An ensemble approach for the estimation of observational error variance

With application to the nominal 1° POP2 ocean model

Alicia R. Karspeck

aliciak@ucar.edu

NCAR, P.O. Box 3000, Boulder, CO 80307-3000, USA

Acknowledgements

This work was funded in part by the NOAA Climate Program under the Climate Variability and Predictability Program grants NA09OAR4310163 and NA13OAR4310138, and by the NSF Collaborative Research EaSM2 grant OCE-1243015. NCAR is sponsored by the NSF and the CCSM project is supported by NSF and the Office of Science Biological and Environmental Research of the DOE. Computing resources were provided by the Climate Simulation Laboratory at NCAR's Computational and Informational Systems Laboratory (CISL). Support with accessing data from the WOD09 was provided by CISL's Data Support Section.

Least-squares algorithms for data assimilation require estimates of both background error covariances and observational error variances. The specification of these errors is an important part of designing an assimilation system because the relative sizes of these uncertainties determines how much an innovation will impact the state estimate. Observational error estimates typically lump together measurement/instrumental errors with representativeness error. In a coarse-resolution ocean model, over much of the ocean, the errors of representation are the dominant contribution to observational errors[1]. Since representativeness errors are meant to account for the unresolved scales and physical processes in the model, the variance of these errors will vary by geographic region and will also vary from model to model.

Here we present a practical approach for estimating model-dependent, spatially varying observational error variances. The method uses ensemble model simulations to compute an expected value and uncertainty associated with the estimator – providing approximate confidence intervals for the true observational error. We illustrate the method for the POP2 global ocean general circulation model.

An ensemble of ocean simulations

We can symbolically formulate the ocean modeling problem as:

$$z_t^k = \mathcal{M} z_{t-1}^k + \mathcal{F}_{t-1}^k$$

where $\mathcal{M}$ is a discrete operator that describes the dynamics (resolved and/or parameterized) of the ocean model system, z_t is the vector of model state variables at time t, and $\mathcal{F}$ is a prescribed, time-dependent forcing from the atmosphere.

Ensemble realizations k of the ocean state will differ from one another due to differences in their initial states and differences in the atmospheric forcing. Let us assume the forcing is a normally distributed (and possibly auto-correlated) process and that an ensemble set of $\mathcal{K}$ atmospheres is available for forcing a set of $\mathcal{K}$ ocean states[2].

The relationship between ensemble members, the "truth," and the observations

A desirable property of any forecast ensemble is that it is "reliable," i.e. that the ensemble members are random draws from the same distribution as the truth. We can slightly alter this definition to account for representativeness error by instead saying that the ensemble members are random draws from the same distribution as the *model-resolvable* component of the truth, which we will hereafter refer to as z^*. This distinction will allow us to estimate the pooled observational error variance that stems both from representativeness error and measurement/instrumental error (hereafter σ_o^2).

A statistical (conceptual) model any member k of a *reliable* ocean ensemble, for z^*, and for observations (y) of the system can be expressed.

$$z^k = \mu + e^k \; ; \qquad e^k \sim \mathrm{N}(0, \Sigma^2)$$
$$z^* = \mu + e^* \; ; \qquad e^* \sim \mathrm{N}(0, \Sigma^2)$$
$$y_j = \mathcal{H}_j z^* + \varepsilon_j^o \; ; \qquad \varepsilon_j^o \sim \mathrm{N}(0, \sigma_o^2)$$

The linear operator $\mathcal{H}$ maps to the time/space location j of an observation[3]. Consistent with the definition of reliability, the vector random variables z^k and z^* are samples from the same distribution. The covariance matrix Σ^2 permits both time and space autocorrelation, however all realizations of e^k are uncorrelated with one another, uncorrelated with e^*, uncorrelated with ε^o, and uncorrelated with μ. The observational errors are also uncorrelated in time and space.

An estimator of the observational error variance

The data for our estimator of σ_o^2 consist of n observations available through time in a fixed geographic box that we believe has constant observational error variance. We can also compute the ensemble model solution at the time/space location of each observation (indexed by j) through $x_j^k = \mathcal{H}_j z^k$.

We can now form the following estimator for the observational error within the regional box:

$$\Phi^2(x,y) = \frac{1}{n}\sum_{j=1}^{n}\left[\, y_j - \langle x \rangle_j \right]^2 - \frac{1}{n}\sum_{j=1}^{n} s_j^2$$

where $< x >$ is the ensemble mean and s^2 is the sample ensemble variance.

Properties of the estimator

Properties of the estimator are derived through the algebra of random variables, but details are not shown here. Happily, estimator is unbiased, i.e.

$$\mathrm{Ex}(\Phi^2) = \sigma_o^2$$

For approximating $Var(\Phi^2)$, we note that in the limit of $\mathcal{K} >> 1$, $\mathcal{H}_j\mu$ and $\mathcal{H}_j\Sigma\mathcal{H}_j^{\mathrm{T}} = \sigma_j^2$ can be approximated by the model ensemble sample statistics (i.e. $\mathcal{H}\mu_j \approx \langle x \rangle_j$ and $\sigma_j^2 \approx s_j^2$). Further assuming (for simplicity only!), that σ_j^2 can be approximated by its average value over all j, we come to an approximation for the variance of $\Phi^{2:}$

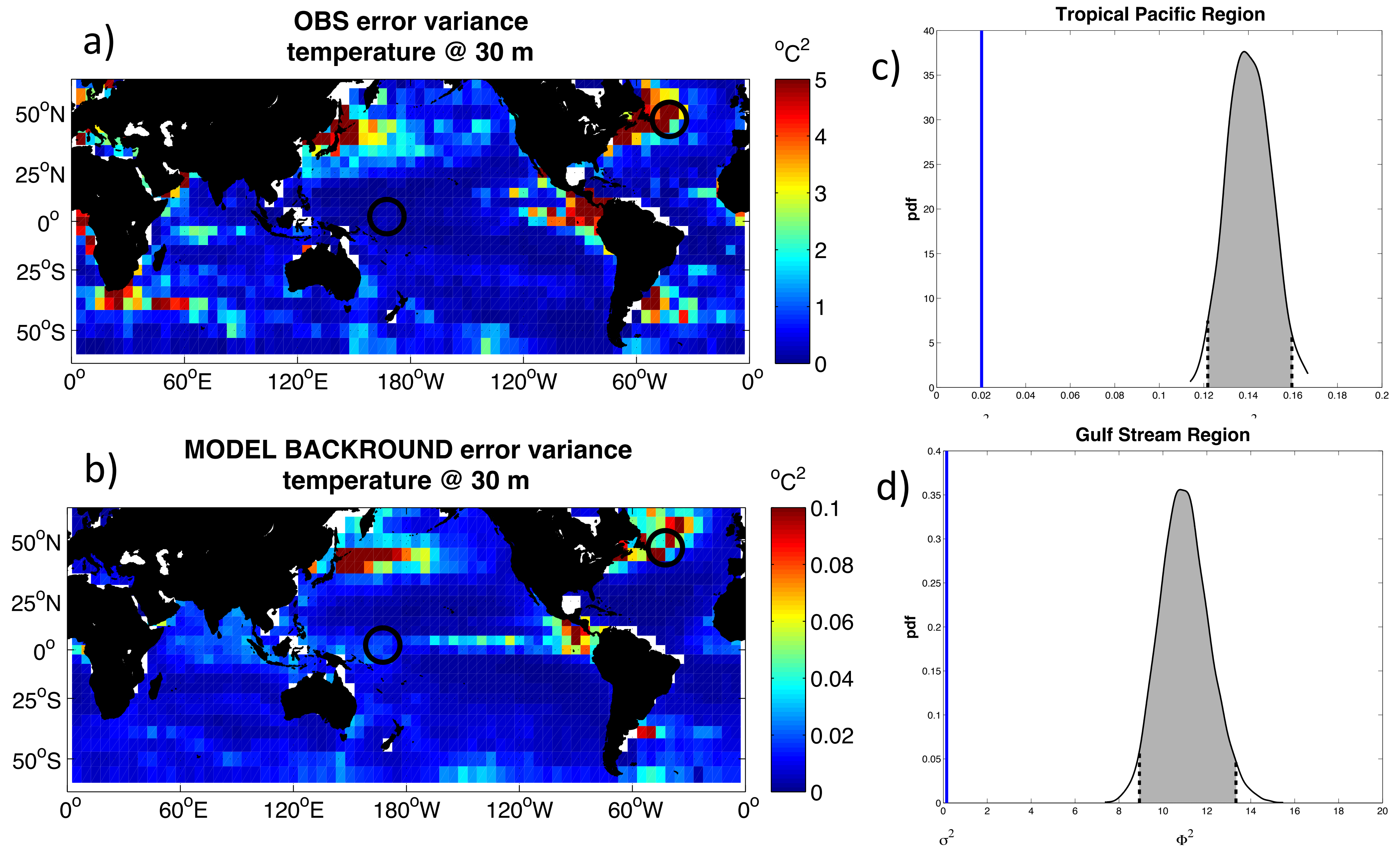
$$Var(\Phi^2) \approx \frac{2}{n}\left(\sigma_o^4 + \frac{n}{v_{eff}}\sigma^4 + 2\sigma_o^2\sigma^2\right)$$

where ν_{eff} is an effective degrees of freedom for the model ensemble within the regional box. For data that are irregularly spaced in time,

$$v_{eff} \approx \frac{n}{1 + n\rho^2}$$

and ρ^2 is the squared autocorrelation for the model background error averaged over every unique pair of observational locations.

Application to the POP2 nominal 1° ocean model



a) Estimate of the POP2 observational error variance based on a 3o member, one year ensemble simulation forced with an ensemble of atmospheric states from a reanalysis. Estimates were computed in 2.5° regional boxes. b) Average model background error variance from same simulation. c) For a region in the tropical Pacific (indicated by circle in a,b), blue line is the average model background variance and gray is the distribution associated with our estimator Φ^2. The distribution is centered at the estimated value, and the dashed lines indicate the range between which we are 95% confident that the true value of observational error variance resides. d) same as c, for the Gulf Stream region.

Discussion

The estimator for observational error variance that we present here capitalizes on the assumption that observational error has no temporal autocorrelation and is also temporally uncorrelated with model background error. We also make use of an ensemble of model simulations to explicitly account for model background error. There are other methods in the literature which attempt to estimate σ_o^2 by exploiting the fact that background errors are uncorrelated in space, while observation errors are not[5]. These should be viewed as complementary approaches (and indeed give similar results). Choosing to filter in space or time are equally valid, although the choice may effect the accuracy of the method.

In terms of the accuracy of this estimator, it can be understood from the (approximate) $Var(\Phi^2)$ that as the total number of observations in the data sample increases, and ρ, σ^2 decreases, confidence that our estimate is near the true σ_o^2 goes up. While the size of σ_o^2 is an irreducible factor in $Var(\Phi^2)$, n, ρ, and σ^2, can, in fact, be altered. Trivially, more data points can be used if they are available. Less trivially, ρ and σ^2 can be reduced through data assimilation. In fact, this is the very goal of data assimilation. If it is done optimally, the model system errors will retain the necessary properties (ie. uncorrelated with the "truth," uncorrelated with the observational errors, and uncorrelated with each other) such that Φ^2 is still a valid estimator.

Footnotes:

[1] Instrument error variance for temperature observations is less than ~.01°C² according to WOD09 documentation.

[2] Certainly, the free atmosphere is not linear, but one can imagine an ensemble reanalysis of atmospheric forcing fields that are nearly-Gaussian around their analyzed mean.

[3] The assumption of normally distributed errors throughout is predicated on $\mathcal{M}$ and $\mathcal{H}$ being linear operators (okay, but not perfect, for 1° ocean model)

[4] Vieba (2010), Metrol. Meas. Syst.

[5] e.g. Rutherford (1972), J. Atmos Sci. ; Richmond et al. (2005),Geophys Res. Lett.