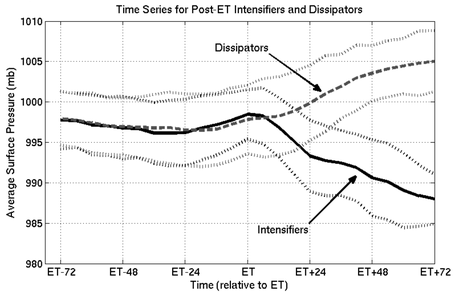
Machine learning techniques to analyze extra-tropical transition

Israel Vaughn, J. Scott Tyo, and Elizabeth Ritchie ivaughn@optics.arizona.edu University of Arizona

Introduction

Extratropical cyclones in the Western Pacific cause a great deal of damage annually. Generally, a tropical cyclone (TC) will track southerly into the mid-latitudes, where it interacts with mid-latitude weather, and subsequently re-intensifies about 51% of the time. When this occurs, the resulting extratropical cyclone is asymmetric, and has a large area of effect. The transition itself, extratropical transition (ET), and the subsequent dissipation or intensification of the extratropical cyclone is difficult to predict (in the figure below there is no difference in surface pressure prior to ET).



This has led many researchers to hypothesize that the transition is a chaotic process. Our current research involves using machine learning methods to both classify (and thus predict) and elucidate the underlying physical features of dissipating vs. intensifying storms. Strong prediction performance implies that ET is *not* chaotic.

Classification

Support vector machines (SVMs) project the data into an infinite dimensional (Hilbert) space, and then find the "best" hyperplane that separates the classes in the infinite space. The hyperplane in the infinite dimensional dual space corresponds to some arbitrarily shaped surface in the data space. Functional analysis then allows

$$\frac{1}{2}\left\|\sum_{i=1}\alpha_i\phi(x_i)\right\|^2 - \sum_i y_i\alpha_i \text{ s.t.} \begin{cases} \sum_i \alpha_i = 0 \\ 0 \le y_i\alpha_i \le C \end{cases}$$

to be minimized over $\boldsymbol{\alpha}$ (a quadratic programming problem), where the inner product $\big(\phi(x_i), \phi(x_j)\big)$ is the kernel function, the $y_i \in \{-1, 1\}$s are the labels, and C is the box constraint. The data is then input into the SVM for training. The training is accomplished using the k-folds method, with a validation test set withheld from the training data.

Data

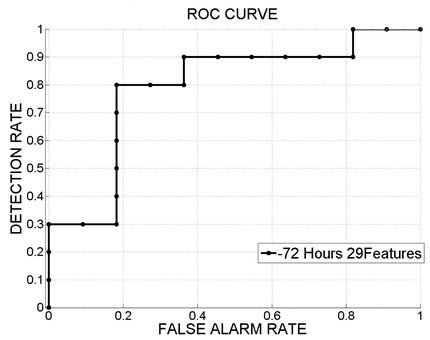
Due to the large physical size of the atmosphere and the difficulty of collecting data for many variables (e.g. wind speed, temperature, etc.) a numerical physics model is backfitted to measured data that is obtained from various sources. The model outputs are then used as the "truth" for input into our machine learning algorithms. We are currently using the GFS-FNL data for a storm centered volume of $\pm 25°$ in latitude, $\pm 30°$ in longitude, 1000-100hPa in pressure for 7 different variables, and times at 6 hour intervals from -72 hours to +72 hours. Our data set consists of 108 storms that underwent ET from 2000-2008; with each storm having a vector of length 5,945,121. Although we are currently investigating other machine learning methods, our primary results have been accomplished via the SVM algorithm [1].

Data Issues

The low number of labeled data (108 storms) in very high dimensions ($n = 5,945,121$) causes over-training ("curse of dimensionality"). *A priori* feature extraction is used to mitigate over-training.

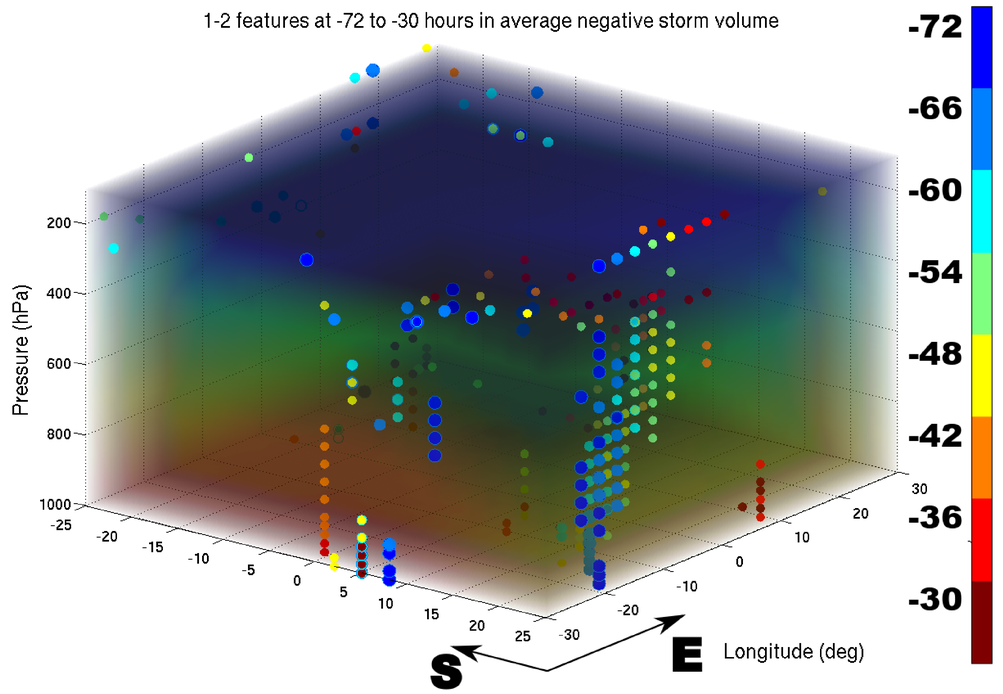
Results from SVM

The dimensionality of the input vectors was reduced using correlation-based feature selection (CFS) developed by Hall before input into the SVM. For the -72 hour time point, the best ROC curve was obtained by using 29 dimensional vectors (from Hall CFS) from equivalent potential temperature, θ_e, volumes (1 for each storm) for the SVM training, and then running the resultant SVM classifier on the withheld test set. θ_e can be thought of as a quantity proportional to energy, and it is computed from the raw outputs of the FNL data.



Finding Structure in ET Data

The following figure shows the best 2 features (from CFS) at various times at each pressure level in the mean θ_e volume of the storms which intensified. Dark blue is at -72h, light blue is at -66h, etc.



Obvious spatio-temporal features are apparent, particularly the air column near $25°$ lat. and $-20°$ long. which moves from the edge into the interior of the volume in time. Subsequent research used 15 CFS features from each pressure level to find the best ROC curve on the validation test set (using area under the ROC or AUC) at each pressure level, at each time point using SVMs. This research indicated pressure levels with the highest AUCs varying from 1000hPa to 700hPA with AUCs varying from 0.71 to 0.86. This research suggests important structure in the data, corresponding to physical spatio-temporal parameters. These parameters are likely important to the underlying physics (PDE flow, etc.)

ET Chaotic?

The results from the SVM classification indicate that the ET process is *not* chaotic on a ± 72 hour time scale. This highlights that the dynamics and physics of these systems are not well understood, and further research is needed to understand the time scales at which predictability breaks down, i.e. at what space and time scales is the system chaotic? We intend to explore this question by using machine learning, specifically RML [2],to extract the *intrinsic* features of the data manifold.

References

[1] S. R. Felker, B. LaCasse, J. S. Tyo, and E. A. Ritchie. Forecasting post-extratropical transition outcomes for tropical cyclones using support vector machine classifiers. *J. Oceanic Atmos. Tech. (Accepted)*, .:., 2010.

[2] T. Lin and H. Zha. Riemannian manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:796–809, 2008.

Funding

This work was supported by the NSF Division of Atmospheric and Geospace Sciences, Award # ATM0730079