Preconditioning of the background error covariance matrix in data assimilation for the Caspian Sea

Rossella Arcucci[1,2], Luisa D'Amore[1,2] and Ralf Toumi[3]

[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 for improving numerical forecasted results by incorporating observed data into prediction models. As a crucial point into DA models is the ill conditioning of the covariance matrices involved, it is mandatory to introduce, in a DA software, preconditioning methods. Here we present first studies concerning the introduction of two different preconditioning methods 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 by using the Regional Ocean Modeling System (ROMS) with observations provided by the Group of High resolution sea surface temperature (GHRSST). We also present the algorithmic strategies we employ.

Keywords: Data Assimilation, ill conditioning, 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 are 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.
The forecasting data which represents sea surface temperature (SST) values into the Caspian Sea 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 within [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. To this aim, we employ the algorithm in [8] 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). The DA is an ill posed inverse problem [9, 10, 11]. The main computational kernel of the DA local problems is the solution of a linear system. Caused by the background error covariance matrices this systems are ill conditioned [12]. In designing data assimilation schemes, it is important, therefore, to implement preconditioning techniques applied to the background error covariance matrices in order to reduce the ill conditioning. Here we employ two different methods:

I. the Empirical Orthogonal Functions (EOFs) method [13] which implement a Truncated Singular Value Decomposition (TSVD) of the background error covariance matrix.

II. the Tikhonov regularization [14] method of the background error covariance matrix as well.

THE S-CASVAR COMPUTATIONAL KERNEL

Hereafter we provide a synthetic formalization of the DD-DA model we implemented in Algoritm 1 for assimilating the data collected into the Caspian sea, which is based on a Problem Decomposition approach [15, 16]
Let t_k, $k = 0, 1, \ldots, n$ be a sequence of observation times and, for each k, let be

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

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 Ω such that $\Omega_i \cap \Omega_j = \Omega_{ij} \neq 0$ if Ω_i and Ω_j are adjacent and Ω_{ij} is called *o*verlapping 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 sistem of N_{sub} non-linear least square problems described in (4)-(5) where J_i in (5) is called cost-function.

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

$$J_i(x_{0_i}^{DA}) = \|\mathbf{x}_{0_i}^{DA} - \mathbf{x}_{0_i}^{M}\|_{\mathbf{B_i}}^2 + \lambda \|\mathcal{H}_\mathbf{i}(\mathbf{x}_{0_i}^{DA}) - \mathbf{y}_i\|_{\mathbf{R_i}}^2 + \mu \left(\mathbf{x}_{0_i}^{DA}/\Omega_{ij} - \mathbf{x}_{0_j}^{DA}/\Omega_{ij}\right)^T \mathbf{V_{ij}}^{-1} \left(\mathbf{x}_{0_i}^{DA}/\Omega_{ij} - \mathbf{x}_{0_j}^{DA}/\Omega_{ij}\right) \tag{5}$$

with λ and μ regularization parameters.
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 $\mathbf{x}_{0_i}$ and $\mathbf{y}_{k_i}$ are the same vectors $\mathbf{x}_0$ and $\mathbf{y}_k$ in (1) and (2) defined on the subdomain Ω_i, $\mathbf{R_i}$ and $\mathbf{B_i}$ are the covariance matrices whose elements provide the estimate of the errors on y_{k_i} and on x_{0_i}, respectively.
The minimum of the cost function J_i in (5) is computed by the LBFGS method [17]. Due to the background error covariance matrix, the Hessian matrix is ill conditioned with an ill determined numerical rank, so a preconditioning methods must be used for improving conditioning of $\mathbf{B_i}$ [12].
Let $d = [y_k - \mathcal{H}(x_k)]$ be the *misfit*, by using the linearization of $\mathcal{H}$ such that $\mathcal{H}(x) = \mathcal{H}(x + \delta x) + H\,\delta x$, where H is the matrix obtained by the first order approximation of the Jacobian of $\mathcal{H}$ and, by setting $v_i = V_i^T \delta x_i$, with V_i such that $\mathbf{B_i} = V_i V_i^T$, the *preconditioned* cost function is [18]:

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

The matrix V_i (see Step 5 of the Algoritm 1) is computed by the subroutine described in Algorithm 2 which implements two preconditioning approaches: the Empirical Orthogonal Functions (EOFs) method and the Tikhonov regularization method. Both methods are based on the singular value decomposition of the error covariance matrix. 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 [19].

Algorithm 1 *the S-CASVAR algorithm on each subdomain* Ω_i

1: Input: y_i and $x_{0_i}^M$
2: Define H_i

3: Compute $d_i \leftarrow y_i - H_i x_{0_i}^M$ % compute the misfit
4: Define R_i
5: Compute $V_i = PECM(''P'', ind, \{x_{k_i}^M\}_{k=1,\dots,m})$ % See Algorithm 2 for details
6: Define the initial value of x_i^{DA}
7: Compute $v_i \leftarrow V_i^T x_i^{DA}$
8: repeat % start of the L-BFGS steps
9: Send and Receive the boundary conditions from the adjacent domains
10: Compute $J_i \leftarrow J_i(v_i)$
11: Compute $gradJ_i \leftarrow \nabla J_i(v_i)$
12: Compute new values for v_i
13: until (Convergence on v_i is obtained) % end of the L-BFGS steps
14: Compute $x_i^{DA} \leftarrow x_{0_i}^M + V_i v_i$

end

Algorithm 2 *the Preconditioning Error Covariance Matrix (PECM) algorithm*

1: Input: $''P''$, ind and $\{x_{k_i}^M\}_{k=1,\dots,m}$
2: Compute $\bar{x}_i \leftarrow mean_k(x_{k_i}^M)$
3: Compute $V_i^k \leftarrow x_{k_i}^M - \bar{x}_i$
4: if $P = EOFS$ then
5: Compute $V_i = TSVD(ind, V_i^k)$ % ind correspond to the TSVD truncation parameter
6: else
7: Compute $V_i = Tikhonov(ind, V_i^k)$ % ind correspond to the Tikhonov regularization parameter
8: endif

end

DISCUSSION

The SST variabilities in the Caspian Sea have different characteristics in the different regions within: 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:

$$\Omega_{NORTH} = \{(64° < lat < 126°, 253° < lon < 275°)\}$$
$$\Omega_{SOUTH} = \{(18° < lat < 61°, 86° < lon < 124°)\}$$

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 [8]. 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. We computed the background error deviance matrix V_i (see Step 5 of Algorithm 1 which is detailed in Algorithm 2) of the covariance matrix from data collected into the selected subdomains in two peculiar months: August 2008 and March 2008 [4]. We estimated the condition number of these matrices which is $\mu(V_i) = O(10^3)$ with a degree of ill posedness is $k = 2$ [14]. We employ as preconditioning methods the EOFs method and the Tikhonov regularization method. The computed condition number of the preconditioned matrices V_i^{EOFs} and $V_i^{Tikhonov}$ is reduced with respect the matrix V_i of two and three orders of magnitude respectively. Experiments reveal a bigger improvement in terms of condition number in employing the Tikhonov regularization method with respect the EOFs method:

$$\frac{\mu(V_i^{EOFs})}{\mu(V_i^{Tikhonov})} \in [10^1, 10^2],$$

Then the Tikhonov regularization method is for the Caspian sea data, is more appropriate than the truncation of the EOFs method. Actually we are validating these approaches into the S3DVAR software we are developing.

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] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems, numerical aspects of linear inversion* (SIAM, 1998).

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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