

Future Directions: What is beyond simple “binning/gridding”

Watson Gregg and Marlon Lewis

6.1 From binning to optimal interpolation

The objective of the approaches described above is to produce an ocean state or field that is as close as possible to reality. This approach should, in principle, take into consideration not only the observational errors, but the spatial and temporal error fields as well. Finally, it would be desirable to ensure that the resulting fields are dynamically consistent, taking into account all the available information - certainly the observational satellite data, but also model dynamics, *in situ* data, physical constraints, and climatology. This extends the options beyond simple binning schemes to include a variety of objective analysis methods, which reach their current state of the art in new four dimensional variational data assimilation schemes.

A relatively simple approach would be to weight the observations (or their transforms) based on their distance from the grid point, an approach that dates to the successive correction methods of Cressman (1959). Say we wish to create an analysis field at a specified location and at a given time $\mathbf{x}_a(j)$, from observations taken at different locations and at different times $\mathbf{y}(i)$. The analysis field in our context would be the Level-3 gridded data product, and the observations would be the remotely-observed normalized radiances (or perhaps some derivation of same, such as chlorophyll concentration, or attenuation length, or their statistical transforms). We might assume that the value or usefulness of the observations might diminish the farther one is away from the grid point j in space and time,

$$\mathbf{x}_a(j) = \frac{\sum_{i=1}^n w(i,j)\mathbf{y}(i)}{\sum_{i=1}^n w(i,j)}, \quad (6.1)$$

where the weights $w(i,j)$ must be determined.

In the simplest case (actually the one taken for the majority of the ocean-colour binning schemes described above), a “zone of influence” is defined in

space and time such that $w(i, j)$ is equal to one, if the observation is within the zone, and zero when it is not. For example, in the SeaWiFS case above, the rectangular tiles or bins of $\sim 85 \text{ km}^2$ constitute the spatial zone of influence, whereas the various daily, weekly and monthly products define the temporal zone,

$$w(i, j) = \begin{cases} 1, & d(i, j) < R \\ 0, & d(i, j) > R \end{cases}, \quad (6.2)$$

where R represents the spatial or temporal extent of the bin, and $d(i, j)$ represents the location of the observation, both relative to the bin centre.

Fixing the dimensions of the zone of influence, and setting the weights implicitly defines the expectation of the statistical nature of the spatial and temporal scales of variability, specifically the autocorrelation function. The weighting, as applied currently to most ocean-colour data sets, is rather crude, first because the implied autocorrelation function is not realistic (sharp cut off at the scale of the zone), and second, because it assumes that the de-correlation scale is somehow related to changes in the area of the bin with latitude, as a function of the Earth's sphericity. Neither is realistic.

A marginally better approach would be to recognize a weighting that falls off as the "distance" of the observation from the grid point increases. Lewis *et al.* (1988) applied such a method to the gridding of the global Secchi disc climatology as,

$$w(i, j) = \max\left(0, \frac{R^2 - d_{i,j}^2}{R^2 + d_{i,j}^2}\right), \quad (6.3)$$

where the "value" or weight given to a given observation falls off as the inverse square of the distance from the grid point. A point falling directly on the grid is given full weight, and observations outside the defined zone of influence are not accumulated for the given bin. Other schemes (*e.g.* weighting like $\exp\left(-\frac{d_{i,j}^2}{R^2}\right)$) can be envisioned. For the Secchi disc climatology, all points falling in a given temporal bin (seasonal and annual) were given a temporal weight of one, and the above only applied to the spatial field. R was taken to be 2 degrees of both longitude and latitude and the data were gridded to a one-degree resolution.

This approach represents a slight improvement in that it more realistically approximates the shape of observed autocorrelation functions, although fixing R as was done, does not reflect the nature of ocean variability. A more accurate approach was taken by Petrie *et al.* (1999) who were interested in mapping the horizontal (x and y), vertical (z) and temporal (t) distributions of nutrients along the North Atlantic coast, given the historical database of observations. They derived a weighting as:

$$w(i, j) = \exp(-d_{i,j}) * \left(1 + d_{i,j} + \frac{d_{i,j}^2}{3}\right), \quad (6.4)$$

where d_{ij} is defined as a non-dimensional distance (pseudo-distance), given by,

$$d_{i,j} = \sqrt{x'^2 + y'^2 + z'^2 + t'^2}, \quad (6.5)$$

and the primed quantities represent the scaled variables, *e.g.* x' is the distance in the x direction divided by the scale in that direction. The x , y , z , and t variables represent the horizontal, vertical and time coordinates. The appropriate scale varies with location and seasonally; the autocorrelation scales must be determined *a priori*. They estimated that the (x, y, z, t) scales for the continental shelf varied from (40 km, 40 km, 15 m, 45 days) in winter for $z = 0 - 30$ m (the z scale increased to 25 m in deeper water), to (30 km, 30 km, 15 m, 30 days) for all other seasons. Such an approach may have some utility for the ocean-colour binning problem, but would require that appropriate decorrelation scales be determined for the global ocean throughout the seasons. A useful extension of these purely statistical approaches would be to combine the observations with models. Models can potentially bridge the satellite observations in space and time, and provide meaningful information that observations cannot. In practice, models are deficient in their representation of processes and interactions, and consequently their outputs stray from the observations. If they can be linked to the observations, then models can provide greatly enhanced understanding of biogeochemical cycling, by identifying the nature of the deficiencies and providing clues to improvement, as well as by nudging model variables toward realism. The general approach of data assimilation, as applied to the production of realistic fields, constrains both data and models, and allows for prediction of ocean biogeochemical processes in a hind cast, analysis, or forecast sense.

Assume a "background" state, which can be the results of a previous model run, persistence or climatology, $\mathbf{x}_b(j)$. We wish to derive the new analysis field, $\mathbf{x}_a(j)$ as above, subject to the assimilation or constraints of observational data, $\mathbf{y}(i)$, taken at places and times not necessarily coincident with the grid points, j ,

$$\mathbf{x}_a(j) = \mathbf{x}_b(j) + \frac{\sum_{i=1}^n w(i, j) \{\mathbf{y}(i) - \mathbf{x}_b(i)\}}{\sum_{i=1}^n w(i, j)} \quad (6.6)$$

The $\mathbf{x}_b(i)$ is the "background" field interpolated to the observational points i . The weights, $w(i, j)$, are as defined above or determined by least-squares criteria or optimal interpolation (see below). Note that more weight can be given to the model (or climatology) by setting the weights less than one for $i = j$. Essentially, this is then a weighted average between the background and the observations.

Such an approach has been used for the production of "blended" fields, which couple satellite observations, *in situ* observations, and a relaxation to either persistence (*e.g.* the previous analysis) or climatology as the background field in the above (Reynolds and Smith, 1994; 1995). This has seen limited use for ocean-colour fields, but with new *in situ* observational platforms now available, could be a useful approach. For example, Gregg and Conkright (2001) have produced a blended ocean-colour product from the CZCS data. This analysis assumes that *in situ* data are valid and uses these data directly in the final product. The satellite chlorophyll data are inserted into the final field using Poisson's equation,

$$\nabla^2 C^b = \Psi, \quad (6.7)$$

where C^b is the final blended field of chlorophyll, and Ψ is a forcing term, which is defined to be the Laplacian of the gridded satellite chlorophyll data ($\nabla^2 S$). *In situ* data serve as internal boundary conditions, and are inserted directly into the solution field C^b

$$C_{\text{ibc}} = I, \quad (6.8)$$

where the subscript ibc indicates internal boundary condition, and I is the *in situ* measured value of chlorophyll. Thus, *in situ* data appear un-adjusted in the final blended product. This method assimilates directly on the basis of spatial variability, as inferred from Poisson's equation. The data field is retained in the analyzed field, while the model retains its spatial variability, adjusted for bias by the data field.

Further improvements are possible. Issues with the above include the lack of dynamical constraints (*e.g.* does not have to respect physical laws), that it is not always clear how to specify the weights *a priori*, and that it is not always easy to treat poor observations (*e.g.* adjust weights). Therefore, a more suitable statistical approach is needed, based on some sort of least squares approach. In a general sense, this optimal interpolation can be given as:

$$\mathbf{x}_a(j) = \sum_{i=1}^n [\alpha(i, j) \{y(i)\}], \quad (6.9)$$

where the optimal weights, $\alpha(i, j)$, represent the least-squares estimates of the true weights. In this method, the weight matrix is chosen to minimize the expected error variance of the analyzed field (Daley, 1991). It differs from the spatial analysis method by allowing the covariance between the model and data to determine the error correlation length scale, and from the blended analysis by use of a statistical approach to defining the weights. The weight matrix now represents the error correlations, and is referred to as the error covariance matrix.

Extensions to the above approaches form the basis on which dynamical biogeochemical models and data (both satellite and *in situ*) can be merged to produce forecasts of ocean biogeochemical processes. These forecasts can, in turn,

provide a "background" against which significantly improved analysis fields can be produced that obey physical/biological constraints, and which avoid some sources of aliasing in the resulting product. For example, the adjoint data assimilation method differs considerably from the methods described above. In essence, this method iterates model parameters, boundary and initial conditions, and forcing functions to minimize a cost function in a least-squares sense. This cost function is a measure of the difference between model output and observations over a specified time and space interval. The model is run forward in time to evaluate the cost function, then an adjoint model, which is forced essentially by the deviations between the model and data, is run backward in time to evaluate the gradient of the cost function. An algorithm is applied that determines how the model should be adjusted to reduce the disparities. The adjustments are made and the entire procedure is run again until minimization results.

The advantage of the method is its inherent diagnosis of the model leading to explicit model improvement. The method has been utilized to considerable success in some biological oceanographic applications (*e.g.* McGillicuddy *et al.*, 1998). More advanced methods include the Kalman filter, which is a four-dimensional state-space approach to variational data assimilation. Building on previous efforts to reduce the state space of the Kalman filter (*cf.*, Cane *et al.*, 1996) in linear reduced-gravity models, the so-called SEEK (Singular Evolutive Extended Kalman) filter has been applied to the tropical Pacific Ocean (Gourdeau *et al.*, 2000; Verron *et al.*, 1999). The approach to reducing the computational burden of the Kalman filter consists essentially in approximating the error covariance matrix by a singular low rank matrix, which leads to making corrections only in those directions for which the error is not naturally attenuated by the system. These directions evolve with time according to the model evolution, yielding to the adaptive nature of the filter. The filter is initialized by a method based on Empirical Orthogonal Functions issued from free runs of the model. This reduction in the error covariance matrix avoids the overwhelming burden of computing the temporal evolution of the prediction error with all the degrees of freedom of the full state vector. An excellent example of this approach, and an exceptionally clear description of the methodology, as applied to a physical/biogeochemical model and the SeaWiFS ocean-colour data set, is given in Natvik and Evensen (2003a, b).

6.2 Ocean-colour products for assimilation

Assimilation data differ from forcing data in that they affect the model biogeochemical distributions in an *a posteriori* and non-causal fashion. They improve model results by enforcing convergence and constraint to observed data distri-

butions. It is critical that data assimilation sources have the highest quality, since they will directly affect model results.

There are several satellite-derived biogeochemically-related output products for potential assimilation into models (Table 6.1). Although chlorophyll is the assimilation data set of primary importance, investigations using other MODIS, MERIS, or GLI data products can provide important information in the assimilation model for biogeochemical cycling. For example, MODIS may contain information on coccolithophores and possible cyanobacteria abundances (Table 6.1) that in the future could be used to refine the model distributions in an assimilation capacity. Furthermore, advances in remote sensing algorithms for the detection of other phytoplankton groups, may provide other opportunities for assimilation.

Table 6.1 Potential data sets for biogeochemical model assimilation, their purpose, and possible sources of data.

Variable	Purpose	Candidate Sensor
Chlorophyll concentration	Total phytoplankton distribution	MODIS/SeaWiFS
Coccolith concentration	Coccolithophore distribution	MODIS
Phycoerythrin	Cyanobacteria distribution	MODIS
CDOM absorption coeff.	CDOM distribution	MODIS/MERIS

6.3 Spatial and temporal resolution

A survey of global ocean carbon and biogeochemical models indicates that the state of the science is generally > 1 degree horizontal resolution (Table 6.2). Future improvements in the next 10 years are expected to require 0.1 degree resolution. Coastal models require much greater resolution, and are typically run at about 0.1 degree. However, future coastal models may be expected to require 0.01 degree (1 km) resolution. Requirements for spatial resolution of ocean-colour data products should meet, or exceed, the highest resolution models to provide maximum usefulness for data assimilation. This is presently about 0.1 degrees, based on coastal model requirements, and may be expected to reach 0.01 degrees in 5 to 10 years.

Daily or monthly assimilation requirements are typical for present efforts, although biogeochemical assimilation is fairly immature. However, these temporal frequencies must be considered the minimum requirement for ocean-colour data products. Monthly data can be quite biased, which can affect assimilation. Often a single observation represents an entire monthly mean for a given bin. If this observation occurs at the beginning or end of the month, or before or after a major bloom/recede event, it is a biased representation of the

Table 6.2 Representative sample of global ocean carbon cycle and biogeochemical models currently in use, and their horizontal grid structure and orientation.

Organisation	Resolution (lon. x lat., degrees)	Rectangular
Alfred Wegener Institute (AWI)	5 x 4, 2.5 x 2	yes
Commonwealth Science and Industrial Research Organization (CSIRO)	5.6 x 3.2	yes
Institute for Global Change Research (IGCR)	4 x 4	yes
Institut Pierre Simon Laplace (IPSL)	2 x 1.5 (0.5 Equator)	yes
Lawrence Livermore National Laboratory (LLNL)	4 x 2	yes
Massachusetts Institute of Technology (MIT)	2.8 x 2.8	yes
Max Planck Institut fuer Meterologie (MPIM)	5 x 5	yes
NASA/Global Modeling and Assimilation Office (GMAO)	1.25 x 0.67	yes
National Center for Atmospheric Research	(NCAR) 3.6 x 1.8 (0.8 Equator)	yes
Nansen Environmental and Remote Sensing Center (NERSC)	2 x 3.2 cos(lat)	no
Physics Institute University of Bern (PIUB)	Zonal basin average	yes
Princeton University /Geophysical Fluid Dynamics Laboratory (GFDL)	3.75 x 4.5	yes
Southampton Oceanography Centre (SOC)	2.5 x 3.75	yes
University of Liege (UL)	3 x 3	yes

Data from Dutay *et al.* (2002)

monthly mean. Therefore daily products are required for assimilation. Future assimilation efforts may take advantage of higher temporal frequencies, such as 3 to 6 hourly, if data are available from merged products. This is despite the loss of coverage from such highly refined temporal products. But models are time-stepped, and can potentially utilize ocean-colour data in assimilation at these intervals. Such high frequency assimilation may be advantageous for models, as they will produce less initialization shock and more coherent model-data matches. Disadvantages are associated with stopping model runs more frequently and re-initialization, which will slow model execution. However, we should look toward model improvement as the ultimate goal.

Ocean-colour data product standardization and ease of use are important to

assimilation efforts, since assimilation is a complex effort and model execution is computer-intensive. Regularly-spaced grids, preferentially rectangular in orientation, are most useful for most model activities, and equal-angle grids are preferred. Data products should contain only one variable and “value-added” data formats, such as HDF, net-CDF, GRIB, T62, tend to require more effort by modellers to utilize data sets. Standard IEEE flat-file formats are useable by virtually all operating platforms at the present time, often requiring only simple modifications in executable scripts or software.
