# The General Data Assimilation Method, its Comparison with the Standard Scheme, and its Application to Dynamical Simulation in the Atlantic

*Konstantin* Belyaev[1,2], *Andrey* Kuleshov[2,*], *Ilya* Smirnov[3], and *Clemente A.S.* Tanajura[4]

[1]Shirshov Institute of Oceanology of Russian Academy of Sciences, Moscow, Russia
[2]Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia
[3]Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
[4] Federal University of Bahia, Salvador, Brazil

**Abstract.** A new data assimilation scheme developed earlier and based on the theory of diffusion stochastic processes and parabolic differential equations is presented and tested. This scheme is applied to the Hybrid Circulation Ocean Model (HYCOM) and altimetry data base Archiving, Validating and Interpolating Satellite Oceanography Data (AVISO) over the Atlantic. Several numerical experiments are conducted and their results are analyzed. It is shown that the method really assimilates data, makes the output oceanic fields closer to observations and, on the other hand, conserves the model integrals and balance. The tested method is also compared with the Ensemble Optimal Interpolation scheme (EnOI) as a counterpart of the standard Kalman filter method and it is shown that the proposed general method has several advantages, in particular, it provides a better forecast and requires less computational consumptions.

## 1 Introduction

Nowadays, data assimilation as a part of mathematical and numerical research is a scientific area of great practical importance that is used in ocean modelling, weather forecast, operational oceanography and many other fields of science. Its main goal is to combine the numerically computed model results and independently observed data as a part of the model phase space in the optimal way. Since the extreme solutions of this problem, namely, ignoring observations at all or replacing only the observed part of model computed variables by observations with no changing all the others, are obviously poor, the optimal solution of these scientific problems is not trivial.

The corresponding studies have been carried out and published in the scientific literature for more than 50 years, since the beginning of the 1960s. A good review of the main achievements in this direction for the last century is presented in [1]. From the beginning of the 2000s, the main progress is related to the development of computer facilities, explosion in observational data network, parallel computations and other technical novelties. This advance leads to the progress in new mathematical methods and algorithms, the construction and development of numerical models of very high resolution, the data exchange all over the world, etc. At the present time, the data assimilation techniques, algorithms and methods have become the essential part of operational oceanography on the ocean shelf and coastal zones, especially in the oil and gas mining zones, as well as in the zones of pipeline transportation. Several national and international scientific projects are specially aimed to find the optimal solution of the data assimilation techniques in conjunction with the regionally configured numerical models. In particular, we can mention the Brazilian REMO project [2], Australian Blue Link project [3], American HYCOM&NCODA project [4] and others.

However, the development and application of new and more advanced data assimilation schemes and methods remains an actual and very important theoretical and practical problem. There is a necessity to have a powerful and, at the same time, relatively portable and economy data assimilation scheme which would be applicable to various numerical ocean and coupled ocean-atmosphere models and would provide a satisfactory and reliable forecast of the ocean characteristics in short and media-term periods. For this reason, many papers are dedicated to the developments and applications of statistical, dynamical or hybrid assimilation methods in the recent years, for instance [5-7].

This study deals with the application of the novel data assimilation method created in [8, 9], hereafter it will be referred to as GKF. Earlier, in [8], this method was used together with the coupled Max Plank Institute Earth System model (MPIESM) [10] (Hamburg, Germany). Here, this method is used in conjunction with the ocean model HYCOM, presented in [11], where we compared

* Corresponding author: andrew_kuleshov@mail.ru

it with the standard Ensemble Optimal Interpolation (EnOI) method [12] as an alternative data assimilation scheme. The twin experiments have been conducted for the same initial conditions and with the assimilation of the equivalent data. The AVISO archive (www.aviso.org) was chosen as the input data for the assimilation. In particular, it was shown that the proposed data assimilation scheme has many advantages in comparison with its EnOI counterpart, including less computational consumptions and better forecast properties.

## 2 The data assimilation method and the numerical algorithm

The basic equations of the scheme are the following:

$$X_{a,n+1} = X_{b,n+1} + K_{n+1}(Y_{n+1} - HX_{b,n+1}), \quad (1)$$

$$K_{n+1} = \sigma_{n+1}^{-1}(\Lambda_{n+1} - C_{n+1})(H\Lambda_{n+1})^T Q_{n+1}^{-1}, \quad (2)$$

$$\sigma_{n+1} = (H\Lambda_{n+1})^T Q_{n+1}{}^{-1}(H\Lambda_{n+1}), \quad (3)$$

where $X_{a,n}, X_{b,n}$ are the vectors of the model state at a calculated moment of time $t_n$, $n$=0, 1,… after and before assimilation, respectively, i.e., the analysis and background, with a dimension $r$, where $r$ is the number of the mesh points multiplied by the number of the model variables; $Y_n$ is the observation vector at the same moment of time, which has a dimension $m$, where $m$ is the number of observation points, multiplied by the number of independently observed variables; it is assumed that $X_{a,0} = X_{b,0} = X_0$ is the known initial condition; $K(r\times m)$ is the gain matrix (analogous to the Kalman gain matrix); $\Lambda_n, C_n$ are the model and observational trends, respectively, for one time step (time-derivative), defined by the formulae: $\Lambda_n = (X_{b,n+1} - X_{a,n})/\Delta t$, $C_n = E(Y_{n+1} - Y_n)/\Delta t$, $\Delta t = t_{n+1} - t_n$. For simplicity, it is assumed that all time moments are equidistant, but it is not necessary. Symbol $E$ stands for the mathematical expectation or ensemble average; $Q_n$ denotes the covariance matrix of the model errors at the moment $n$ with a dimension $m\times m$, i.e. $Q_n = E(Y_{n+1} - HX_{a,n})(Y_{n+1} - HX_{a,n})^T$. Finally, $H$ is the observational projection matrix with a dimension $m\times r$. As usual, the superscript $T$ denotes the transpose of a vector and/or a matrix.

This scheme with all necessary and sufficient conditions was introduced in [8]. In [9], it was shown that this scheme generalizes the standard Kalman scheme method which follows from (1)-(3), if $C_n = 0$ and $X_{a,n}$ coincides with the ensemble average.

As is seen from (1), this algorithm can be applied to arbitrary numerical model with any physically reasonable initially conditions and it gives the output result (analysis) at any moment $t_n$, $n = 1,2\ldots$. To provide its correct usage, it is sufficient to set up two parameters which are supposed to be known, namely, the observational drift vector $C_n$ and the error covariance matrix $Q_n$. To set up the vector $C_n$, it is sufficient to take the consecutive analysis fields $X_{a,l}$, $l = 0,1,...,n$ and create this vector as $C_{n+1} = X_{a,n} - n^{-1}\sum_{l=0}^{n} X_{a,l}$. In [9], it is proved that if the condition $E(Y_{n+1} - HX_{a,n}) = 0$ holds, than this construction of drift vector $C_n$ really estimates the observational trend $C_n = E(Y_{n+1} - Y_n)/\Delta t$. To obtain the error covariance matrix, it is necessary to set up the ensemble of the model outputs of the observed vector at time step $t_{n+1}$, $Y_{n+1,l}, l = 1,..., N_e$, such as $N_e^{-1}\sum_{l=1}^{N_e}(Y_{n+1,l} - HX_{a,n}) = 0$, where $N_e$ is the number of ensemble members, and calculate the covariance, according to the formula

$$Q_{n,ij} = N_e^{-1}\sum_{l=1}^{N_e}(Y_{l,n+1} - HX_{a,n})(Y_{l,n+1} - HX_{a,n})^T.$$

This ensemble can be set up from the previous run(s) of the model.

## 3 Computational experiments and their results

There are several numerical experiments that have been carried out with the AVISO data and HYCOM model. The model [11] has been configured as follows: its version 2.2.14 has a spatial resolution of approximately 0.25 degree in the Atlantic in both West-East (OX axis) and South-North (OY axis) directions and 21 density layers of equal density from top to bottom (OZ axis). The model computes 4 barotropic variables (sea level, two velocity components and barotropic pressure on the sea surface) and 105 baroclinic components, namely, 5 variables on each given density layer: temperature, salinity, two velocity components, and layer thickness. After assimilation all of those variables change with respect to (1)-(3). The satellite data and model domain are shown in Fig.1.

For the comparison with the GKF, the standard EnOI scheme with the same data and the same model outputs used to create the ensemble was applied. For EnOI, it is necessary to set up the ensemble from the previous model output(s) not only for the observed component $Y$, but also for the model computed variables $X$. Let this ensemble be $X_l, l = 1,2,...,N_e$ and its average be $\bar{X} = N_e^{-1}\sum_{l=1}^{N_e} X_l$. Once it is done, the analysis is obtained according to the formulae

$$X_{a,n+1} = X_{b,n+1} + K_{n+1}(Y_{n+1} - HX_{b,n+1}),$$

$$K = \alpha BH^T(HBH^T + R)^{-1},$$

$$B = N_e^{-1}\sum_{l=1}^{N_e}(X^l - \bar{X})(X^l - \bar{X})^T,$$

where we used the same notations as in formulae (1)-(3). The instrumental error covariance matrix $R$ and the empirical scalar α are defined "manually" from heuristic considerations.

Figure 2 shows the model output of the sea level after 10 days of consecutive assimilation with both methods and control run, i.e. the model run without assimilation. It is seen that the model results obtained with using the GKF scheme (Fig. 2a) are similar to the control results (Fig. 2c), however, showing more pronounced and intensive dynamics, especially in Golf Stream zone and in Brazil-Maldivian confluence zone where the amplitude of the sea level can reach one meter. On the contrary, the results obtained with using the EnOI scheme (Fig. 2b) distinguish substantially and produces less pronounced dynamics with smaller amplitudes not more than 0.3 m.

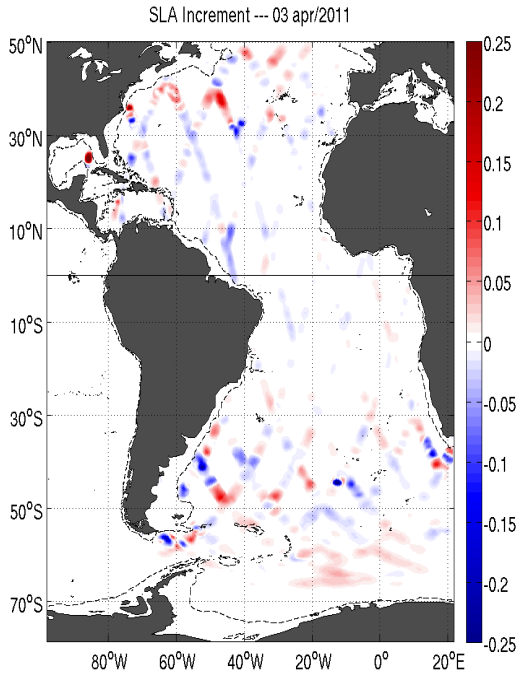
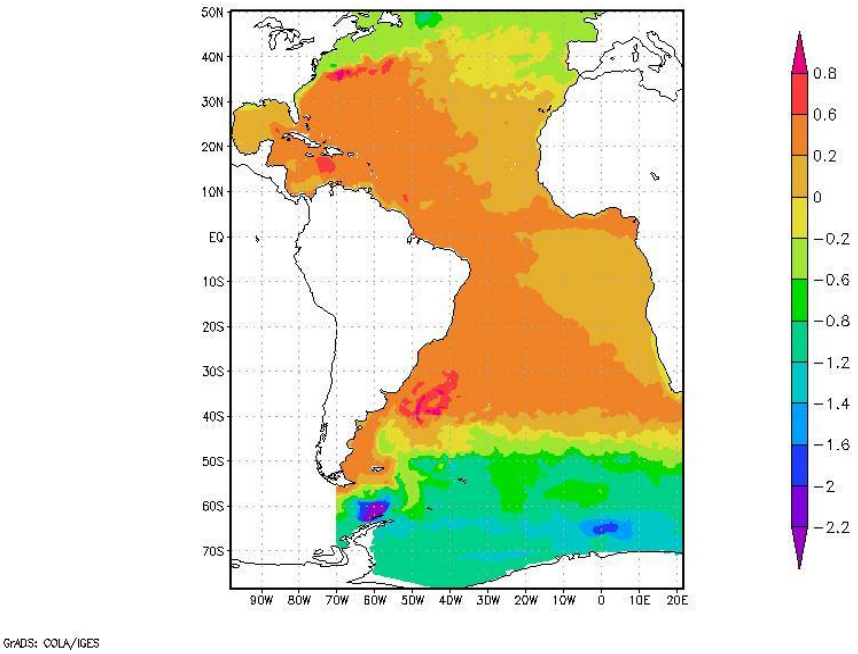
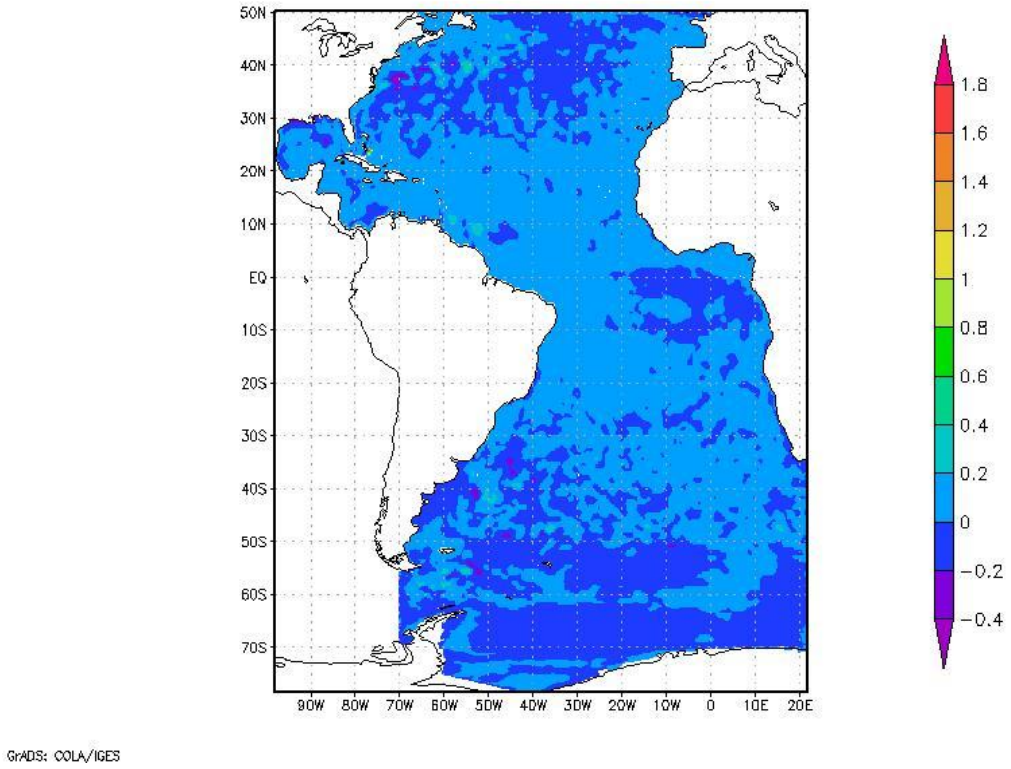



**Fig. 1.** Model domain and satellite tracks over the Atlantic. Red points show where data value exceeds the model one; blue points show the opposite.

(a)



(b)




(c)

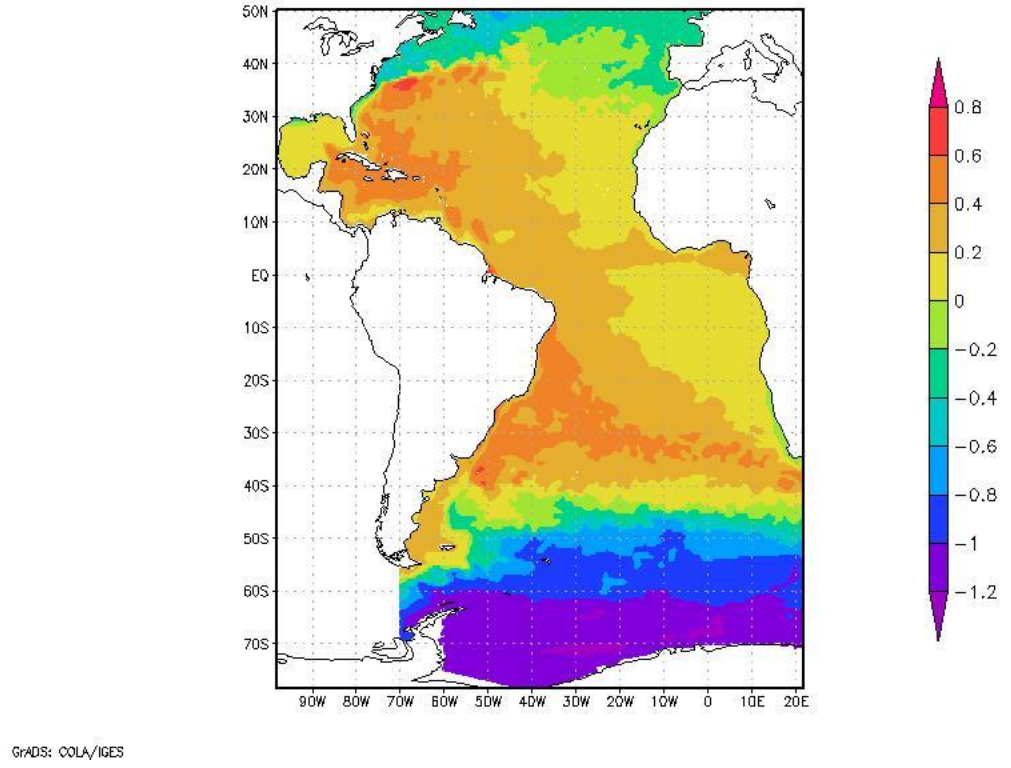



**Fig. 2.** Model of the sea level after 10 days of integration: (a) GKF–method, (b) EnOI–method, (c) – control.

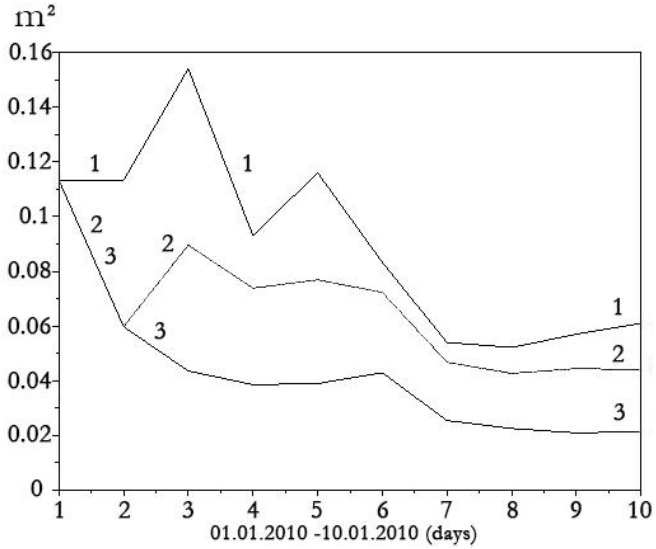



**Fig. 3.** 24h forecast error variance for 3 model runs. Line 1 is the model control error, line 2 is the error of the EnOI–method, line 3 is the error of the GKF–method.

However, Fig. 3 shows that both assimilation methods and model itself work correctly and produce a reasonable forecast as compared with the independent data taken 24 hours later. It is readily seen from Fig.3 that the model run itself approaches to observations because of atmosphere external forcing. Since the atmosphere is real, the model converges to the observed sea level even without assimilation. However, when using the data assimilation as an additional forcing, the convergence to the real independent data proves to be better and faster. The EnOI scheme gives this convergence from 0.06 to 0.04 $m^2$ in ten days run, while the GKF provides the convergence from 0.06 to 0.02 $m^2$ for the same time period.

## 4 Conclusions

The presented data assimilation method really assimilates data and provides a better 24h forecast than the alternative EnOI scheme does. In addition, it constructs the physically reasonable model fields and better follows the synoptic variability than the model itself and produces more intensive and pronounced dynamics than the EnOI scheme. It is possible to state that this scheme can be used in operational regime since it is computationally feasible, physically consistent and reliable.

This research was supported by Russian Science Foundation, project no. 14-11-00434.

## References

1. M. Ghil, P. Malnotte-Rizzoli, Adv. Geophys., **33**, 141-266 (1991).
2. M.O. Lima, M. Cirano, M.M. Mata et al., Ocean Dynamics, **66**, 1-12 (2016)
3. A. Schiller, G.B. Brassington (Eds.), *Operational Oceanography in the 21st Century* (Springer, 2011)
4. J.A. Cummings, O.M. Smedstad, *In Data Assimilation for Atmospheric, Oceanic and Hydrolo gic Applications, Vol. II* (Springer-Verlag, Berlin, Heidelberg, 303-343, 2013)
5. K. Belyaev, C.A.S. Tanajura, J.J. O'Brien, J. Math. Sci., **99** (4) 1393-1402 (2000)
6. C.A.S. Tanajura, K. Belyaev, Appl. Math. Model., **33** (5), 2165-2174 (2009)
7. K. Belyaev, C.A.S. Tanajura, N. Tuchkova, Oceanology, **52**, 523-615 (2012)
8. K. Belyaev, A. Kuleshov, N. Tuchkova, *MATEC Web of Conferences*, **76**, 05003 (2016)
9. K. Belyaev, A. Kuleshov, N. Tuchkova, C.A.S. Tanajura, Math. and Comp. Modelling of Dynamical Systems, **24** (1), 12-25 (2018)
10. H. Haak, *Simulation of Low-Frequency Climate Variability in the North Atlantic Ocean and the Arctic*, *V.1* (Max Planck Institute for Meteorology, 2004)
11. R. Bleck, Ocean Model., **4**, 55-88 (2002)
12. G. Evensen. *Data Assimilation*, *The Ensemble Kalman Filter*, *2nd ed*. (Springer, Berlin, 2009)