Analysis of ROMS Estimated Posterior Error Utilizing 4DVAR Data Assimilation

Joseph Patrick Horton

Mathematics Department

California Polytechnic State University

San Luis Obispo, California

June 2011

APPROVAL PAGE

TITLE:	Analysis of ROMS Estimated Posterior Error Utilizing 4DVAR Data Assimilation
AUTHORS:	Joseph P Horton
DATE SUBMITTED:	June 2011

____________________________ ____________________________

Senior Project Advisor Signature

____________________________ ____________________________

Mathematics Department Chair Signature

Abstract

The appropriateness of the approximate error calculated by the Regional Ocean Modeling System (ROMS) is analyzed using Four-Dimensional Data Assimilation (4DVAR) performed on a numerical model of the San Luis Obispo Bay. An effective method of sampling data to minimize the actual error associated with the assimilated numerical model is explored by using different data sampling methods. An idealized state of the SLO bay region ("Real Run") is created to be used as the real ocean, then a numerical model of this region is created approximating this Real Run; this is known as the "Simulated State". By taking samples from the Real Run then running 4DVAR on the Simulated State using this input, the exact error of the assimilation step is compared directly with the Real Run using the Assimilated State. Once the exact errors are determined, comparison with the exact error with the estimated error calculated by ROMS to evaluate the appropriateness of 4DVAR on the sample region.

Acknowledgements

I would like to thank Dr. Paul Choboter for helping patiently with this research project by reiterating carefully the complex data assimilation methods and creating the Alternate State simulation to utilize for the Assimilation State. Thanks go to the College Based Fee which provided funding to give Math Undergraduates a lounge full with computers accessible to me at any hour of the day. And lastly, thank you to the Cal Poly Computer Science Department for granting temporary access to the parallel computing power necessary for numerical fluid dynamics. This project could not have been completed without this aid and access.

Contents

List of Figures

I. Introduction

Numerical models of the ocean are used for a variety of reasons from better weather forecast systems and climate models to further understanding upwelling currents and hurricane formation. In 2009 the funding was awarded for the purchase of an autonomous underwater vehicle (AUV) to research ocean dynamics with specific interest in the San Luis Obispo (SLO) bay. This AUV was equipped to gather information about temperature to provide researchers with information on the state of the waters in the SLO bay. This information is then used by researchers to try to better approximate the actual state of the ocean using a method known as data assimilation.

When performing simulations, you do not have access to the true state of the real ocean, making it necessary to use a simulated model of the ocean with imperfect conditions to be used in replacement of the true state. After collecting data from the real ocean, this simulated model is used with the data to perform the assimilation step, where the model is adjusted dependent on the collected data to obtain an Assimilated Model closer to the true state of the ocean.

The Regional Ocean Modeling System (ROMS, http://www.myroms.org) is utilized for this type of analysis by researchers. ROMS is a free-surface, primitive equation ocean model which is utilized for a variety of simulations. ROMS evolves velocity, temperature, salinity and sea surface height in time using a finite-difference method of the equations of motion. [HAB08] Sophisticated numerical analysis techniques are implemented including a split-explicit time-stepping method treating the fast barotropic (2D) and slow baroclinic (3D) modes separately for improved efficiency. [SM05] ROMS includes a set of tools based on four dimensional data assimilation methods to improve the accuracy of simulations by including observational data. [LMA07]

Using ROMS, researchers perform data assimilation by beginning with an idealized version of the ocean's state produced through numerical simulation. Then, using a method known as four-dimensional variational analysis (4DVAR), the data collected at specific positions and times by the AUV will be used to better approximate the actual state of the ocean. During 4DVAR the expectations of the variance is minimized in a type of multidimensional regression. Since this state cannot be perfectly accurate, ROMS calculates the approximate posterior error covariance of the model to determine the approximate error at each coordinate within the simulation.

This project focuses on the appropriateness of the approximate error calculated by ROMS specific to the SLO bay, with specific interest towards determining effective methods of sampling data to minimize the actual error associated with the assimilated numerical model. To explore this topic we created an idealized state of the SLO bay region ("Real Run"), then created a numerical model of the region approximating this Real Run in the same way the numerical model is created during data assimilation on the real ocean ("Simulated State"). By sampling from the Real Run simulation we can calculate the exact error of the assimilation step by comparing directly with the Real Run using our Simulated State simulation. Once the exact errors are determined we can then compare the exact error with the estimated error determined by ROMS to determine the appropriateness of this method on the specific sample region.

To create the Real Run simulation we began simulating from the calendar day October 1st and simulate to October 31st. Using the information from October 31st as the initial conditions for the Real Run we simulated from October 31st 12:00 AM to November 1st 12:00 AM forced by a set of wind conditions. The Simulation Run utilizes the same wind conditions as the Real Run, but with all initial velocities set to zero. The initial temperature and salinity were set to be horizontally uniform, depending only on depth. By using the same wind conditions as the Real Run, we allow the model to develop similarly to the Real Run.

During the Assimilated Run, we take a sample of data points from the Real Run to allow our model to have a utilizable set of data. We can choose these points in any fashion, however in a real life implementation you are restricted to having the AUV travel to extract the data; often times researchers will only sample a handful of points and deem this appropriate for dramatically affecting the system's accuracy. In our test we take a large number of points from the Real Run model all at the same time in an effort to determine if the model is greatly effected by sampling a large amount of data. Later, the sample regions will be more feasible for real life and used in the same way to determine the system's reaction to a small amount of data being fed into the system.

By treating this Real Run as the actual state of the ocean, we can create a set of synthetic observations to be used in ROMS's assimilation step. During the assimilation step, we use ROMS to read in the data from the large synthetic observations at 12:00 PM October 31st to attempt to get our Assimilation Run approximating the Real Run. After reading in this data, the actual error calculated by the difference between temperature at each point in the Assimilated Run and the Real Run is compared with the output error estimate calculated by ROMS.

II. 4DVAR

Four-dimensional variational analysis (4DVAR) is an assimilation technique used to examine observations that are distributed in time and space. 4DVAR assimilation uses observational data and numerical forecast data to minimize the cost function

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_f)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_f) + (H[\mathrm{x}] - \mathbf{y})^T \mathbf{R}^{-1} (H[\mathrm{x}] - \mathbf{y}) \tag{1}$$

where $\mathbf{x}$ and $\mathbf{x_f}$ represent all the variables predicted by ROMS including temperature, salinity, u and v velocity components, and sea surface height at all the grid points throughout the numerical domain. Also included in the equation: $\mathbf{y}$ is the observed values, H is the analysis estimate of the observed values, $\mathbf{R}$ is the Observation error covariance, and $\mathbf{B}$ is the forecast error covariance. [Kep07]

To minimize the cost function $J(\mathbf{x})$ (1) we first must discuss the Best Linear Unbiased Estimate (BLUE). Letting y_1 and y_2 be observations of a true state x_t we can express y_1 and y_2 as $y_1 = x_t + \epsilon_1$ and $y_2 = x_t + \epsilon_2$. From probability theory we know $E(\epsilon_1^2) = \sigma_1^2$ and $E(\epsilon_2^2) = \sigma_2^2$ where the operation $E(x)$ is the expectation of a variable x and σ_i^2 is the variance of the ith observation. It can be shown that the best estimate x_a of x_t is

$$x_a = \frac{\sigma_2^2 y_1 + \sigma_1^2 y_2}{\sigma_1^2 + \sigma_2^2} \tag{2}$$

This is a linear combination of observations such that x_a is an unbiased estimator of x_t which minimizes σ_a^2 also known as an efficient estimator. [BC99] The same estimate is found by minimizing

$$J(x) = \frac{(x - y_1)^2}{\sigma_1^2} + \frac{(x - y_2)^2}{\sigma_2^2} \tag{3}$$

Hence, for Gaussian Errors, the BLUE is the maximum likelihood estimate (MLE) of x_t. For a generalized $\mathbf{y} = (y_1, y_2, ..., y_n)^T$

$$J(x) = (x - \mathbf{y})^T \mathbf{P}^{-1} (x - \mathbf{y}) \tag{4}$$

where $\mathbf{P}$ is the error covariance matrix of $\mathbf{y}$. [Kep07]

Since $J(\mathbf{x})$ (1) is in quadratic form, in order to minimize we need the gradient

$$\triangledown J(\mathbf{x}) = 2\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_f) + 2\mathbf{H}^T \mathbf{R}^{-1} (H[\mathrm{x}] - \mathbf{y}) \tag{5}$$

to be equal to 0. [BC99] Here, $\mathbf{H}$ is the Jacobian (tangent linear) of H.

To solve $\nabla J(\mathbf{x}) = 0$ directly you would need to manipulate large matrices with a nonlinear H. Instead, we will use iterative minimization to find the full three-dimensional state of the ocean using observation to constrain a bad IC model of the ocean. With the modeling assumption that each variable temperature, salinity, etc is independent, you can represent $\mathbf{B}$ with a diagonal matrix, lessening the computational load of the iterative process.

By iterating the model from time one (t_1) to any number n time stamps, we can add terms to (1) to yield

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_f)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_f) + \sum_{\mathbf{i}=0}^{n} (H_i[\mathbf{x_i}] - \mathbf{y_i})^T \mathbf{R}^{-1} (H_i[\mathbf{x_i}] - \mathbf{y_i}). \tag{6}$$

where $H_i[\mathbf{x}_i]$ is the model applied to the state $\mathbf{x}_i$ and $\mathbf{y}_i$ is the observed values at the time t_i. Further explanation to the extension of (1) to observations at an arbitrary set of times can be found in [Kep07].

4DVAR is capable of handling a large number of observations at the same time while also handling nonlinear observations from H. This method creates a three-dimensional regression model in time to allow for more accurate numerical methods of complex dynamical systems. [Kep07]

III. ROMS

The Regional Ocean Modeling System (ROMS) [http://www.myroms.org] was utilized during the project to perform all of the data models. Access to this was through a secure shell server utilizing bash command navigation and file transfer techniques. Among the files necessary for creating distinct runs were the initial conditions, the wind conditions, bathymetry of the SLO bay [http://topex.ucsd.edu/WWW html/srtm30 plus.html] and the coastline structure of the bay [http://www.ngdc.noaa.gov/mgg/coast/] (Figure 1). We used a region of size 146 by 194 by 20, where 20 represents depth. To gather the initial conditions for the Real Run, we first performed a numerical run spanning from 12:00 AM October 1, 2009 to 12:00 AM October 31, 2009, using the state at 12:00AM October 31st as the initial state for our as the initial conditions of the Real Run (Figure 2). The split-explicit time-stepping was set for an internal time step at sixty seconds, while the surface fast time step was set to one second.

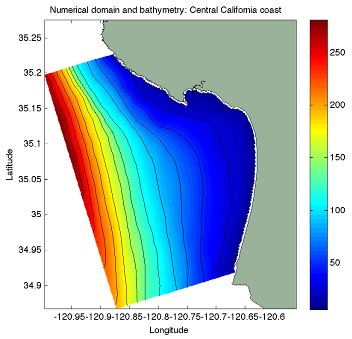

Figure 1. The SLO coastline structure is the gray mass. The bathymetry contours approximately every 20 meters show the shape of the ocean floor, gradually rising as the coastline is approached.

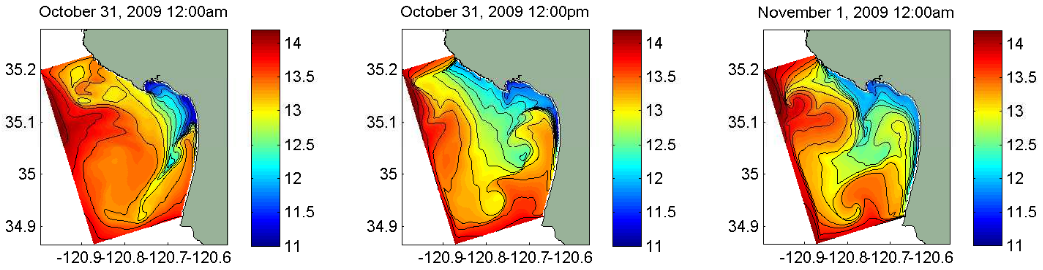

Figure 2. Time Progression of Surface Temperature for Real Run

For the Assimilation Run, the variables were assumed to be spatially uniform at the initial time with temperature and salinity depending on depth. This idealized model is consistent with methods used in a real assimilation with the real ocean, providing the Assimilated Run with an initial starting state (Figure 3). As the model is run forward in time, information is fed into the system from the sample in the first simulation at twelve hours into the model, corresponding to the moment the data was taken from the Real Run.

Then, utilizing ROMS 4DVAR assimilation technique we sampled data at the time stamp coordinated with half of the runtime (12:00 PM October 31, 2009) in a sample region size 4 by 7 taken at the surface and near the center of the sample region. We will refer to this sample as the Medium Sample (Figure 4a). By performing this we can determine if the system has a quick and effective response to the assimilation technique as expected and desired.

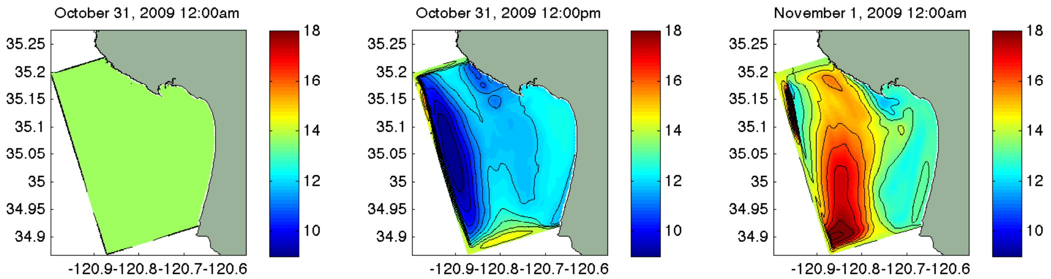

FIGURE 3. Time progression of surface temperature of Assimilation Run.

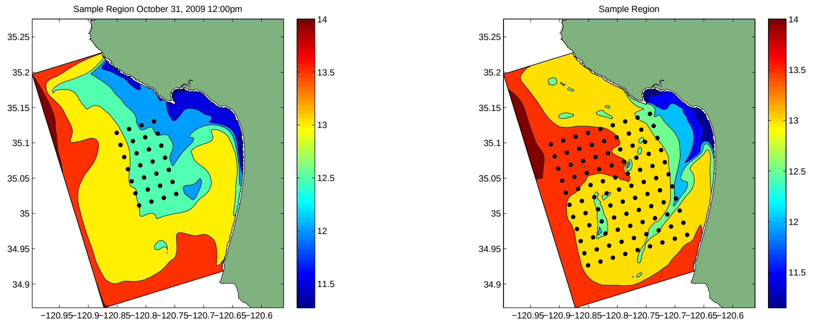

FIGURE 4. Graphic Display of Sample Region at Real Solution Surface of Medium and Exhaustive sample respectively. Each black dot represents the location of one sample, the contours represent the Real Run at October 1, 12:00pm and October 1, 2:00am respective to the time at which each sample was taken.

Another set of data we call the Exhaustive Run was analyzed using a sample at two hours into the Reality Run, using a region size 9 by 11 at each even depth 2 through 20, where 20 represents the surface of the model (Figure 4b). By performing this run we tried to ensure that the model would quickly converge to the solution by simultaneously increasing our sample greatly and put in the data very early on to allow the process of assimilation to occur more effectively.

IV. Comparisons and Analysis

We first compared the actual differences of posterior error visually using Figures 5 and 6. For the Medium case there are three regions of high errors located at the bottom left , the upper right corner of the region and a thin region of high error magnitude located at the upper left region; the upper left region is difficult to distinguish in Figure 5 due to resolution. This output appears strange because the actual error at the sample region is relatively high, while the error away from the sample region with exception of the two regions of high error are much lower. In the Exhaustive Run, the real error at the surface appears much higher than that of the Medium Run with a lower error through the center of the region.

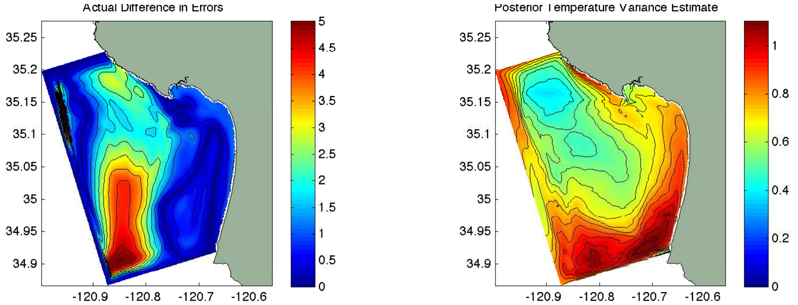

Figure 5. Posterior error estimation of surface temperature on November 1, 2009 12:00 AM with Medium Sample Run

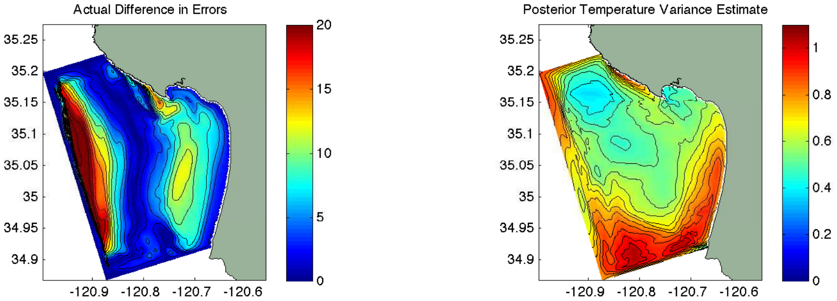

Figure 6. Posterior error estimation of surface temperature on November 1, 2009 12:00 AM with Exhaustive Sample Run

Utilizing MATLAB, we calculated the relevant error statistics for both models. Letting T^R_{ijk} be the temperature sampled from the Real Run at time 24 hours, and let T^A_{ijk} be the temperature sampled from the Assimilation Run at time 24 hours, we calculate the total sum of the real error term at just the surface (SERR) (7a), the total sum error over the whole model (ERR) (7b), and the correlation coefficient between the Real Run and the Assimilated Run (CORR).

$$(7) \qquad SERR = \sum_{i=1}^{146}\sum_{j=1}^{194} \left|T^A_{ij20} - T^R_{ij20}\right| \qquad ERR = \sum_{i=1}^{146}\sum_{j=1}^{194}\sum_{k=1}^{20} \left|T^A_{ijk} - T^R_{ijk}\right|$$

These values reveal that the Medium Sample Run was superior in every category compared to the Exhaustive Sample Run (Table 1). These results are surprising because it was expected that the exhaustive case would lead to a must faster convergence to the real version of the solution yet the Medium Sample Run yielded much less error for both the surface and globally. The dramatic difference in error terms could potentially have occurred because our sampling method in the Exhaustive Run captured data at every depth of the model rather than just the surface.

Table 1. Error Statistics of Assimilation Run

Sample Method	SERR	ERR	CORR
Medium	3.3030e+04	6.9051e+05	0.9646
Exhaustive	1.3786e+05	3.5439e+06	0.5480

The actual cost function (6) was calculated after each inner loop step of the Assimilated Run and are not surprising in relation with the output error values and correlation coefficient (Table 2). We can see that the cost function of the Medium Run converge after about 4 inner loops (Figure 7a), while the cost function of the Exhaustive Run jumps around and didn't appear to converge after ten iterations (Figure 7b).

Table 2. Cost Function Calculations

Inner Loop Count	Medium (1.0e+02)	Exhaustive (1.0e+05)
0	4.906031267238034	0.264562338539813
1	0.022404474313908	0.045987511404171
2	0.021143185201053	0.799819074394525
3	0.021139325450831	2.834135628923240
4	0.021139319099248	2.676981427144457
5	0.021139319099248	0.289690154221484
6	0.021139319099248	0.167254342008194
7	0.021139319099248	0.104937210682415
8	0.021139319099248	0.070209748663164
9	0.021139319099248	0.321546971696976
10	0.021139319099248	0.074190667391438

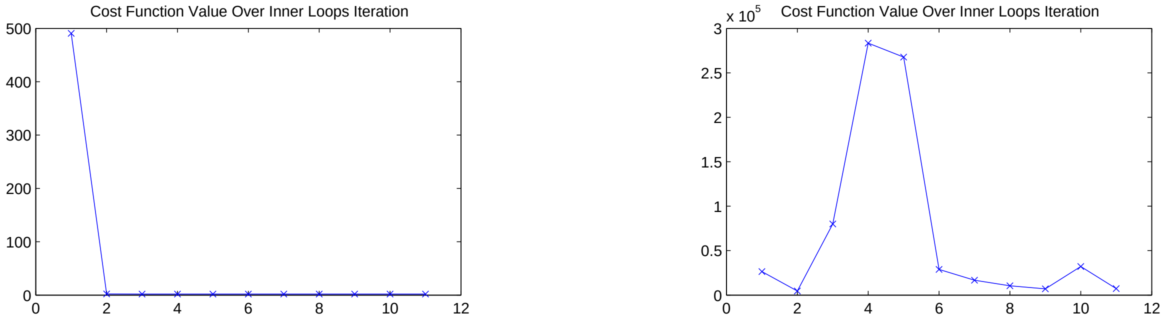

Figure 7. Cost Function of Medium and Exhaustive sample run respectively

V. Concluding Remarks

This project explores the effectiveness of a 4DVAR technique being applied to a three-dimensional dynamical model by comparing the actual error with that of the estimated error provided by ROMS. The results of the actual errors when compared to the estimated error variance indicates that the error provided by ROMS is an over-idealized version of the error, underestimating the error of the model.

The results of the multiple sample sets indicate that the error of the model is affected drastically by different sampling methods. The Medium Sample of utilizing the 4DVAR method achieved a final correlation coefficient of 0.9646 and total error of 6.9051e+05. In comparison, the Exhaustive Run's correlation coefficient of 0.5480 and total error of 3.5439e+06 indicating that more complete sampling does not seem to improve the total error of such a dynamic system. Although the cost function for the Exhaustive Run doesn't seem to converge after 10 iterations, the initial condition of the post assimilation Exhaustive Run interprets the general behavior of surface temperature better than the Medium Run, indicative that more data being fed into the system leads to better general behavioral results (Figure 8).

This project has potential for further study in testing effective sampling methods; it would be of interest to take samples of the data based upon factors such as greatest variability within the model, regions of greatest range of temperature at the surface, or by taking a single point early in the assimilation process. Further exploration can be done by changing the method of data selection to be more similar to an AUV to determine if feeding information into the system at different times could improve the Assimilated Run more effectively than other sampling techniques. Since we had

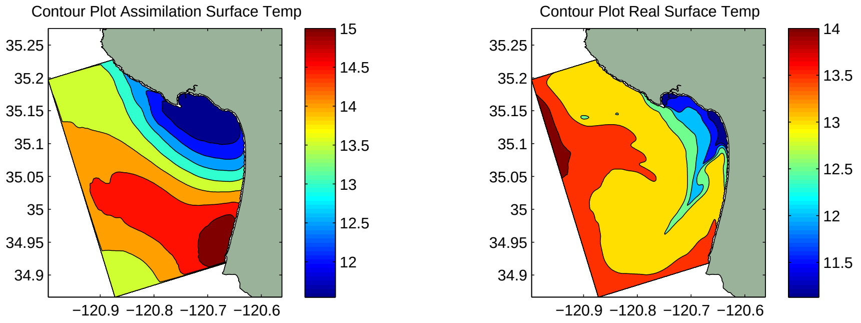

Figure 8. Initial Surface Temperature Conditions of Exhaustive Assimilation Run and Reality Run

such a disparity in error between the Exhaustive and Medium Runs, it would be of interest also to limit the number of changes we make between the two sampling techniques.

References

[BC99] F. Bouttier and P. Courtier. Data Assimilation Concepts and Methods. *in: Meteorological Training Course Lecture Series*, page 22, March 1999.

[HAB08] D.B. Haidvogel, H. Arango, W.P. Budgell, B.D. Cornuelle, E. Curchitser, E. Di Lorenzo, K. Fennel, W.R. Geyer, A.J. Hermann, L. Lanerolle, J. Levin, J.C. McWilliams, A.J. Miller, A.M. Moore, T.M. Powell, A.F. Shchepetkin, C.R. Sherwood, R.P. Signell, J.C. Warner, and J. Wilkin. Ocean Foreccasting in terrain-following cocordinates: Formulation and skill assessment of the Regional Ocean Modeling System. *J. Computational Physics*, 227:2595–3624, 2008.

[Kep07] J. D. Kepert. Maths at work in meteorology. In N. Heaps, editor, *Gazette of the Austrailian Mathematical Society*, pages 150–155. Australian Mathematical Society, 2007.

[LMA07] Emanuele Di Lorenzo, Andrew M. Moore, Hernan G. Arango, Bruce D. Cornuelle, Arthur J. Miller, Brian Powell, Boon S. Chua, and Andrew F. Bennett. Weak and strong constraint data assimilation in the inverse Regional Ocean Modeling System (roms): Develpment and application for a boroclinic coastal upwelling system. *Ocean Modelling*, 16:160–187, 2007.

[SM05] A.F. Schepetkin and J.C. McWilliams. The regional oceanic modeling system (roms): a split-explicit, free-surface, topography following coordinate oceanic model. *Ocean Modeling*, 9:347–404, 2005.