Supporting Information for "Oceanic eddy detection and lifetime forecast using machine learning methods"

Mohammad D. Ashkezari,[1] Christopher N. Hill,[1] Christopher N. Follett,[1]

Gaël Forget,[1] and Michael J. Follows[1]

Contents of this file

1. Text S1 to S4
2. Figures S1 to S6
3. Tables S1 to S4

Introduction The supporting information presented here consists of more details regarding our numerical method for oceanic eddy detection and lifetime prediction. We provide a complementary set of metrics that evaluates the classification algorithms used for eddy

---

Corresponding author: Mohammad D. Ashkezari, Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA. (DeMo@mit.edu)

[1]Department of Earth, Atmospheric and
Planetary Sciences, Massachusetts Institute
of Technology, Cambridge, MA 02139 USA.

detection. Also, a full list of features in connection with the regression model used for lifetime forecast is presented in this document.

Text S1. This paper uses the python implementation of machine learning methods for eddy identification; a publicly available and well-documented library [*Online Access1*, @]. The ultimate goal of such methods is to develop a model that generalizes to the out-of-sample data. Since these methods remain a relatively new analytical tool in oceanography, it may be useful to illustrate the typical workflow through an example.

Various machine learning models were examined for this study and, ultimately, Support Vector Machines (SVM) classifier was used to identify the eddy cores. SVM is an advanced supervised machine learning algorithm that is used both for classification and regression problems. The support vector machine classifier attempts to find an optimized hyperplane decision boundary with N-1 dimensions where N is the dataset dimensions or the number of features associated with each class. The decision boundary is determined so that the margin of separation between the different classes is maximized. As an example, Fig. S1 shows a simple two-class (cyan and pink circles) problem with two features. Therefore, the decision boundary will be a simple line. The data points closest to the decision boundary are known as "support vectors" (hence name support vector machines) and are critical to constrain the maximum separating margin. The problem treated in this paper is directly analogous to Fig. S1 but with 21×21 features. Once the SVM model is trained and its performance is established (see Sec. 3.1 and details below), it provides an efficient means to classify extensive data sets (see Sec. 3.2).

The training phase and performance evaluation of the SVM classifier is carried out through

a classic cross-validation procedure [*Online Access2*, @]. The model performance is evaluated by standard metrics that are commonly used within the machine learning community (see Tab. S1 and descriptions), and are supported by the machine learning libraries such as scikit-learn (python), R, and Matlab Machine Learning Toolbox.

Typically, machine learning models are controlled by a number of settings that adjust the learning rate and quality of the model. These settings are enforced by a series of "hyper-parameters" such as those of the Radial Basis Function (RBF) kernel in SVM. When training an SVM with RBF kernel, results may be sensitive to two hyper-parameters named C and gamma. Therefore, finding values that result in an optimized model with stable performances is the critical part of the training procedures. We applied a standard grid search method [*Online Access3*, @] on the train-test dataset to select the values ($C = 100$ and $\gamma = 10^{-5}$) that maximize the model performance.

It should be emphasized that the introductory section presented here only offers a big picture of the machine learning solutions. Fortunately, numerous machine learning textbook presentations [*Bishop*, 2006; *Schlkopf and Smola*, 2001] are available that will provide the interested reader with additional details regarding machine learning methods and their inner workings. In addition, there is a large number of free numerical packages that offer machine learning libraries with examples and tutorials [*Online Access4*, @; *Online Access5*, @; *Online Access6*, @] illustrating how machine learning methods are operated in

practice.

Table S1 lists a series of classification metrics to demonstrate the quality of predictions made on test datasets discussed in Sec. 3.1. These scores are associated with identification of eddy structures from those of non-eddies. The classification metrics are defined as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{1}$$

Intuitively, accuracy can be interpreted as the proportion of correct predictions made by the classifier, precision is the ability of the classifier not to miss-classify a negative sample (non-eddy) as positive (eddy) and recall is the ability of the classifier to identify all of the positive samples. F1 is another measure for classifier performance, which can be understood as a weighted average of the precision and recall scores. Finally, AUC represents the Area Under the Curve of true positive rate versus false positive rate as the classification threshold is varied [*Fawcett*, 2006].

Text S2. Once the eddy cores are identified, the tracking algorithm constructs a dataset of eddy trajectories (see Sec. 4). Figure S2 shows distribution of eddy lifetime, dis-

placement, and mean velocity. The lifetime and displacement distributions exhibit close similarity with previously reported results discussed in *Chelton et al.* [2011-b]; *Mason et al.* [2014] . A more detailed statistical summary is presented in Tab. S2.

Text S3. In Sec. 3.3 of the article the distributions of relative vorticity associated with the detected eddy cores at Hawaiian regions using our method and another recent SLA-based method [*Faghmous et al.*, 2015] were compared. Here we repeat the comparison procedure for Gulf Stream region, a dynamically and geographically different domain (delimited by 22°N-47°N and 85°W-48°W). Figure S3 compares the identified eddy cores using the two methods at a given day. Figure S4 shows that our method separates the vorticity distributions of the detected eddy cores (panel **a**) while misclassification is observed in the other method (panel **b**). Panel **c** shows the results of our method using different training sets, and again it appears that our detection algorithm is largely insensitive to the training region choice.

Text S4. As mentioned in the article, we selected 900 smoothly-evolved eddy trajectories with a nearly balanced population of cyclonic and anti-cyclonic eddies for train-test purposes of the regression model used for lifetime forecast (see Sec. 5). Figure S5 shows the distribution of lifetime, displacement, and mean velocity of the selected eddies (statistical summaries may be found in Tab. S3). Each eddy, and at a given day, is associated with a 10-dimensional numerical vector consisting of a list of features that we considered to be relevant to eddy lifetime. Table S4 lists the associated lifetime features. Once the model

is trained, it is examined on the test set. Figure S6 demonstrates the predictions made by the trained model on a test set.

Table S1. A list of classification scores for SVM and random forest classifiers. The uncertainties indicate one standard deviation on multiple realizations.

Classifier	Accuracy	Precision	Recall	F1	AUC
SVM	0.92 ± 0.03	0.99 ± 0.01	0.87 ± 0.04	0.91 ± 0.04	0.92 ± 0.03
Random Forest	0.97 ± 0.02	0.96 ± 0.03	0.98 ± 0.02	0.97 ± 0.02	0.97 ± 0.02

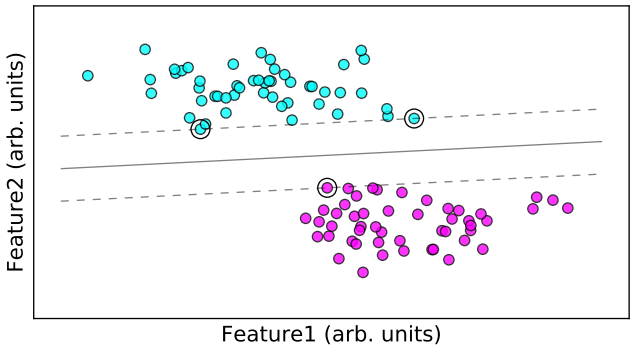

Figure S1. A two-dimensional dataset classified using SVM algorithm (class1: cyan, class2: pink). The decision boundary is determined so that the separating margin is maximized. Notice that the dashed lines touch only a few data points (circled). These points are known as the support vectors and are critical in the fitting procedure of the model.

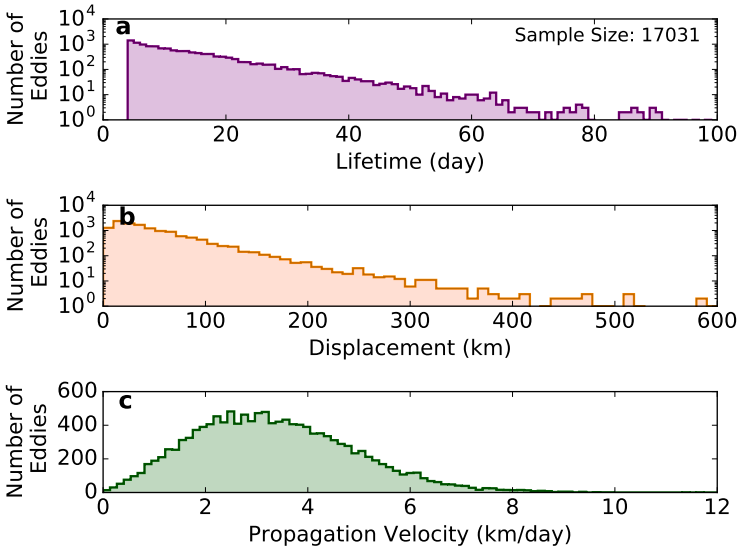

Figure S2. Histograms of a few properties of detected eddies around the Hawaiian islands at the North Pacific Subtropical Gyre. Panel **a** shows the distribution of eddy lifetime. Panel **b** shows the distribution of eddy displacement, and panel **c** illustrates the distribution of mean velocity of the eddy cores. It should be stressed that some of the identified eddies may have formed or died outside of the chosen domain, which may result in low biases in lifetime and displacement.

Table S2. A list of eddy characteristics derived from the tracking algorithm. Notice that cyclonic and anti-cyclonic eddies have similar sample size.

	Mean Lifetime (day)	**Max Lifetime (day)**	**Mean Displacement (km)**	**Max Displacement (km)**	**Mean Velocity (km/day)**	**Max Velocity (km/day)**
All Eddies	14.34	231	53.86	1008	3.35	12.01
Cyclonic Eddies	14.48	231	53.84	1008	3.30	11.67
Anti-Cyclonic Eddies	14.21	128	53.87	734	3.40	12.01

Table S3. A list of eddy characteristics derived from the population of the smoothly-evolving eddy trajectories selected for training and evaluation of the lifetime predictive model.

	Mean Lifetime (day)	**Max Lifetime (day)**	**Mean Displacement (km)**	**Max Displacement (km)**	**Mean Velocity (km/day)**	**Max Velocity (km/day)**
All Eddies	25.02	231	94.67	1008	3.40	9.78
Cyclonic Eddies	25.70	231	96.60	1008	3.34	7.43
Anti-Cyclonic Eddies	24.36	117	92.81	483	3.45	9.78

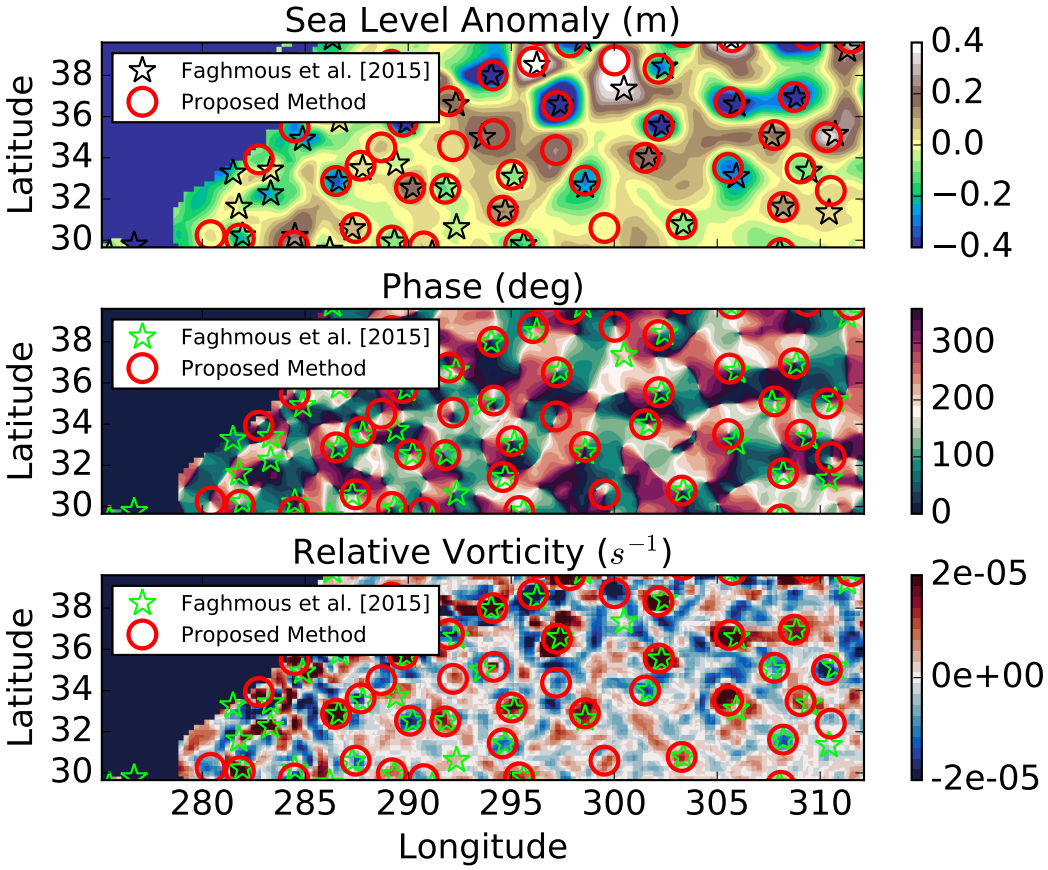

Figure S3. Comparison of the detected eddy cores using the proposed method and that of *Faghmous et al.* [2015]. Panels **a**, **b**, and **c** compare the identified eddy cores superimposed on SLA, surface velocity phase, and relative vorticity, respectively.

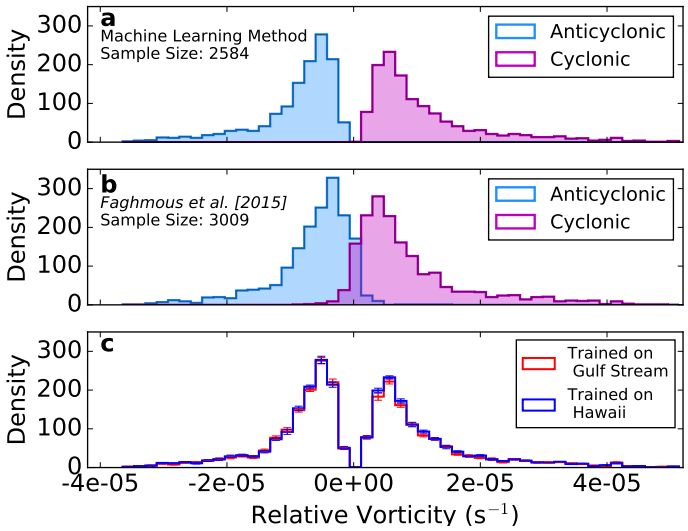

Figure S4. Panels **a**, **b** show the distribution of relative vorticity at the location of detected eddy cores at the Gulf Stream region and using the method presented in this study and by *Faghmous et al.* [2015], respectively. Panel **c** compares the distribution of relative vorticity of the detected eddies at the Gulf Stream region using two different training sets. The error bars indicate one standard deviation on multiple realizations.

Table S4. Features that are computed for an eddy at each day during its life time.

Feature	**Description**
Phase Integral	Integral of sinosoidal function of the phase values within a radius of 30 km
Velocity	Instantaneous velocity of the eddy core
Acceleration	Instantaneous acceleration of the eddy core
Mean SLA	Mean value of the Standardized SLA within a radius of 30 km
Std. Dev. SLA	Standard deviation of the Standardized SLA within a radius of 30 km
Mean Vorticity	Mean value of the Standardized relative vorticity within a radius of 30 km
Std. Dev. Vorticity	Standard deviation of the Standardized relative vorticity within a radius of 30 km
Age	Current age of the eddy
Location: Latitude	Current location (latitude) of the eddy
Location: Longitude	Current location (longitude) of the eddy

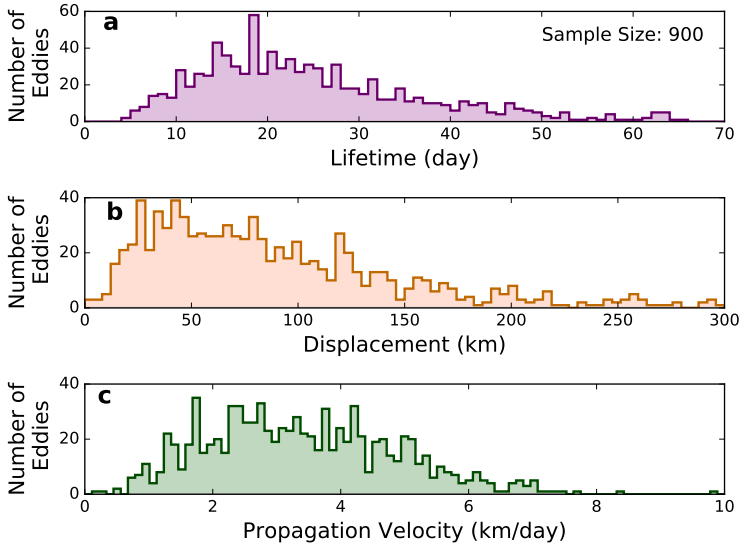

Figure S5. Properties of the selected smoothly-evolved eddies around the Hawaiian islands at the North Pacific Subtropical Gyre. Panels **a, b, c** show the associated distribution of eddy lifetime, displacement, and mean velocity.

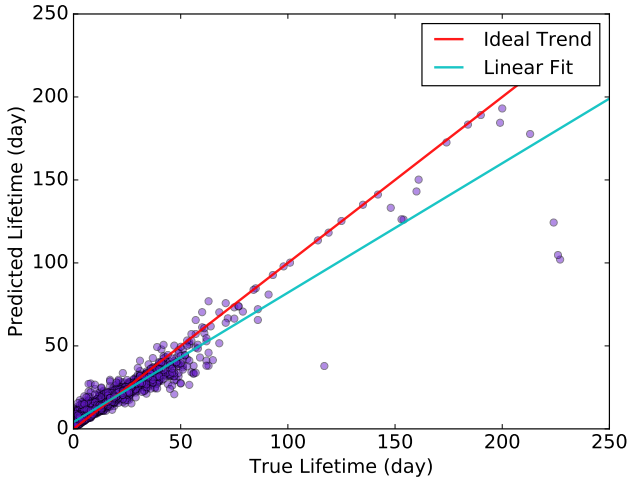

Figure S6. Lifetime predicted by the random forest regressor model as a function of true lifetime. The cyan line shows a least-square linear fit on the predictions. If the model performance is enhanced, one expects that the cyan line approaches the red line which presents a perfect model.

References

Bishop, C. M. (2006), Pattern Recognition and Machine Learning, *Springer*.

Chelton, D., M. Schlax, and R. Samelson (2011-b), Global observations of nonlinear mesoscale eddies, *Progress in Oceanography*, *91*, 167–216.

Faghmous, J. H., I. Frenger, Y. Yao, R. Warmka, A. Lindell, and V. Kumar (2015), A daily global mesoscale ocean eddy dataset from satellite altimetry, *Scientific Data*, doi: 10.1038/sdata.2015.28.

Fawcett, T. (2006), An introduction to ROC analysis, *Pattern Recognition Letters*, *27*, 861–874.

Mason, E., A. Pascual, and J. C. McWilliams (2014), A new sea surface height-based code for oceanic mesoscale eddy tracking, *Journal of Atmospheric and Oceanic Technology*, *31*, 1181–1188.

Online Access1: http://scikit-learn.org/

Online Access2: http://scikit-learn.org/stable/modules/cross_validation.html

Online Access3: http://scikit-learn.org/stable/modules/grid_search.html

Online Access4: http://www.svm-tutorial.com/

Online Access5: http://scikit-learn.org/stable/modules/svm.html

Online Access6: http://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html#bsr5b42

Schlkopf, B., and A. J. Smola (2001), Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, *MIT Press*.