Méthode d'ensemble pour l'estimation des variances d'erreur d'ébauche dans un système 3D-Var appliqué à l'océan global

A. Weaver[1], N. Daget[1] et M. A. Balmaseda[2]

[1] *CERFACS, Toulouse*
[2] *ECMWF, Reading*

Outline

Outline

General context

- This study employs a variational data assimilation system (OPAVAR) developed at CERFACS for climate research applications.
- The system has been applied in the European projects ENACT and ENSEMBLES:
  - to provide initial conditions for climate hindcasts/forecasts with coupled ocean-atmosphere models;
  - to reconstruct the 1960-present history of the ocean state ("reanalysis").
- In ENSEMBLES, a 3D-Var version of OPAVAR was used to produce a 9-member ensemble of multi-decadal ocean analyses:
  - produced using perturbed atmospheric surface fluxes;
  - used to sample uncertainty in ocean initial conditions for probabilistic forecasts on seasonal to deceadal time-scales.

Objective of this study

- An important feature of an ensemble data assimilation system is its capacity to provide flow-dependent information on analysis and background error.
- This information can be exploited to improve the estimate of the background-error covariance matrix (**B**) on each assimilation cycle.
  - In the ENSEMBLES experiments, we made no attempt to use the ensemble to update **B**.
- The objective here is explore the possibility of using the ENSEMBLES data assimilation system to improve **B**.

How to use the ensemble information?

- Construct a low-rank approximation to **B** directly from the sample covariance of the ensemble of model forecast states.
  (Houtekammer and Mitchell 2001; Keppenne and Reinecker 2002; Ott *et al.* 2004; Buehner and Charron 2007; Oke *et al.* 2007).
  - Covariance localization is necessary to minimize spurious effects due to sampling error.

- or -

- Use the ensemble indirectly to calibrate parameters of a (localized) covariance model in a full-rank (operator) representation of **B**.
  (Fisher 2003; Žagar *et al.* 2005; Belo Pereira and Berre 2006; Berre *et al.* 2006; Küçükkaraca and Fisher 2006).
  - A flexible covariance model (inhomogeneous, anisotropic) is required to make best use of the ensemble information.
  - Full-rank covariance matrix representations are important for fitting detailed structures in the observations.

Objective of this study

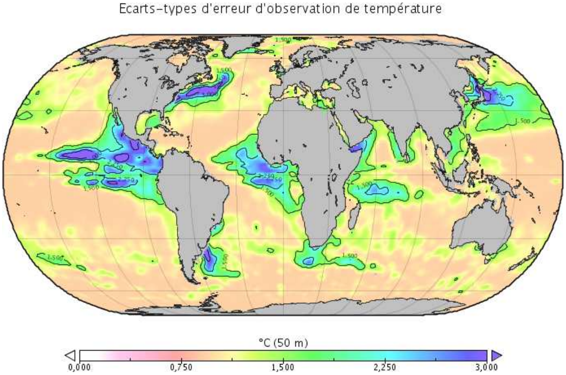
- In this study, we adopt the covariance model approach.
- In particular, we investigate the potential of an ensemble of ocean states to provide useful flow-dependent estimates of the background-error **variances** in the OPAVAR 3D-Var system.
- This approach will be compared with a simpler approach for incorporating flow dependence in the variances.
- This study is a first step towards making more comprehensive use of an ensemble for calibrating additional parameters of the covariance model.

Outline

Basic features of the assimilation system

- The ocean model is the global ORCA2 configuration of OPA8.2 (Madec *et al.* 1998).
- The surface forcing fields are derived from ERA40 (Uppala *et al.* 2005).
- The assimilation method is a multivariate 3D-Var version of OPAVAR (Weaver *et al.* 2005).
- First-Guess at Appropriate Time (FGAT) and Incremental Analysis Updates (IAU) are employed.
- The data are quality-controlled temperature and salinity profiles from EN2v1 data-base (Ingleby and Huddleston 2007).
- The observation-error variances are estimated from a model-data comparison prior to assimilation (the Fu *et al.* (1993) method).

Example of temperature σ^o at 50 m



$$J[\delta \mathbf{w}] \;=\; \frac{1}{2}\delta\mathbf{w}^{\mathrm{T}}\,\mathbf{B}_{(\mathbf{w})}^{-1}\,\delta\mathbf{w} + \frac{1}{2}(\mathbf{H}\delta\mathbf{w} - \mathbf{d})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{w} - \mathbf{d})$$

where

$$\mathbf{d} = \begin{pmatrix} \mathbf{d}_0 \\ \vdots \\ \mathbf{d}_i \\ \vdots \\ \mathbf{d}_N \end{pmatrix} = \begin{pmatrix} \mathbf{y}_0^o - \mathbf{H}_0\mathbf{w}^b(t_0) \\ \vdots \\ \mathbf{y}_i^o - \mathbf{H}_i\mathbf{w}^b(t_i) \\ \vdots \\ \mathbf{y}_N^o - \mathbf{H}_N\mathbf{w}^b(t_N) \end{pmatrix} \quad \text{and} \quad \mathbf{H} = \begin{pmatrix} \mathbf{H}_0 \\ \vdots \\ \mathbf{H}_i \\ \vdots \\ \mathbf{H}_N \end{pmatrix}.$$

- $\delta\mathbf{w} = (\delta T, \delta S)^{\mathrm{T}}$ is the vector of temperature and salinity increments.
- $\mathbf{y}_i^o = (T_i^o, S_i^o)^{\mathrm{T}}$ is the vector of temperature and salinity observations.
- Increments for sea-surface height and velocity are obtained *a posteriori* using balance constraints applied to the analysis increment $\delta\mathbf{w}^a$.

$$\mathbf{B}_{(\mathbf{w})} = \mathbf{K}_{(\mathbf{w})}\, \mathbf{D}^{1/2}_{(\widehat{\mathbf{w}})}\, \mathbf{F}_{(\widehat{\mathbf{w}})} \mathbf{F}^{\mathrm{T}}_{(\widehat{\mathbf{w}})}\, \mathbf{D}^{1/2}_{(\widehat{\mathbf{w}})} \mathbf{K}^{\mathrm{T}}_{(\mathbf{w})}$$

where

$$\mathbf{F}_{(\widehat{\mathbf{w}})} = \begin{pmatrix} \mathbf{F}_{TT} & 0 \\ 0 & \mathbf{F}_{S_U S_U} \end{pmatrix}, \quad \mathbf{D}^{1/2}_{(\widehat{\mathbf{w}})} = \begin{pmatrix} \mathbf{D}^{1/2}_{T} & 0 \\ 0 & \mathbf{D}^{1/2}_{S_U} \end{pmatrix}, \quad \mathbf{K}_{(\mathbf{w})} = \begin{pmatrix} \mathbf{I} & 0 \\ \mathbf{K}_{ST} & \mathbf{I} \end{pmatrix}$$

- $\widehat{\mathbf{w}} = (T, S_U)^{\mathrm{T}}$ where S_U corresponds to "unbalanced" salinity.
- $\mathbf{K}_{(\mathbf{w})}$ is a multivariate balance operator: $\widehat{\mathbf{w}} \mapsto \mathbf{w}$.
- $\mathbf{F}_{(\widehat{\mathbf{w}})} \mathbf{F}^{T}_{(\widehat{\mathbf{w}})}$ is a quasi-Gaussian 3D univariate correlation operator, modelled using a diffusion operator (see Mirouze and Weaver poster).
- $\mathbf{D}_{(\widehat{\mathbf{w}})}$ is a variance matrix (for $\widehat{\mathbf{w}}$), whose estimation is the focus of this study.

Outline

Reference formulation: σ^b parameterized using $\mathbf{x}^b$

For temperature:

$$\sigma_T^b = \begin{cases} \max\left(\tilde{\sigma}_T^b,\ \sigma_T^{ml}\right) & \text{in the mixed layer,} \\ \max\left(\tilde{\sigma}_T^b,\ \sigma_T^{do}\right) & \text{below the mixed layer,} \end{cases}$$

where

$$\tilde{\sigma}_T^b = \min\left(\left|\left(\partial T/\partial z|_{T=T^b}\right)\ \delta z\right|,\ \sigma_T^{max}\right).$$

For unbalanced salinity:

$$\sigma_{S_U}^b = \begin{cases} \sigma_{S_U}^{max} & \text{if } z \geq z_{max} \\ \sigma_{S_U}^{max}\,\alpha(z) & \text{if } z < z_{max} \end{cases}$$

where z_{max} is the depth of maximum $\left|\left(\partial S/\partial T|_{T=T^b}\right)\right|$ and

$$\alpha(z) = 0.1 + 0.45 \times \left\{1 - \tanh\left(2 \ln\left(z/z_{max}\right)\right)\right\}.$$

Advantages:

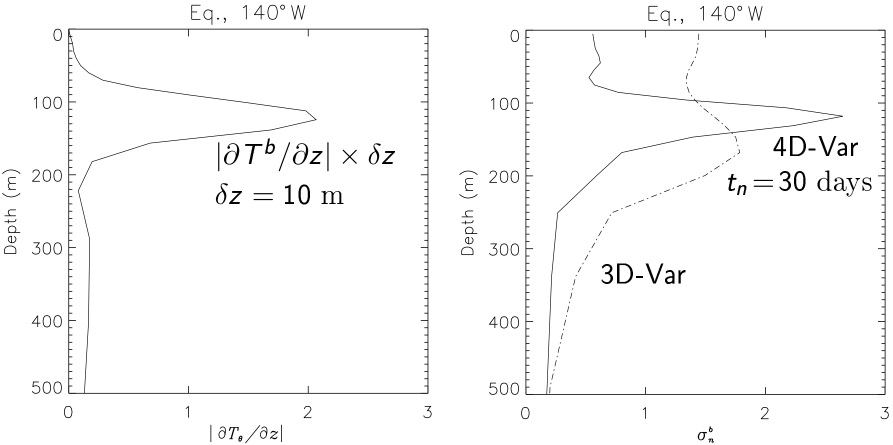
- A simple and inexpensive way to incorporate flow dependence in σ^b.
- Captures some of the dynamical effects implicit in 4D-Var (Weaver *et al.* 2003).

Disadvantages:

- Somewhat adhoc (especially for $\sigma^b_{S_U}$), and requires careful tuning of several parameters (δz, σ_T^{ml}, σ_T^{do}, σ_T^{max}, $\sigma_{S_U}^{max}$).
- Does not account for changes in the accuracy of σ^b resulting from data assimilation on previous assimilation cycles (i.e., σ^b is independent of the observing system).

Implicitly-evolved temperature σ^b in 4D-Var

(from Weaver *et al.* 2003)



$\mathbf{P}^b(t_n) = \mathbf{B}$ in 3D-Var FGAT

$\mathbf{P}^b(t_n) = \mathbf{M}(t_0, t_n)\,\mathbf{B}\,\mathbf{M}(t_0, t_n)^{\mathrm{T}}$ in 4D-Var (cf. EKF)

Ensemble formulation of σ^b

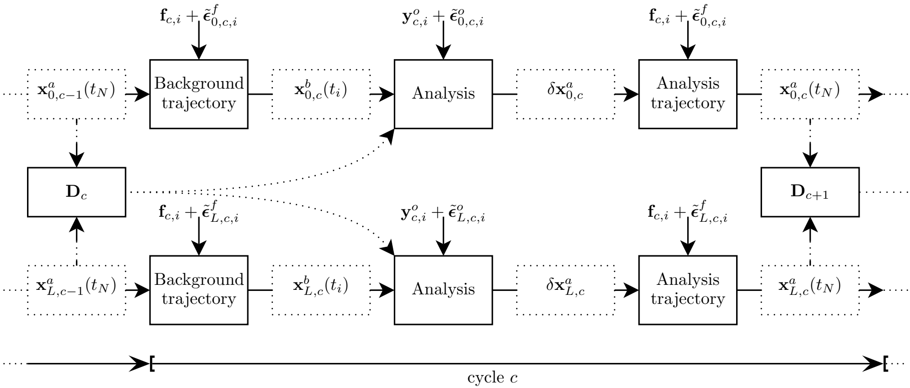
Estimate σ^b from the difference between background states of successive ensemble members, $l = 0, \ldots, L-1$:

$$\begin{aligned} \mathsf{D}_{(\widehat{\mathbf{w}})} &= \mathrm{diag}\left\{(\sigma^b_T)^2, (\sigma^b_{S_U})^2\right\} \\ &= \mathrm{diag}\left\{\frac{1}{2(L-1)} \sum_{l=0}^{L-1} \left[\mathsf{K}^{-1}_{(\mathbf{w})}\left(\mathbf{w}^b_l(t_0) - \mathbf{w}^b_{l+1}(t_0)\right)\right]\right. \\ &\qquad \left. \times \left[\mathsf{K}^{-1}_{(\mathbf{w})}\left(\mathbf{w}^b_l(t_0) - \mathbf{w}^b_{l+1}(t_0)\right)\right]^{\mathrm{T}}\right\} \end{aligned}$$

where

$$\mathsf{K}^{-1}_{(\mathbf{w})} = \begin{pmatrix} \mathsf{I} & 0 \\ -\mathsf{K}_{ST} & \mathsf{I} \end{pmatrix}$$

and $\mathbf{w}^b_L(t_0) = \mathbf{w}^b_0(t_0)$.



- The background-error variance matrix ($\mathbf{D}_c$) used for the analysis on cycle c is estimated from the sample variance matrix computed from the ensemble of background states ($\mathbf{x}^b_{l,c}(t_0)$) at the start of cycle c.
- In our set-up, $\mathbf{x}^b_{l,c}(t_0) = \mathbf{x}^a_{l,c-1}(t_N)$.

Ensemble formulation of σ^b

Advantages:

- The analysis and forecast state perturbations from the cycled ensemble system will evolve like the true analysis and forecast errors. (Berre *et al.* 2006; Daget *et al.* 2008)
- If the perturbations to the system input parameters (observations, initial state, model error, forcing and boundary terms) are random samples drawn from the pdf of the true errors then

$$\begin{aligned}
\mathbf{P}^b(t_0) &\equiv E[(\mathbf{x}^b(t_0) - \mathbf{x}^t(t_0))(\mathbf{x}^b(t_0) - \mathbf{x}^t(t_0))^{\mathrm{T}}] \\
&\approx \frac{1}{L-1} \sum_{l=1}^{L-1} \left(\mathbf{x}_l^b(t_0) - \mathbf{x}^b(t_0)\right) \left(\mathbf{x}_l^b(t_0) - \mathbf{x}^b(t_0)\right)^{\mathrm{T}} \\
&\approx \frac{1}{2(L-1)} \sum_{l=0}^{L-1} \left(\mathbf{x}_l^b(t_0) - \mathbf{x}_{l+1}^b(t_0)\right) \left(\mathbf{x}_l^b(t_0) - \mathbf{x}_{l+1}^b(t_0)\right)^{\mathrm{T}}
\end{aligned}$$

where $\mathbf{x}_L^a(t_0) = \mathbf{x}_0^a(t_0) = \mathbf{x}^a(t_0)$.

The disadvantages are essentially practical.

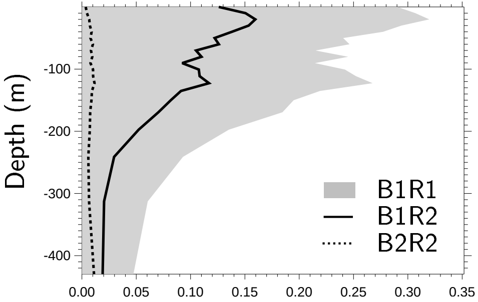
- An ensemble-based data assimilation system is expensive.
  - A model integration and variational analysis must be performed separately for each ensemble member.
- Designing an effective perturbation strategy is tricky.
  - Which input parameters to perturb?
  - How to obtain an adequate spread with few ensemble members?
- Constructing appropriate perturbations to the system input parameters is difficult.
  - We can only approximately sample the pdf of the true errors.

- A 9-member ensemble.
- The perturbed input parameters:
  - the surface forcing fields (heat flux, fresh-water flux, wind-stress);
  - the temperature and salinity observations;
  - the background state;
  - model error is neglected.
- Construction of the perturbations:
  - the forcing perturbations are derived from differences between different forcing analysis products (Balmaseda *et al.* 2008);
  - the observation perturbations are drawn from a Gaussian pdf with covariance matrix $\mathbf{R}$;
  - the background state is perturbed implicity via the cycling procedure;
- Reduction of sampling error:
  - A 90-day (9-cycle) sliding window is used, giving an effective ensemble size of 81 on each cycle for estimating σ^b.
  - Intraseasonal variability in σ^b is thus filtered out.

Outline

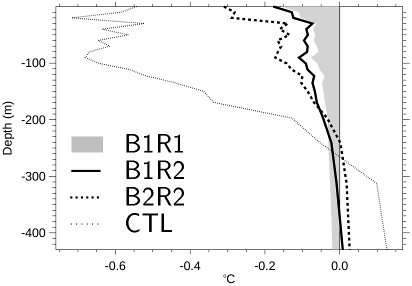
The numerical experiments

- The experimental design follows the common reanalysis procedures used in the ENSEMBLES and ENACT projects (Davey *et al.* 2006).
- The experiments are performed for the 9-year period from 1 January 1993 to 31 December 2001.
- A 10-day assimilation cycle is used.
- The experiments:
  - **CTL** : no data assimilation.
  - **B1R1** : parameterized σ^b, and σ^o defined using globally-averaged estimates from Ingleby and Huddleston (2007).
  - **B1R2** : parameterized σ^b, and σ^o estimated from Fu *et al.* method.
  - **B2R2** : ensemble σ^b, and σ^o estimated from Fu *et al.* method.
- Results will be displayed for temperature only (results for salinity are qualitatively similar).

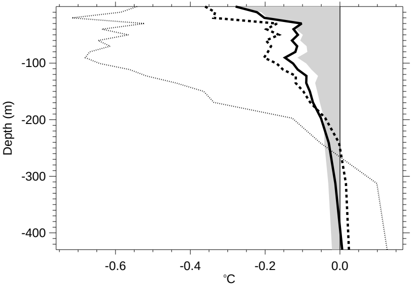


$$w_T = \frac{(\sigma_T^b)^2}{(\sigma_T^b)^2 + (\sigma_T^o)^2}$$

- Neglecting correlations, w_T is the average weight for an innovation.
- Both σ_T^b and σ_T^o have been computed at observation points, and averaged over the 1994-2000 period and the global domain.

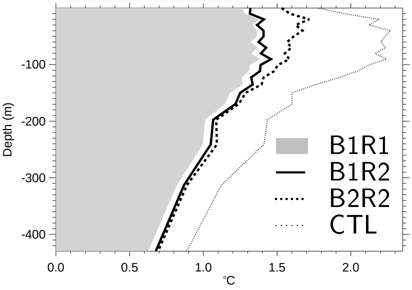


$\overline{\mathbf{r}}$

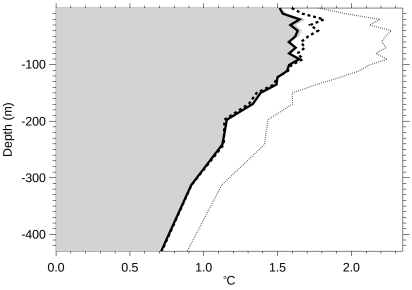


$\overline{\mathbf{d}}$

- $\mathbf{r} = \mathbf{y}^o - \mathbf{H}\mathbf{w}^a$ (residual) and $\mathbf{d} = \mathbf{y}^o - \mathbf{H}\mathbf{w}^b$ (innovation).
- $\overline{\mathbf{z}}$ indicates spatial (global) and temporal (1994-2000) average.
- Mean bias in CTL is reduced substantially in all assimilation expts.

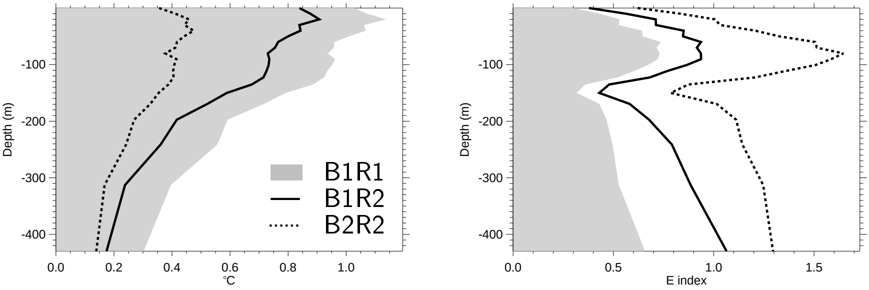


sd(**r**)



sd(**d**)

- $\mathbf{r} = \mathbf{y}^o - \mathbf{H}\mathbf{w}^a$ (residual) and $\mathbf{d} = \mathbf{y}^o - \mathbf{H}\mathbf{w}^b$ (innovation).
- $\mathrm{sd}(\mathbf{z}) = \sqrt{\overline{(\mathbf{z} - \overline{\mathbf{z}})^2}}$
- All assimilation expts. improve the fit to the observed variability.
- The "error growth" in the 10-day forecast is smallest for B2R2.



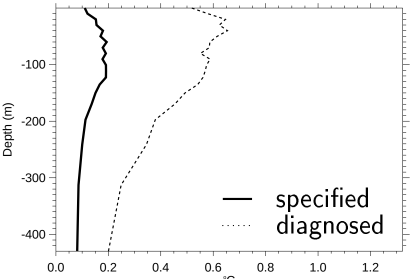
$\mathrm{rms}(\mathbf{H}\delta\mathbf{w}^a)$

$$\mathsf{E} = \frac{\mathrm{rms}(\mathbf{d}_{\mathrm{CTL}}) - \mathrm{rms}(\mathbf{d})}{\mathrm{rms}(\mathbf{H}\delta\mathbf{w}^a)}$$

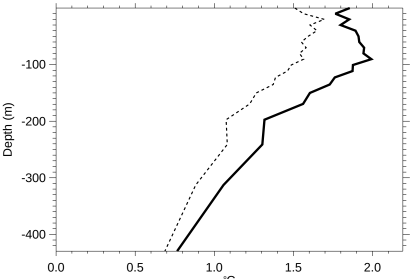
- $\mathrm{E} = \dfrac{\text{10-day forecast error improvement with respect to CTL}}{\text{"work done" by assimilation method to reduce forecast error}}$
- $\mathrm{E} > 0$ ($\mathrm{E} < 0$) implies assimilation is beneficial (detrimental).
- E increases (decreases) if $\mathbf{d}$ or $\delta\mathbf{w}^a$ decreases (increases).

Specified versus diagnosed σ^b and σ^o for temperature in B2R2

(method of Desroziers *et al.* 2005)



σ^b_T



σ^o_T

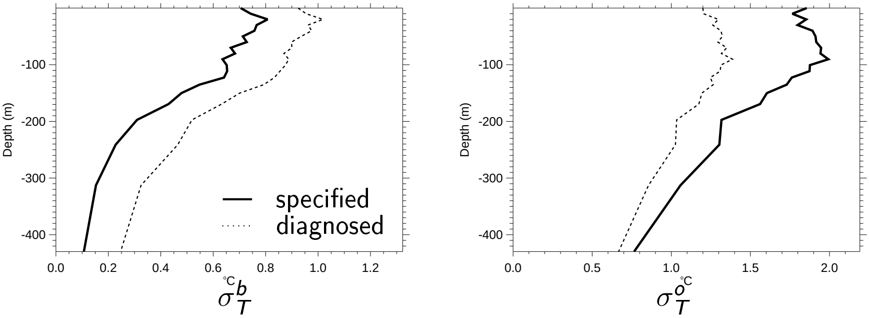
- If **B** and **R** are good estimates of the true background- and observation-error covariance matrices then

$$E[\mathbf{d}\,(\mathbf{H}\delta\mathbf{w}^a)^{\mathrm{T}}] \approx \mathbf{H}\mathbf{B}_{(\mathbf{w})}\mathbf{H}^{\mathrm{T}}$$

$$E[\mathbf{d}\,(\mathbf{d} - \mathbf{H}\delta\mathbf{w}^a)^{\mathrm{T}}] \approx \mathbf{R}$$

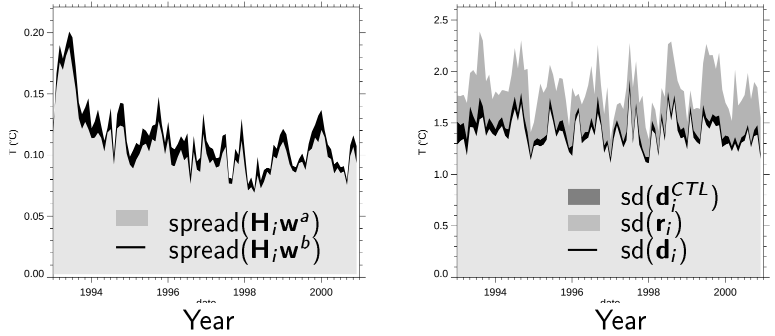
- σ^b_T is underestimated, and σ^o_T is overestimated.

(method of Desroziers *et al.* 2005)



- σ_T^b is underestimated (to a lesser extent than in B2R2).
- σ_T^o is overestimated (to a greater extent than in B2R2).

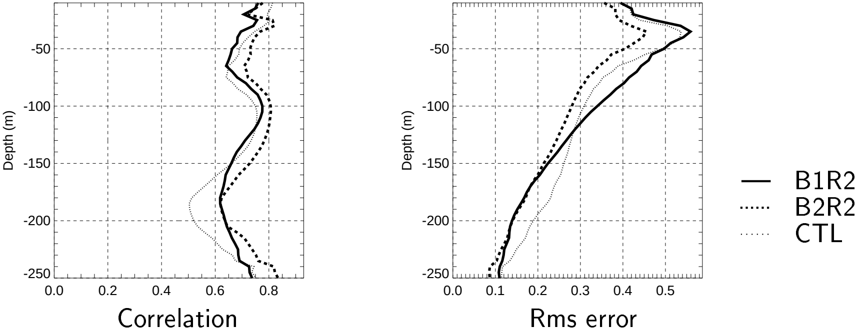
Experiment B2R2



- $\text{spread}\{\mathbf{H}_i\mathbf{w}^{a,b}\} = \sqrt{\frac{1}{L-1}\sum_{l=0}^{L-1}\left(\mathbf{H}_i\mathbf{w}_l^{a,b}(t_i) - \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{H}_i\mathbf{w}_l^{a,b}(t_i)\right)^2}$
- Spread of the analysis $<$ spread of the background.
- No evidence of ensemble collapse.
- Spread($\mathbf{H}_i\mathbf{w}^{a,b}$) is approximately a factor 10 smaller than sd($\mathbf{r}_i$), sd($\mathbf{d}_i$).

Current-meter data from TAO moorings

(Example from the eastern Pacific (110°W))



- B2R2 outperforms B1R2 (and B1R1) at all moorings.
- B2R2 outperforms CTL in the central and eastern Pacific, but slightly worse in the western Pacific.

Outline

Conclusions

- Both the parameterized and ensemble σ^b formulations produce a significant reduction in the rms of the innovations, with the parameterized σ^b slightly better above 150 m.
- The ensemble σ^b analyses are better "balanced" and closer to independent data (sea-level anomalies from T/P and current-meter data from TAO) than those produced with the parameterized σ^b.
- Statistical consistency checks suggest that the ensemble σ^b are underestimated. The parameterized σ^b are also underestimated but to a lesser extent.
- The apparent underestimation of the ensemble spread points to the need to increase the ensemble size and/or to improve the ensemble generation strategy.
- Producing an ensemble of analyses is costly but may be justified if these analyses can be used simultaneously for probabilistic forecasting as well as background-error covariance estimation.

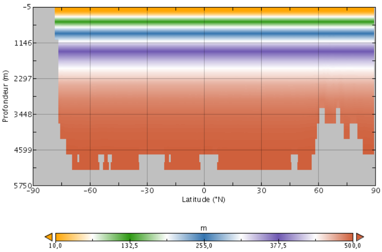
Future work

- Implementation of the ensemble ocean analysis procedure in the NEMOVAR system (OPAVAR's successor) and application to higher resolution global configurations.
  - A contribution to the NEMOVAR project, supported by LEFE-ASSIM.
- Improvements and extensions to the ensemble generation strategy (e.g., to include explicit model-error perturbations).
- Exploitation of the ensemble information for estimating other aspects of the background-error covariance matrix
  - Geographically-dependent length scales for our diffusion-based correlation model.
    (Pannekoucke *et al.* 2008; Daget 2008, *PhD thesis*)
  - Testing the validity of assumptions made in the balance operator.
    (Daget 2008)

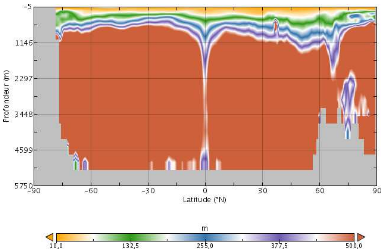
Postdoctoral position available at CERFACS: see
http://www.cerfacs.fr/Open_Positions/index_offres.php

Ensemble-estimation of vertical length scales

Example for temperature (from Daget 2008, PhD thesis)



Fixed scales currently used



Ensemble estimates