ENSEMBLE DATA ASSIMILATION SIMULATION EXPERIMENTS FOR THE COASTAL OCEAN: TOWARDS AUTOMATIC LOCALIZATION

R. Hoffman[*1]*, J. Poterjoy*[2]*, S. V. Vinogradov*[1]*, and S. M. Leidner*[1]

[1]Atmospheric and Environmental Research, Inc., Lexington, MA;
[2]Millersville University, Millersville PA.

ABSTRACT

A coastal ocean data assimilation system was tested in simulation earlier for sensitivity to the different types of observational data. The system couples an advanced ensemble Kalman filter algorithm to a detailed and sophisticated primitive equations coastal ocean model. It is found that assimilating only one type of data, say temperature, greatly slows down the approach to asymptotic behavior of the analysis of the other variables. Assimilating temperature alone does not help to infer salinity and vice versa.

We examined correlations between simulated state variables on various locations and depths within the baseline experiment. Distributions of correlation coefficients surrounding an analysis point could be used to determine the optimal localization domain for each particular relationship. Given the large amount of dynamical and bathymetric variability within this model domain, correlation structures of mixed shapes and sizes were observed. In many instances, the parameterized localization actually used in the baseline experiment was either too small or too large to capture the actual correlations. Correlations between temperature and salinity were found to be very small, consistent with the results of the data impact experiments.

Index Terms— Kalman filtering, data impact, data assimilation, coastal ocean model, localization

1. INTRODUCTION

A coastal ocean data assimilation system is being developed. The goal is to combine large and disparate datasets with ocean numerical models, producing accurate analyses, forecasts, and respective uncertainty estimates for any littoral region. A modular interface combines the Estuarine and Coastal Ocean Model (ECOM) and the Local Ensemble Transform Kalman Filter (LETKF) into a highly scalable, portable and efficient ocean data assimilation system. The ECOM is a state-of-the-art, three-dimensional, hydrodynamic ocean model developed as a derivative of the Princeton Ocean Model [1]. The LETKF, a recent adaptation of ensemble Kalman filtering techniques, works particularly well for very large non-linear dynamical systems in both sparse and dense data regimes, and provides efficient algorithms for error estimation and quality control [2]. In simulation experiments for highly idealized data distributions in the New York Harbor Observing and Prediction System (NYHOPS) the filter quickly converges, eliminating bias and greatly reducing rms errors [3]. This behavior is robust to changes in ensemble size, data coverage, and data error. The work of [3] was extended to subsets of observed variables [4]. The experiments compared are: assimilate all variables (h, T, u, v, S), only the free surface elevation h, only the temperature T, only the salinity S, and the combination (h, T, S). Only the baseline experiment assimilates currents (u, v). Observations in each case are generated by adding random errors to the "truth" and selecting each datum randomly.

Ensemble data assimilation provides a good way of determining the background error correlations. Since distant correlations are expected to be relatively insignificant, the LETKF limits the data region considered in a process called localization. That is only observations from a local volume are used [5]. Questions still remain concerning how the localization size varies for different models and for different regions within a model domain. How strong is the relationship between the errors of a single variable and all other variables at different points and levels throughout a model domain? How should the localization volume be tuned?

To answer these questions as well as to examine the T–S correlations in an effort to explain the results of the data impact experiments, we examined the actual correlations within the forecast and analysis ensembles of the baseline experiments of [3]. Patterns of sample correlations show large variability depending on the variables correlated and the position within the model domain. Distributions of correlation coefficients (r) between a single forecasted or analyzed point variable and the remaining field were used for visualization. Using larger ensembles, averaging in time, and eliminating the start of the data assimilation experiment all help to control spurious correlations in remote locations.

*Presenting author; ross.n.hoffman@aer.com

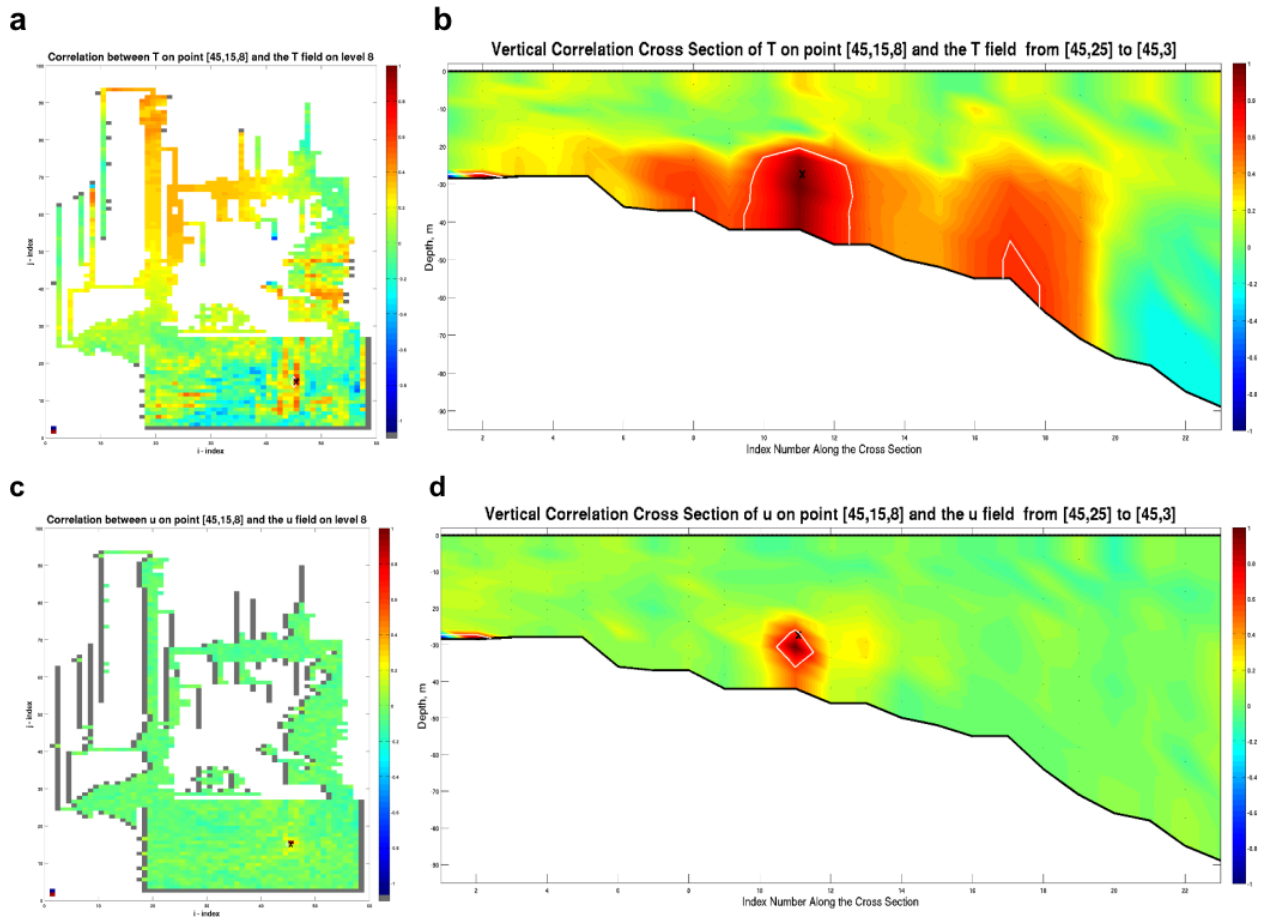

Fig. 1. The horizontal correlation between the point marked by $\times$, and the remaining model domain are illustrated for (a) T–T and (c) u–u correlations. Regions which are colored red represent large positive correlations, where blue shades indicate strong negative correlations and a white contour is plotted for $|r| = 0.6$. The key point is at location 45, 15, level 8. The vertical correlation structure is shown along a cross section with $i = 45$ extending from the coast to the southern open boundary for (b) T–T and (d) u–u correlations.

2. CORRELATIONS STRUCTURES

The NYHOPS domain is ideal for this particular type of study, where the relationships between variables at an assortment of locations in a diverse model domain are expected to vary substantially. The correlation structures differ greatly depending on location and variable type. For visualization, a particular (or key) variable at a particular (or key) location is correlated with all variables at all grid points and levels in the model domain. Then horizontal or vertical slices through the domain of correlations of a chosen variable with the key variable are plotted. Typically, correlations greater than 0.6 are considered significant. As examples, Fig. 1 shows T–T and u–u correlations calculated for the last time step of the data assimilation experiment. In this case, typical of locations in the open ocean, correlations are strong at grid points nearest to the key location and gradually taper off with distance. Length scales for T–T correlations are much larger than for u–u correlations. We find for correlation of a variable with itself length scales are largest for h, then S, T, and (u, v) (cf. Fig. 1 to Fig. 2). Correlations shown here are calculated from the analysis ensemble. Correlations were calculated for both ECOM analyses and backgrounds (forecasts). Differences between these are large at several isolated locations for the first analysis time, but then quickly decay with time. By the eighth analysis (at the end of day 1) the two results have converged and a strong agreement exists between analysis and background correlations.

The ensemble Kalman filter uses a limited sample to estimate a large number of correlations. Some of these correlations are expected to appear to be significant, but are in fact spurious. Increasing the sample, reduces this problem. For example, Fig. 2 plots S–S correlations for ensemble size $k = 16$ and 64. (Except for this plot, all correlations displayed here are calculated from the $k = 64$ experiment.)

Another way to reduce spurious correlations is to average

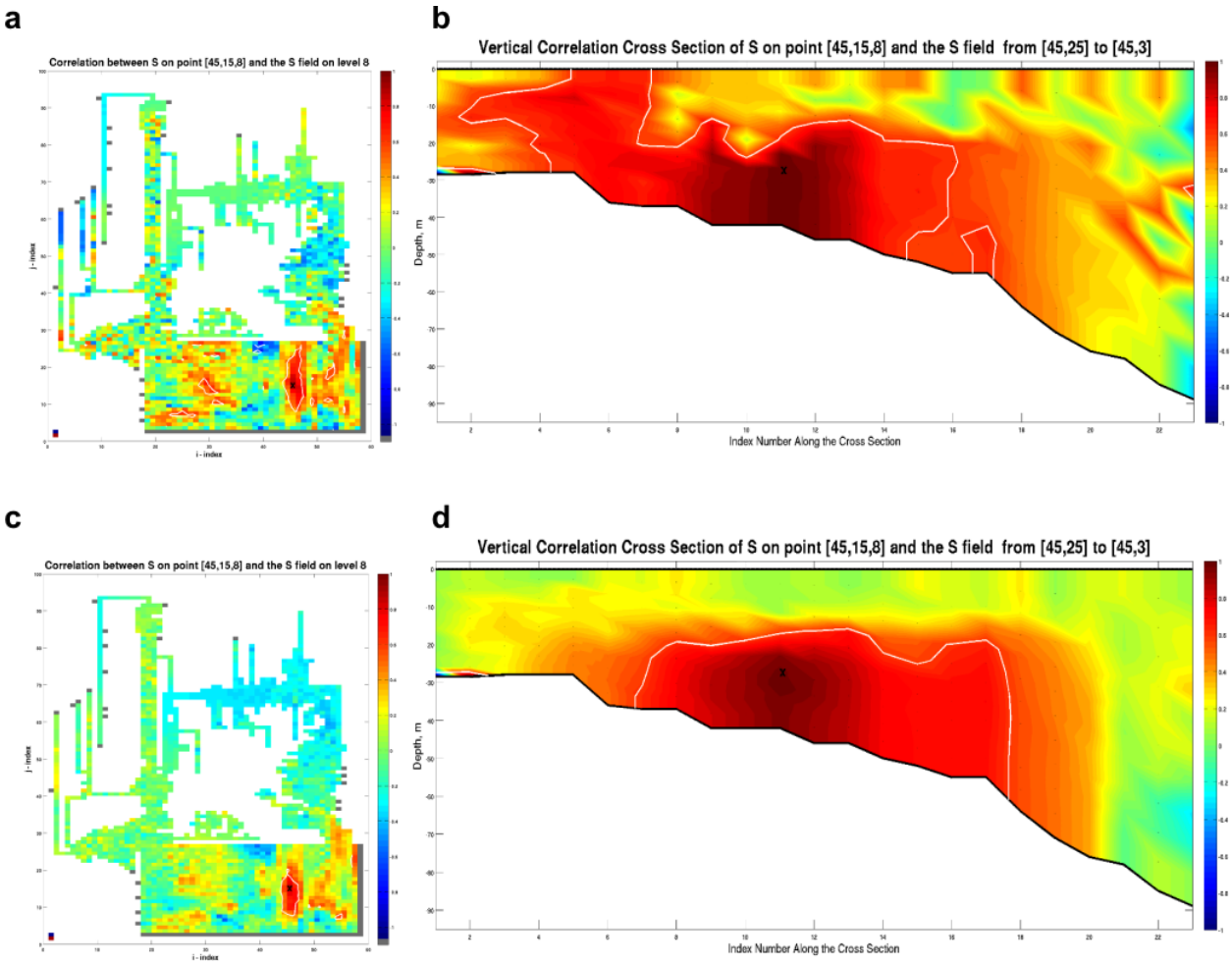

Fig. 2. As in Fig. 1 but for S–S correlations for $k = 16$ (top) and $k = 64$ (bottom). The key point is again located at 45, 15, level 8 (marked by the $\times$).

the correlations over time (*i.e.*, over the different analyses, once the data assimilation has reached asymptotic behavious). Averaging in this way smooths out transient features in the correlations, but can provide a much cleaner estimate of the localization volume. First we note that spurious correlations diminish with time as the LETKF analyses converge towards the true system state. To illustrate this, Fig. 3 shows typical analysis ensemble correlation at 12 and 96 hours since the start of the experiment. As mentioned before, little change is observed in the correlations by day 3; therefore, correlations averaged over the last two days of the experiment can provide reliable and accurate estimates of the localization volumes.

3. DATA TYPE IMPACT EXPERIMENTS

Figure 4 (reproduced from [4]) summarizes the results of the impact experiments described above. Each panel in the figure plots the vertical profile of a single statistic (bias, error, or spread) for a single variable (T or S) during the last half of the assimilation experiments when asymptotic behavior is obtained. A different symbol and color is used for each experiment, as well as for the free running forecast (FRF). The baseline experiment (denoted "all") and the FRF are from [3]. Examination of the figure shows that S errors are large unless S is observed and similarly for T, indicating that the cross correlations between forecast errors of T and S are small, in spite of the fact that advection is expected to be critical to the evolution of both. The error statistics show given observations of one variables, T or S, then observing additional variables does not improve the analysis of the first variable, although the ensemble spread, which may be considered an estimate of the analysis error, is slightly smaller when more variables are observed. We note that the T bias of the FRF is large, while the S bias is small. These last findings reflect the biases of the initial ensemble.

To better understand the results of Fig. 4 the spatial distribution of the errors was examined [4]. Not surprisingly in the highly heterogeneous New York Harbor there are a vari-

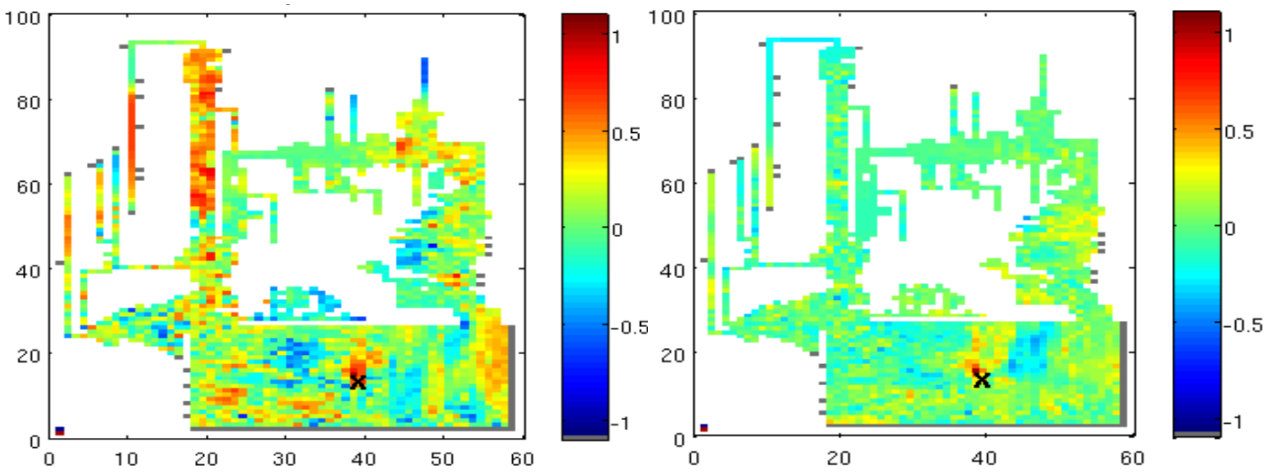

Fig. 3. The evolution of a S–S sample correlation field over time. The key point is located at 38, 15, level 5 (marked by $\times$). Correlations are shown at the end of 12 hours (left) and 96 hours (right) of assimilation.

ety of flow regimes and therefore the data assimilation system responds to different treatments in a variety of ways. Here we examine the cross-correlations between T and S in the data assimilation of the $k = 64$ experiment that used all data.

For nearly all regions within the model domain, the analysis ensemble cross-correlations between T and S at an analysis point are quite minimal. For a key point in the vast open water regions of the New York Bight, the cross-correlation structure reaches barely significant values over a fairly large scale area, but not centered at the key analysis grid point (Fig. 5). Note that the cross section of Fig. 5 runs directly through the region of maximum correlations, but not through the key location. The same phenomenon is present within the Long Island Sound and along coastal regions. However, this type of correlation structure is probably spurious and should not be considered significant.

4. CONCLUDING REMARKS

Ensemble data assimilation provides a promising path for making use of remotely sensed ocean data such as sea surface temperature, ocean color, turbidity, surface currents, free surface elevation, and sea surface salinity. As a practical matter, the ensemble approach provides the best way for observations of one variable to affect the analysis of other correlated variables, either collocated or nearby, at the same level or through the depth of the water column. With many sub-models available in the ECOM for biogeochemistry, sediment transport, water quality, waves, and particle tracking, there are opportunities to extend the assimilation to non-standard data such as ocean color and turbidity, chemical tracers, wave energy, and locations of drifting buoys and autonomous underwater vehicles. These opportunities exist because the LETKF method is completely general in the sense that when the observation errors can be assumed to be Gaussian, any observation of a physical parameter that has a known functional dependence on the variables of the dynamical model, can potentially be usefully assimilated.

Since spurious correlations tend to diminish after 1 to 2 days of simulation, results observed during days 3 and 4 of this experiment were taken to be a good estimate of true relationships between variables. Given the large amount of dynamical and bathymetric variability within this model domain, correlation structures of mixed shapes and sizes were observed. Correlations differ greatly for different variables and for different locations within the NYHOPS domain. In general, correlation length scales substantially decrease from h to S to T to (u, v). In the Hudson River, correlation structures tend to be relatively compact. In the case of u and v only correlations between the analysis point and one or two nearby grid points are significant for this part of the domain. The dynamics of this region (*i.e.*, shallow, fast moving water) is most likely what causes these structures. On the other side of the spectrum, correlations at points within small bay regions of the model domain are rather large. Cross-correlations between variables are much more complicated.

In many instances, the parameterized localization domain was either too small or too large to capture the actual correlations. In previous experiments the LETKF localization radius is set to two horizontal grid lengths and one or two vertical levels. This appears to be deficient for many cases. The accuracy and efficiency of this prediction system can potentially be improved through larger localization volumes for analyses of variables with large scale correlations (S, T, h, especially in the open ocean), and smaller volumes for cases involving compact correlation structures (u, v, especially in rivers).

Results from this study provide incentive to pursue an automated solution to optimal localization within the LETKF/ECOM system that tailors a unique localization volume for each analysis region. Since we found a wide

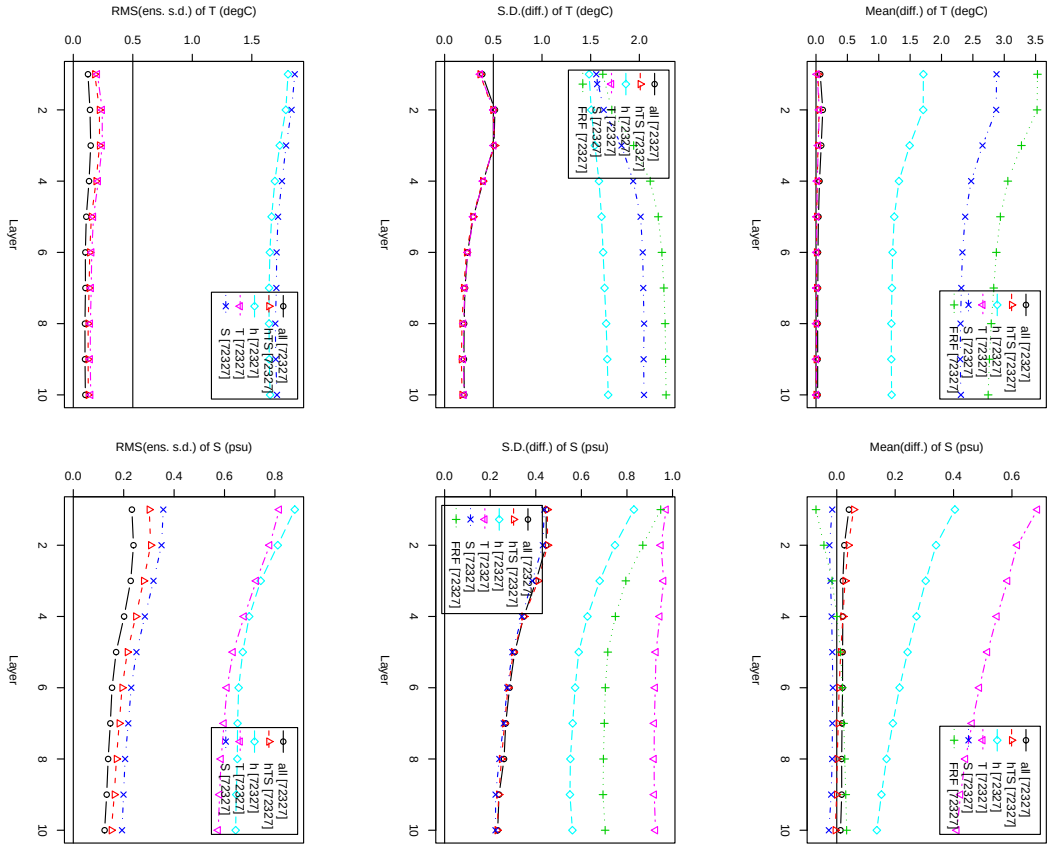

Fig. 4. The vertical profiles for temperature (top) and salinity (bottom) bias (right), analysis errors (middle), and ensemble spread (left) in the data type impact experiments. The color codes are black for all, red for (h, T, S), aqua for H, magenta for T, and blue for S. Units are °C and psu. The expected observational error for T of 0.5 °C is plotted on two of the panels. The expected observational errors for S is 1 psu. Number in brackets in the legends are all equal to 72327 and give the average number of values used to calculate the statistic plotted at each point in the curve.

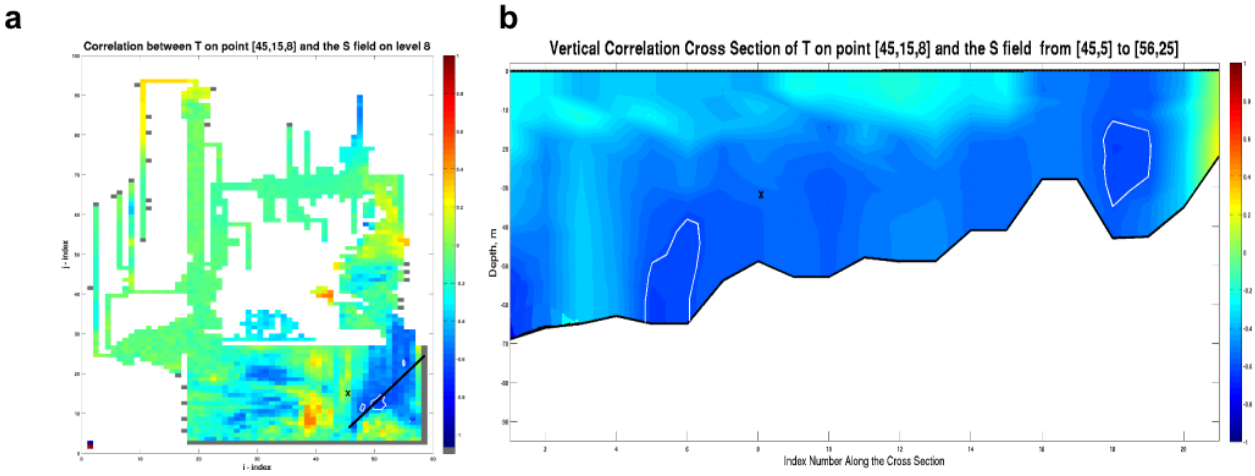

Fig. 5. The ensemble cross-correlations between S and T at grid point 45, 15, level 8 (marked by $\times$) are averaged over days 3 and 4 of the simulation. In the horizontal correlation field (a), the line passing diagonally through the open ocean region shows the location of the vertical cross section of the correlations (b).

variability in correlation structures, it follows that the correct localization should also be variable. If successful, this method can be applied to a variety of other applications, including non-traditional variables where correlation distributions are too complicated to resolve with simple *a priori* localization volumes.

5. AKNOWLEDGMENTS

The work of the one author, J. Poterjoy, was supported by the National Science Foundation under the University of Oklahoma REU program, Grant No. ATM-0648566.

6. REFERENCES

[1] A. F. Blumberg, L. A. Khan, and J. P. St. John, "Three-dimensional hydrodynamic simulations of the New York Harbor, Long Island Sound and the New York Bight," *J. Hydrologic Eng.*, vol. 125, pp. 799–816, 1999.

[2] Istvan Szunyogh, Eric J. Kostelich, Gyorgyi Gyarmati, Eugenia Kalnay, Brian R. Hunt, Edward Ott, Elizabeth Satterfield, and James A. Yorke, "A local ensemble transform Kalman filter data assimilation system for the NCEP global model," *Tellus A*, vol. 60, no. 1, pp. 113–130, 2008, doi:10.1111/j.1600-0870.2007.00274.x.

[3] Ross N. Hoffman, Rui M. Ponte, Eric J. Kostelich, Alan Blumberg, Istvan Szunyogh, Sergey Vinogradov, and John M. Henderson, "A simulation study using a local ensemble transform Kalman filter for data assimilation in New York Harbor," *J. Atmos. Oceanic Technol.*, vol. 25, no. 9, pp. 1638–1656, Sept. 2008.

[4] R. Hoffman, R. Ponte, E. Kostelich, A. Blumberg, I. Szunyogh, and S. Vinogradov, "Ensemble data assimilation simulation experiments for the coastal ocean: Impact of different observed variables," in *Proc. Int. Geoscience and Remote Sensing Symp. (IGARSS)*, Boston, Massachusetts, 6-11 July 2008, IEEE, New York.

[5] Edward Ott, Brian R. Hunt, Istvan Szunyogh, Aleksey V. Zimin, Eric J. Kostelich, Matteo Corazza, Eugenia Kalnay, D. J. Patil, and James A. Yorke, "A local ensemble Kalman filter for atmospheric data assimilation," *Tellus A*, vol. 56, no. 5, pp. 415–428, Oct. 2004.