^{DA}) & on \quad \Omega_i \\ 0 & on \quad \Omega - \Omega_i \end{cases} \tag{4}$$

$$J_i(x_{0_i}^{DA}) = \|x_{0_i}^{DA} - x_{0_i}^{\mathcal{M}}\|_{B_i}^2 + \lambda_i \|\mathcal{H}_i(x_{0_i}^{DA}) - y_i\|_{R_i}^2 + \mu_i \left(x_{0_i}^{DA}/\Omega_{ij} - x_{0_j}^{DA}/\Omega_{ij}\right)^T {B_{ij}}^{-1} \left(x_{0_i}^{DA}/\Omega_{ij} - x_{0_j}^{DA}/\Omega_{ij}\right) \tag{5}$$

with $\lambda_i$ and $\mu_i$ regularization parameters [12].
$x_0^{DA}$ in (4) is the *analysis* (i.e. the estimation of the vector $x_{0_i}^{DA}$ at time $t_0$). The variables $x_{0_i}$ and $y_{k_i}$ are the same vectors $x_0$ and $y_k$ in (1) and (2) defined on the subdomain $\Omega_i$, $R_i$ and $B_i$ are the covariance matrices whose elements provide the estimate of the errors on $y_{k_i}$ and on $x_{0_i}^{\mathcal{M}}$ respectively, and $B_{ij}$ is the background error covariance matrix defined on $\Omega_{ij}$.
The minimum of the cost function $J_i$ in (5) is computed by the LBFGS method [13]. Due to the background error covariance matrix, the Hessian matrix is ill conditioned, so a preconditioning methods must be used for improving conditioning of $B_i$ [14].
Let $d_K = [y_k - \mathcal{H}_k(x_k)]$ be the *misfit*, by using the linearization of $\mathcal{H}_k$ such that $\mathcal{H}_k(x) = \mathcal{H}_k(x + \delta x) + H_k\, \delta x$, where $H_k$ is the matrix obtained by the first order approximation of the Jacobian of $\mathcal{H}_k$ and, by setting $v_i = V_i^T \delta x_i$, with $V_i$ such that $B_i = V_i V_i^T$, the *preconditioned* cost function is:

$$J_i(v_i) = \frac{1}{2} v_i^T v_i + \lambda_i \frac{1}{2}(H_i V_i v_i - d_i)^T R_i^{-1}(H_i V_i v_i - d_i) + \mu_i \frac{1}{2}(V_{ij} v_i^+ - V_{ij} v_i^-)^T (V_{ij} v_i^+ - V_{ij} v_i^-) \tag{6}$$

where $V_{ij}$ is such that $B_{ij} = V_{ij} V_{ij}^T$ and $v_i^+ = v_i$ on $\Omega_{ij}$ and $v_j^- = v_j$ on $\Omega_{ij}$.

**Algorithm 1** *the S3DVAR algorithm on each subdomain* $\Omega_i$

1: Input: $y_i$ and $x_{0_i}^{\mathcal{M}}$
2: Define $H_i$
3: Compute $d_i \leftarrow y_i - H_i x_{0_i}^{\mathcal{M}}$ % compute the misfit
4: Define $R_i$ starting from the observed data $y_i$
5: Define $V_i$ starting from a temporal sequence of hystorical data $\{x_{k_i}^{\mathcal{M}}\}_{k=0,\ldots,M}$
6: Setting of $\lambda_i$ % It balances the weigth of the observations with respect the background data
7: Setting of $\mu_i$ to join up the solutions on the boundaries
8: Define the initial value of $\delta x_i^{DA}$
9: Compute $v_i \leftarrow V_i^T \delta x_i^{DA}$
10: repeat % start of the L-BFGS steps
11: Send and Receive the boundary conditions from the adjacent domains
12: Compute $J_i \leftarrow J_i(v_i)$
13: Compute $gradJ_i \leftarrow \nabla J_i(v_i)$
14: Compute new values for $v_i$
15: until (Convergence on $v_i$ is obtained) % end of the L-BFGS steps
16: Compute $x_i^{DA} \leftarrow x_{0_i}^{\mathcal{M}} + V_i v_i$

**end**

# DISCUSSION

The SST variabilities in the Caspian Sea have different characteristics in the different regions. In the Southern Caspian, the SST reaches a high of $25-29°C$ in the summer months and has a low of $7-10°C$ in the winter. The Northern Caspian experiences a more drastic change in SST throughout the year, with a high of $25-26°C$ in the summer and a below freezing point in the winter. Here we focus on the North Caspian and South Caspian separately by considering two different subdomains

$$\Omega_{NORTH} = \{(64° < lat < 126°, 253° < lon < 275°)\}$$

and

$$\Omega_{SOUTH} = \{(18° < lat < 61°, 86° < lon < 124°)\}$$

Here we focus on the main computational issues we faced by implementing the Algorithm 1. The architecture we use for developing is a Multiple-Instruction, Multiple-Data (MIMD) architecture made of 8 nodes which consist of distributed memory DELL M600 blades connected by a 10 Gigabit Ethernet technology. Each blade consists of 2 Intel Xeon@2.33GHz quadcore processors sharing the same local 16 GB RAM memory for a total of 8 cores per blade and of 64 total cores. Here we do not provide scalability results as the computational model we are using is been already proved to be fully scalable [11].

All the routines we refer are implemented by using the Linear Algebra PACKage (LAPACK) library which provides a documentation and description of all the parameters [15].

The background data (defined in (1)) we consider are provided by the software ROMS [3]. The satellite observations (defined in (2)) provided by the GHRSST give us information about the SST every day of the selected months at 12:00am according with the data provided by ROMS. The computed values of the misfits (see Step 3 of Algorithm 1) present an order of magnitude of the errors of $O(10^{-2})$.

We computed the background error deviance matrix $V_i$ (see Step 5 of Algorithm 1) of the covariance matrix from data collected into the selected subdomains in two peculiar months: August 2008 and March 2008 [4]. The preconditioning approach for $V_i$ we used is the EOFs method [16] which is based on a Truncated Singular Value Decomposition (TSVD) of the matrix. We studied the spectrum of the matrix $V_i$, then we fixed 20 EOFs.

The chosen starting point for assimilating data is been fixed as the first of August and the first of March respectively for both subdomains.

As the DA problem is an inverse ill posed problem [17, 18, 19], a very important topic is the choice of the regularization parameters in (5) then in (6) (see Step 6 and Step 7 of Algorithm 6). Results we carried out show as the solution of the S3DVAR software depends on these parameters in terms of both accuracy (e.g. values of the misfits) and efficiency (e.g. number of L-BFGS steps). The computed values of the misfits after the DA present an improvement into the order of magnitude of the errors which is $O(10^{-3})$. The results show that the number of LBFGS steps decrease as the values of the regularization parameters decrease. For example, the number of LBFGS steps is $n_{iter} = 5$ for values of $1 < \lambda < 0.5$ and it decreases for values of $\lambda < 0.125$ and $\mu = 0$ which imply no interaction among the subdomains. Actually we are working on the optimal parametes tuning which balance accuracy and efficiency results.

## ACKNOWLEDGMENTS

This work was developed within the research activity of the H2020-MSCA-RISE-2016 NASDAC Project N. 691184. The computing architectures are located at the University of Naples Federico II, Naples, Italy.

## REFERENCES


[1] H. Tamura-Wicks, R. Toumi, and W. P. Budgell, *Sensitivity of Caspian sea-ice to air temperature.* (Quarterly Journal of the Royal Meteorological Society, Soc. 141., 2015), pp. 3088–3096.

[2] J. F. Nicholls and R. Toumi, *On the lake effects of the Caspian Sea.* (Quarterly Journal of the Royal Meteorological Society, Soc. 140., 2014), pp. 1399–1408.

[3] ROMS, *Web page: www.myroms.org.*

[4] R. Ibrayev, E. Ozsoy, C. Schrum, and H. Sur, *Seasonal variability of the caspian sea three-dimensional circulation, sea level and air-sea interaction.* (Ocean Science Discussions 6, 2009), pp. 1913–1970.

[5] G. of High resolution sea surface temperature (GHRSST), *Web page: www.ghrsst.org.*

[6] L. D'Amore, R. Arcucci, L. Marcellino, and A. Murli, *HPC computation issues of the incremental 3D variational data assimilation scheme in OceanVar software* (Journal of Numerical Analysis, Industrial and Applied Mathematics, vol.7, 2013), pp. 91–105.

[7] L. D'Amore, R. Arcucci, L. Marcellino, and A. Murli, *A parallel three-dimensional variational data assimilation scheme* (AIP Conference Proceedings 1389, 2011), pp. 1829–1831.

[8] G. Fox, R. Williams, and P. Messina, *P.C.: Parallel. Computing Works!* (Morgan Kaufmann Publishers Inc., Los Altos, CA, 1994).

[9] R. Arcucci, L. D'Amore, and L. Carracciuolo, *On the Problem Decomposition of Scalable 4D-Var Data Assimilation Models* (HPCS-IEEE, 978-1-4673-7812-3, 2015), pp. 589–594.

[10] L. D'Amore, R. Arcucci, L. Carracciuolo, and A. Murli, *DD-OceanVar: a Domain Decomposition fully parallel Data Assimilation software in Mediterranean Sea* (Procedia Computer Science 18, 2013), pp. 1235–1244.

[11] L. D'Amore, R. Arcucci, L. Carracciuolo, and A. Murli, *A Scalabale Variational Data Assimilation* (Journal of Scientific Computing, vol. 61, 2014), pp. 239–257.

[12] R. Arcucci, L. D'Amore, and L. Carracciuolo, *A scalable numerical algorithm for solving Tikhonov regularization problems* (Lecture Notes in Computer Science, vol. 9574, 2016), pp. 45–54.

[13] J. Nocedal, R. Byrd, P. Lu, and C. Zhu, *L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimizatio* (ACM Transactions on Mathematical Software, Vol. 23, No. 4, 1997), pp. 550–560.

[14] N. Nichols, *Mathematical Concepts in Data Assimilation* (W. Lahoz et al. (eds), Data Assimilation, Springe, 2010).

[15] LAPACK, *Web page: www.netlib.org/lapack/.*

[16] E. N. Lorenz, *Empirical orthogonal functions and statistical weather prediction.* (Sci.Rep. No. 1, Statistical Forecasting Project, M.I.T., Cambridge, MA, 1956).

[17] L. D'Amore, L. Marcellino, and A. Murli, *Image sequence inpainting: Towards numerical software for detection and removal of local missing data via motion estimation.* (Journal of Computational and Applied Mathematics Vol 198, Issue 2, 2007), pp. 84–98.

[18] R. Campagna, L. D'Amore, and A. Murli, *An efficient algorithm for regularization of Laplace transform inversion in real case.* (Journal of Computational and Applied Mathematics Vol 210, Issue 1-2, 2009), pp. 1913–1970.

[19] L. D'Amore, R. Campagna, A. Galletti, L. Marcellino, and A. Murli, *A smoothing spline that approximates Laplace transform functions only known on measurements on the real axis.* (Journal of Computational and Applied Mathematics Vol 28, Issue 2, 2012), pp. 396–413.