Calculating T and S uncertainties using a high-res ocean model

Ian Fenty

Contents

1 Motivation

To objectively quantify the degree of consistency between a set of independent observations of a system state (data) and a reconstructed estimate of that state from a model (state estimate), one must define measures that define consistency in terms of the differences between the data and the reconstruction. These measures may be defined with respect to the statistics of the model-data residuals.[1] Examples of statistical properties of the residuals, ϵ, included the expected value of their mean and the squared deviation from the mean,

$$\begin{aligned} E[\epsilon] =& E\left[x_o - x_m\right] = \mu_\epsilon \\ E[(\epsilon - \mu_\epsilon)^2] =& Var(\epsilon) = \sigma_\epsilon^2 \end{aligned}$$

Where $E[x]$ is the expected value of the random variable x, μ_ϵ is the expected mean of the residuals, and σ_ϵ^2 is the expected squared deviation of the residuals from their mean.

Defining consistency Defining consistency in terms of the statistics of model-data residuals requires the specification of a function, f, that takes as inputs these expectation values and the actual model-data residuals,

$$f = f\left(\mu_\epsilon, \sigma_\epsilon^2, \epsilon\right)$$

The evaluation of the arbitrary function f for a set of expectation values and model-data residuals can provide an objective measure of the consistency of the state estimate. In addition to f, additional conditions may be specified for the model-data residuals, such as homoscedasticity of their

[1] These are typically treated as a set of uncorrelated random variables.

variance.[2]

A common requirement for model-data consistency is for the model and data to have identical means (zero expected value of the residual, $\mu_\epsilon = 0$) and for the residuals to be uncorrelated with a finite variance,

$$\begin{aligned} E[\epsilon] =& \mu_\epsilon = 0 \\ E[(\epsilon - \mu_\epsilon)^2] =& E[(\epsilon - 0)^2] = E[\epsilon^2] = \sigma_\epsilon^2 \end{aligned}$$

Specification of several residual statistics In large ocean state estimation problems one does not typically define a single value of μ_ϵ and σ_ϵ^2 over the entire domain in time and space and for all observations independent of type or instrument. Indeed, one often defines a set of μ_ϵ and σ_ϵ^2 with each member corresponding to one or more observation types, instrument, or locations in time and space. Because the number of measurements in each $(\mu_\epsilon, \sigma_\epsilon^2)$ set may be few and the number of $(\mu_\epsilon, \sigma_\epsilon^2)$ sets are so large, it is impractical to simply analyze the individual distributions of model-data residuals for each set to determine overall model-data consistency.

Instead, progress is made by scaling each model-data residuals by its corresponding $(\mu_\epsilon, \sigma_\epsilon^2)$ set and then combining these scaled residuals into a single, new distribution with properties that can be more easily evaluated against a consistency criterion.

One such function used to evaluate model-data consistency of the scaled residuals is the

[2] For example, requiring that there is no correlation between the magnitude of simulated SST vs observed SST residuals as a function of ocean surface temperature or salinity.

reduced chi-squared statistic, χ_r^2,

$$\chi_r^2 = \frac{1}{N-n-1}\sum_{i=1}^{N}\frac{\left[\epsilon(i)-\mu_\epsilon(i)\right]^2}{\sigma_\epsilon^2(i)}$$

Where N is the number of observations, and n is the number of free parameters in the distribution ($n = 1$ for a standard normal distribution with a zero-mean and unknown standard deviation). The χ_r^2 statistic is proportional to the sum of the variances of the actual residuals scaled by their expected values. Model-data consistency then amounts to evaluating the ratio of the χ_r^2 for the actual residuals to the χ_r^2 expected if the residuals had distributions consistent with μ_ϵ and σ_ϵ^2. Model-data consistency is achieved when this ratio evaluates to unity.

Of course, one is free to define additional or alternative functions to quantify model-data consistency.

2 Defining model-data consistency

Determining the actual measures to define model-data consistency is a central problem in ocean state estimation because there are many possible causes of model-data difference that must be taken into account. The causes of model-data difference include errors associated with imperfect measurements (data errors), a model's inability to represent the true ocean state (representation error), and errors associated with a model's imperfect encoding of the physics of the real system (model error). All of these possible causes have consequences for the model-data residuals and the their statistical moments.

Failure to properly define these error measures can lead to the problem of over- or underfitting the model to the data. In an adjoint-based state estimation methodology, failing to properly specify these measures can lead to suboptimal iterative adjustments of the model control variables. In practice, incorrectly defined consistency measures may cripple the optimization machinery by causing control variable adjustments that fail to make useful improvements of the state estimate.

Here, I focus on calculating the expected second order moments of model-data residuals associated with model representation error, the inability of the model to represent the true spatial and temporal variability of ocean fields. Even in state-of-the-art models, representation errors can be much larger than measurement errors over much of the global ocean. Indeed, given the turbulent, nonlinear nature of the ocean's meso- and sub-mesoscale eddy field and the set of model control variables used by ECCO-class adjoint ocean state estimation systems, model representation errors will be the dominant contribution to model-data misfits for as long as we seek ocean reconstructions that faithfully reproduce the general trajectory and the statistical properties of global ocean variability and not its exact evolution.

If the statistics of a model's temperature and salinity distributions defined over a given volume and time approach the statistics of the true distributions as model resolution increases, one may analyze the output of a high resolution model to estimate the variances of the true state.

The approach I take is to estimate the variances of the expected model-data difference due to model representation errors for a relatively coarse model (llc90, $\Delta x = 45$ km) by analyzing the simulated temperature and salinity fields from a model with a much finer spatial resolution (llc2160, $\Delta x = 2$ km). Ocean grid cell areas in the finer model are approximately 570 times smaller than the coarser.

3 Uncertainty 1: Spatial Sampling

To compare a model state against *in situ* observational data, one must interpolate the model state to the observation's location in space. The full spatial variability of temperature and salinity fields in the real ocean within a given volume cannot be perfectly represented in models with finite spatial resolution. The temperature and salinity values associated with model grid cells represent the means of those fields within the volumes defining those cells. Consequently, even if the model perfectly reproduced the volume-averaged representation of the true field, one would expect a nonzero variance in the residuals between the model field interpolated to the observation location and the observation itself. Determining whether a model's state is consistent with the observational data therefore requires a characterization of the expected statistics of these residuals.

The true ocean field $\Psi(t, x)$ within volume V at time t has a spatial mean $\overline{\Psi(t, V)}$ and its distribution has a second central moment $\sigma^2(t, V)$,

$$\begin{aligned} \overline{\Psi(t,V)} &= \iiint_V (\Psi(t,x) dV \\ \sigma^2(t,V) &= \iiint_V \left[\Psi(t) - \overline{\Psi(t,x)} \right]^2 dV \end{aligned}$$

The distribution of $\Psi(t, x)$ in V is unlikely to be Gaussian but we ignore high-order moments here.

Given that a model state at a given time represents the *volume average* of the true field over each model grid cell, one reasonable model-data consistency requirement is that the residuals of the model-data differences in each grid cell to have a zero expected mean and a nonzero expected variance. The expected variance of the model-data residuals should be no less than the spatial variance of the true field within that same volume. For simplicity, I assume that the spatial variance of the true field is stationary (time independent).

To estimate the true spatial variance of the ocean T and S fields, we analyze the spatial distributions of *high-resolution model fields* within volumes that are defined by each *coarse-resolution model grid cell*. Because at any given time the spatial distribution of the high-resolution field within V may be an under- or over-estimate the true spatial distribution, we seek a time-invariant estimate of $\sigma^2(x)$ by combining multiple estimates of $\sigma^2(t, x)$ from the high-resolution model from within a time window of length τ.

The appropriate choice of τ and the method of combining these distributions to estimate the true variance depends in part on the temporal and spatial variability of the underlying field. If temporal changes in the spatial mean of the high-resolution T and S fields in V are small over τ, then one can simply analyze a sample of the high-resolution model output within V and τ. On the other hand, if the temporal variability of the spatial mean of T and S are large over τ, one must combine several separate variance estimates calculated over shorter time intervals within which the T and S spatial means can be assumed constant. For example, because extratropical sea surface temperatures undergo a large seasonal cycle, estimates of $\sigma^2(x)$ must be made over time periods that are much shorter than a seasonal cycle to avoid including temporal variability into the estimate of spatial variability.

Here, I make separate estimates of $\sigma^2(x)$ by analyzing the spatial distribution of the high-resolution model output in V over daily time intervals over a time period τ and then combine these separate variance estimates using the method of pooled variance. I make the assumption that that the spatial variance of the true field within V is constant over τ.

Method The method used to estimate the expected magnitude of the model-data residuals associated with spatial sampling errors associated with the model's finite spatial resolution is as follows:

1. For each volume V that defines a llc90 grid cell, define the set of points $X_V = \{x_1, x_2, ..., x_{n_x}\}$ whose n_x members are the subset of the llc2160 model grid points falling within V. In most llc90 grid cells, $n_x \approx 520$.

2. Define a time window τ that spans n_t high-resolution model time steps over which to sample the high resolution model field $\psi_{llc2160}$.

3. For every high resolution model time step $t \in \tau$ and llc90 grid cell volume V, extract a subset of the high-resolution model output, $\psi_{llc2160}(t, X_V)$.

4. For every such subset, find its mean, $\overline{\psi_{llc2160}(t, V)}$, and make an unbiased estimate of the

spatial variance of the true field, $\hat{\sigma}^2_{spatial}(t, V)$,

$$\begin{aligned}\overline{\psi_{llc2160}(t, V)} =& \frac{1}{n_x} \sum_{i=1}^{n_x} \psi_{llc2160}(t, X_V(i)) \\ \hat{\sigma}^2_{spatial}(t, V) =& \frac{1}{n_x - 1} \sum_{i=1}^{n_x} \left[\psi_{llc2160}(t, X_V(i)) - \overline{\psi_{llc2160}(t, V)} \right]^2\end{aligned}$$

5. Combine the n_t separate $\hat{\sigma}^2(t, V)$ to make a time-independent spatial variance estimate, $\hat{\sigma}^2_{spatial}(V)$, using the method of pooled variance,

$$\hat{\sigma}^2_{spatial}(V) = \frac{\sum_{t=1}^{n_t} (n_x - 1)\, \hat{\sigma}^2_{spatial}(t, V)}{\sum_{t=1}^{n_t} (n_x - 1)}$$

I chose $\tau = 90$ days and $t = 1$ day for each calculation of $\hat{\sigma}^2_{spatial}(V)$.

Also, I tried 5 different 90 day windows, with the first starting from 3/1.

4 Uncertainty 2: Unresolved Sub-seasonal Temporal Variability of the Volume-Mean Fields

Over time period τ, the volume-mean of the true field within volume V has a (temporal) mean $\overline{\Psi(\tau, V)}$ and a distribution with a second central moment $\sigma^2(\tau, V)$,

$$\begin{aligned}\overline{\Psi(\tau, V)} &= \int_\tau \overline{\Psi(t, V)}\, dT \\ \sigma^2(\tau, V) &= \int_\tau \left[\overline{\Psi(t, V)} - \overline{\Psi(\tau, V)} \right]^2 dT\end{aligned}$$

Over what timescales can the llc90 model reproduce temporal variations of the true volume-mean ocean temperature and salinity? As the llc90 model is not eddy resolving, it cannot represent temporal variability of the volume-mean associated with eddy processes. Thus, a reasonable model-data consistency requirement is that model-data residuals have an expected mean value of zero and an expected variance that is the same as the (temporal) variance of the volume-mean of the true temperature and salinity fields over eddy timescales. Since the eddy temporal variability timescale is not well-defined, I chose 90 days.

Method The method used to estimate the expected magnitude of the model-data residuals associated with model representation errors associated with the inability of the model to reproduce fluctuations on mesoscale eddy timescales and shorter is given by,

1. For each volume V that defines a llc90 grid cell, define the set of points $X_V = \{x_1, x_2, ..., x_{n_x}\}$ whose n_x members are the subset of the llc2160 model grid points falling within V. In most llc90 grid cells, $n_x \approx 520$.

2. Define a time window τ that spans n_t high-resolution model time steps over which to sample the high resolution model field $\psi_{llc2160}$. τ should be approximately equal to the timescale of ocean variability that we expect the model to be able to reproduce.

3. For every high resolution model time step $t \in \tau$ and llc90 grid cell volume V, extract a subset of the high-resolution model output, $\psi_{llc2160}(t, X_V)$.

4. Find the volume mean of every such subset, $\overline{\psi_{llc2160}(t, V)}$,

$$\overline{\psi_{llc2160}(t, V)} = \frac{1}{n_x} \sum_{i=1}^{n_x} \psi_{llc2160}(t, X_V(i))$$

5. Find the temporal mean over τ of these volume mean subsets, $\overline{\psi_{llc2160}(\tau, V)}$,

$$\overline{\psi_{llc2160}(\tau, V)} = \frac{1}{n_t} \sum_{t=1}^{n_t} \overline{\psi_{llc2160}(t, V)}$$

6. Make an unbiased estimate of the temporal variance over τ of these volume means, $\hat{\sigma}^2_{eddy}(\tau, V)$,

$$\hat{\sigma}^2_{eddy}(\tau, V) = \frac{1}{n_t - 1} \sum_{t=1}^{n_t} \left[\overline{\psi_{llc2160}(t, V)} - \overline{\psi_{llc2160}(\tau, V)} \right]^2$$

7. Repeat Step 2 a total of n_τ times, in each case using different start times in order to make n_τ estimates of $\sigma^2_{eddy}(\tau, V)$. Collect these separate estimates as $\sigma^2_{eddy}(\tau_i, V) \;\; \tau_i \in \{1, 2, \ldots n_\tau\}$

8. Combine these n_τ separate $\hat{\sigma}^2_{eddy}(\tau_i, V)$ estimates made with different start times to make a final, start-time independent estimate, $\hat{\sigma}^2_{eddy}(V)$, using the method of pooled variance,

$$\hat{\sigma}^2_{te}(V) = \frac{\sum_{i=1}^{n_\tau} (n_t - 1)\, \hat{\sigma}^2_{eddy}(\tau_i, V)}{\sum_{i=1}^{n_\tau} (n_t - 1)}$$

I chose $\tau = 30$ days and n_τ = 12 to make $\hat{\sigma}^2_{te}(V)$.

5 Uncertainty 3: Improper Sub-seasonal Temporal Variability of the Volume-Mean Fields

The motivating idea for Uncertainty 2 was that no model can perfectly represent the full range of temporal variability of the true volume-averaged field and, consequently, one must account for this source of additional expected variance in the model-data residuals. Implicit in Uncertainty 2 was the idea that since the llc90 model is able to capture the evolution of the mean field on timescales longer than some τ, it is sufficient to characterize the expected second order moments of model-data residuals due to eddy fluctuations on timescales shorter than τ. Indeed, it would be sufficient to only characterize the expected variability of the true eddy field if the model fields exhibited little variability on timescales shorter than τ. However, in some parts of the ocean this is not true, the llc90 model does exhibit significant sub-seasonal variability. The timescale of this significant sub-seasonal llc90 variability, τ_{llc90}, may be quite different than that of the true ocean since the model cannot explicitly simulate the full range of dynamical motions. Thus, it is not reasonable to expect that such sub-seasonal model variability reflect reality nor can one reasonably expect that all such variability be "corrected" through the model control variables. Consequently, when comparing a model state against *in situ* observational data, one must also account for the expected model-data residuals resulting from this type of model representation error.

Over the n_t llc90 model time steps within time period τ_{llc90}, the llc90 model field $\psi_{llc90}(t, V)$ associated with volume V has a (temporal) mean $\overline{\psi_{llc90}(\tau_{llc90}, V)}$ and a distribution with a second central moment $\sigma^2_{llc90}(\tau_{llc90}, V)$,

$$\begin{aligned}\overline{\psi_{llc90}(\tau_{llc90}, V)} =& \frac{1}{n_t}\sum_{t=1}^{n_t} \psi_{llc90}(t, V)\\ \sigma^2_{mod}(\tau_{llc90}, V) =& \frac{1}{n_t - 1}\sum_{t=1}^{n_t} \left[\psi_{llc90}(t, V) - \overline{\psi_{llc90}(\tau_{llc90}, V)}\right]^2\end{aligned}$$

Using the same arguments as were made in Uncertainty 2, one may pick a typical eddy turnover timescale or a timescale of eddy translation across the llc90 model grid cell for τ_{llc90}. However, as the llc90 model follows its own (non-eddying) dynamics and it is not immediately obvious what the best value of τ_{llc90} should be.

Nevertheless, a reasonable model-data consistency requirement is that model-data residuals due to the model's incorrect variability on sub-seasonal timescales (less than τ_{llc90}) have an expected mean value of zero and an expected variance that is the same as the (temporal) variance of the model fields over τ_{llc90}

Method The method used to estimate the expected magnitude of the model-data residuals associated with model representation errors associated with the model's incorrect sub-seasonal variability is as follows,

1. Define a time window τ_{llc90} that spans n_t coarse-resolution model time steps over which to sample the coarse resolution model field ψ_{llc90}. τ_{llc90} should be approximately equal to the typical timescales of model variability that are associated with the model's incomplete representation of sub-seasonal dynamics.

2. For every coarse-resolution model time step $t \in \tau$ and llc90 grid cell volume V, extract a the coarse resolution output, $\psi_{llc90}(t, V)$.

3. Find the temporal mean over τ of the coarse-resolution model field, $\overline{\psi_{llc90}(\tau, V)}$,

$$\overline{\psi_{llc90}(\tau_{llc90}, V)} = \frac{1}{n_t} \sum_{t=1}^{n_t} \overline{\psi_{llc90}(t, V)}$$

4. Make an unbiased estimate of the temporal variance over τ of the model field, $\hat{\sigma}^2_{mod}(\tau_{llc90}, V)$,

$$\hat{\sigma}^2_{mod}(\tau_{llc90}, V) = \frac{1}{n_t - 1} \sum_{t=1}^{n_t} \left[\psi_{llc90}(t, V) - \overline{\psi_{llc90}(\tau_{llc90}, V)} \right]^2$$

5. Repeat Step 1 a total of n_τ times, in each case using different start times in order to make n_τ estimates of $\sigma^2_{mod}(\tau_{llc90}, V)$. Collect these separate estimates as $\sigma^2_{mod}(\tau_i, V) \;\; \tau_i \in \{1, 2, \ldots n_\tau\}$

6. Combine these n_τ separate $\hat{\sigma}^2_{mod}(\tau_i, V)$ estimates made with different start times to make a final, start-time independent estimate, $\hat{\sigma}^2_{mod}(V)$, using the method of pooled variance,

$$\hat{\sigma}^2_{mod}(V) = \frac{\sum_{i=1}^{n_\tau}(n_t - 1)\,\hat{\sigma}^2_{mod}(\tau_i, V)}{\sum_{i=1}^{n_\tau}(n_t - 1)}$$

6 Combining Uncertainties

If the three sources of model-data residuals are uncorrelated then the expected squared deviation of the residuals from their (expected zero) mean is,

$$\sigma^2_{residuals} = \sigma^2_{spatial} + \sigma^2_{eddy} + \sigma^2_{model}$$