Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions

Maike Sonnewald,[1,2,3] Carl Wunsch,[1,2], and Patrick Heimbach[3]

Supporting Information for "Insert Title"

=Authors=

$^{=number=}$=Affiliation Address=

Corresponding author: Maike Sonnewald, Harvard University, 26 Oxford Street, Cambridge, MA 02138. (maike_s@mit.edu)

[1]Massachusetts Institute of Technology,
77 Massachusetts Ave, Cambridge, MA
02139, USA.

[2]Harvard University, 26 Oxford Street,
Cambridge, MA 02138.

[3]Institute for Computational Engineering
and Sciences, Institute for Geophysics,
Jackson School of Geosciences, The
University of Texas at Austin, 201 East
24th Street, Austin, TX 78712, USA.

Abstract.

Dynamically similar regions of the global ocean are identified using a barotropic vorticity (BV) framework from a twenty-year mean of the ECCO state estimate at 1° resolution. An unsupervised learning algorithm, k-means, objectively clusters the standardized BV equation, identifying five unambiguous regimes. Cluster 1 covers 43±3.3% of the ocean area. Surface and bottom stress torque are balanced by the bottom pressure torque (BPT) and the non-linear torque. Cluster 2 covers 24.8 ± 1.2%, where the beta effect balances the BPT. Cluster 3 covers 14.6±1.0%, characterized by a 'Quasi-Sverdrupian' regime where the beta effect is balanced by the wind and bottom stress term. The small region of Cluster 4 has baroclinic dynamics covering 6.9 ± 2.9% of the ocean. Cluster 5 occurs primarily in the Southern Ocean. Residual 'dominantly non-linear' regions highlight where the BV approach is inadequate, found in areas of rough topography in the Southern Ocean and along western boundaries.

1. Motivation

Before the advent of modern observational and modeling techniques, understanding of the physical/dynamical state of the ocean focused on large-scale quasi-laminar descriptions such as Sverdrup balance, abyssal recipes, or Stommel-Arons flows (**???**). Recent advances in instrumentation and modeling capability have revealed a complex spatial and temporal variability of oceanic physics. It is possible that every location in the ocean has a unique physical state depending upon many factors, including local topography, meteorology, proximity to eastern and western boundaries, or latitude, rendering any global scale interpretation lacking in general applicability.

The ocean is spatially and temporally diverse, but delineating spatial and temporal commonalities and continuities are central to understanding emergent patterns that lead to a global geography of dynamical regimes. The dominant physics underlying the emergent patterns become evident when common features are identified. The purpose of this note is to explore unsupervised learning as a method for depicting and understanding the gross features of global oceanic physics. The study is restricted to the barotropic vorticity (BV) balance of a time-mean global circulation as calculated from a non-eddy resolving state estimate (**?**). The approach appears to be both interesting and useful and is readily generalized to far more complex oceanic states. Unsupervised learning refers to that the physically significant patterns have not been labeled a-priori, as with supervised learning in e.g., neural networks. The non-eddy resolving case is explored in this initial work, but has relevance as the resolution is similar to that of ocean models deployed within climate model simulations in support of the CMIP5 and CMIP6 efforts (??).

The presence of differing dynamical regimes is already suggested by the known structures of the wind-stress forcing and the geometry of the ocean basins, including the underlying topography. Classifying and identifying regions in the world ocean is done here using the variations in the dominant terms of the BV budget, following the demonstration that the global budget can be closed. ?? have assessed the dynamics of the BV budget focusing on the North Atlantic Ocean variability and sensitivity to resolution, respectively. Here the procedures are global.

Distinct geographical physics were demonstrated by e.g. ?, using altimeter data to show differing spatial regimes of geostrophic turbulence. Their global patterns are presumably connected to circulation structures, planetary waves, and topography. Similarly, ?? used linear statistical models finding patterns suggestive of global regimes. The present work extends the pattern determination methodology, as working only with data from the surface limits the application to a comprehensive assessment of global dynamical regimes.

Objective classification via k-means clustering, allows unbiased identification of patterns in data. This form of unsupervised machine learning is common in many fields ranging from pharmaceutical to engineering (???). ? applied a similar method to identify regions with distinct biological activity, and k-means have been used to identify key regions for data collection to build maps of nitrate in the Southern Ocean (Liang, pers com). Applications in the earth sciences have been investigated both in the prognostic and diagnostic sense (??).

2. Methods

2.1. The ECCO Version 4 State Estimate

The BV equation is applied to the version 4, release 2, of the Estimating the Circulation and Climate of the Ocean (ECCOv4) global state estimate described by ??, see also ??. The state estimate has a nominally 1° resolution, available at: ecco.jpl.nasa.gov. A least-squares with Lagrange multipliers approach is used to obtain the state estimate. The result is a *free-running* version of the MIT General Circulation Model (MITgcm, ?), with adjusted initial and boundary conditions and internal model parameters. The ECCO state satisfies basic conservation laws for enthalpy, energy, salt, volume, and momentum while remaining largely within error estimates of a diverse set of global data (???). Regions without data are filled in a dynamically consistent way using the dynamics, still relying on parameterizations but avoiding the use of untested statistical hypotheses e.g., ?.

2.2. Barotropic Vorticity

The momentum and continuity equations of an ocean in a thin shell on a rotating sphere are:

$$\partial_t \mathbf{u} + f\,\mathbf{k} \times \mathbf{u} = -1\rho_0 \nabla_h p + 1\rho_0 \partial_z \tau + \mathbf{a} + \mathbf{b}, \partial_z p = -g\rho, \nabla_h \cdot \mathbf{u} + \partial_z\, w = 0 \qquad (1)$$

Pressure, gravity, density and vertical shear stress are denoted p, g, ρ and τ respectively, with ρ_0 the reference density; the three dimensional velocity field $\mathbf{v}$ = (u, v, w) = ($\mathbf{u}$, w); the gradient $\nabla = (\nabla_h, \partial_z)$; the unit vector is denoted $\mathbf{k}$; planetary vorticity is a function of latitude ϕ in $f\,\mathbf{k} = (0, 0, 2\,\Omega \sin\phi)$; the viscous forcing by vertical shear is denoted $\partial_z \tau$; the non-linear torque are $\mathbf{a}$ and the horizontal viscous forcing $\mathbf{b}$ includes sub-gridscale parameterizations. Assuming a steady state, the vertical integral from the surface $z = \eta(x, y, t)$ to the water depth below the surface $z = H(x, y)$ is,

$$\beta V = 1\rho_0 \nabla p_b \times \nabla H + 1\rho_0 \nabla \times \tau + \nabla \times \mathbf{A} + \nabla \times \mathbf{B}, \tag{2}$$

where $\nabla \cdot \mathbf{U} = 0, \mathbf{U} \cdot \nabla f = \beta V$, the bottom pressure is denoted p_b, $\mathbf{A} = \int_H^\eta \mathbf{a} \mathrm{d}z$ and $\mathbf{B} = \int_H^\eta \mathbf{b} \mathrm{d}z$. The curl operator $\nabla \times$ yields a scalar, representing the vertical component of the operator. The LHS of equation (**??**) is the planetary vorticity advection term, while the RHS of equation (**??**) is the bottom pressure torque (BPT), the wind and bottom stress curl, the non-linear torques and the viscous torque, respectively. The sub-gridscale parameterization introduces a torque, which is included in the viscous torque term. The non-linear torque is composed of three terms:

$$\nabla \times \mathbf{A} = \nabla \times \left[\int_{-H}^{\eta} \nabla \cdot (\mathbf{u}\mathbf{u}) \mathrm{d}z \right] + [w\zeta]_{z=H}^{z=\eta} + [\nabla w \times \mathbf{u}]_{z=H}^{z=\eta} . \tag{3}$$

The RHS of equation (**??**) represents the curl of the vertically integrated momentum flux divergence, the non-linear contribution to vortex tube stretching and the conversion of vertical shear to barotropic vorticity. Horizontal viscous forcing includes that induced by sub-gridscale parameterizations. Twenty-year averaged fields of the BV equation are used after a Laplacian smoother is applied, with an effective averaging range of three gridcells.

2.3. Unsupervised learning: K-means clustering

Various combinations of terms dominate the BV equation in different regions, and determining the dominant spatial patterns is the present goal. Clustering determines patterns of variability, as discussed in textbooks e.g. ?. If groups of dominant terms are present that differ significantly, the “clusters” are robustly identified. The terms in the BV equation were each scaled to have zero mean and unit variance globally. This normalization and feature scaling is applied, centering to the mean and component wise scaling the individual terms of the BV equation. The k-means algorithm involves an iterative minimization of the sum of squares of the Euclidean distance partitioning of the hyperspace given by the terms in the BV equation:

$$J = \Sigma_{j=1}^{\mathrm{K}} \Sigma_{i=1}^{n} ||\mathbf{x}_i^j - \mathbf{c}_j||^2 \tag{4}$$

where the number of K clusters is a parameter fixed a priori, representing the initial guesses randomly scattered among the parameter space space given by the normalized BV equation. $\mathbf{x}$ is a vector field on the discretized sphere, with each element $\mathbf{x}_i$ representing a 5-dimensional vector on the model's horizontal grid, such that index i uniquely identifies a grid point on the sphere, with (lon,lat) $= (\phi_i, \theta_i)$. The components of each vector $\mathbf{x}_i$ correspond to the 5 terms in the BV budget. Each cluster $j = 1, \ldots,$ K is represented by the 6-dimensional characterizing vector c_j, and the k-means classification attributes each vector $\mathbf{x}_i$ to a unique cluster c_j, thus $\mathbf{x}_i = \mathbf{x}_i^j$. The distance between a data point $\mathbf{x}_i^j$ and $\mathbf{c}_j$ is given by $||\mathbf{x}_i^j - \mathbf{c}_j||^2$. Each data point is associated with the closest K-cluster, the position of $\mathbf{c}_j$ is recalculated, and the association reassessed until the solution converges.

Assumptions regarding the covariance of the data are discussed in the appendix.

The solution is sensitive to the initialization and choice of K, and the algorithm partitions the parameter subspace using linear hyperplanes. This linearity constraint means that higher numbers of K can both assist in partitioning the subspace more appropriately, and isolate noise. The appendix demonstrates the data's small sensitivity to the algorithm's initial random seed, and the impact of varying K. The appropriate value of K is determined as K>35 using the Akaike and Bayesian Information Criteria (AIC and BIC). Information criteria are a measure of the quality of a statistical model, weighting increased precision against the cost of additional parameters. The AIC and BIC indicate robust regimes as they both asymptote, suggesting no information is gained by further increasing K. K=50 is used for the remaining analysis.

3. Results

The closure in ECCOv4 for the 20-year average of the BV terms in equation (**??**) is illustrated in Figure **??**. Individual terms are of order $\pm 10^{-9}$ m s^{-1} and the residuals have magnitudes $<< 1\%$ for 36% of the ocean—a very small value (Figure **??**), less than $\pm 10^{-12}$ m s^{-1}. These small residuals permit going forward with confidence. Some numerical issues do exist on the continental shelf and in shallow water generally, but these regions only amount to 3% of the area of the global ocean and will be ignored.

Figure **??**b illustrates where the beta term is important from equation (**??**). This term is balanced by the BPT term shown in Figure **??**c, and the wind and bottom stress BV terms shown in Figure ??d (the bottom stress term is small). The remainder is largely found in the non-linear BV contributions seen in Figure ??e, with the lateral viscous dissipation largely being an order of magnitude smaller, apart from in localized regions in the Southern Ocean. The wind and bottom stress BV terms in Figure ??d are largely zonally symmetric, with large patterns of negative BV in the Southern Ocean, and large gyre patterns visible in the Pacific and Atlantic basins. The BPT term in Figure ??c is associated with interactions with steep bathymetry. For example, in the Southern Ocean a large positive patch leads towards the Antarctic-Pacific ridge, with a negative patch beyond. This structure is consistent with vortex stretching as the ridge is crossed. Along Western Boundaries, BPT is positive to the west and negative just adjacent to the east, consistent with studies such as ?. The BV of the non-linear torque is concentrated along the western edge of basins where WBCs are found, but it is less spatially coherent than the BPT term. The Southern Ocean stands out as a region of large activity, particularity in the Atlantic sector. Lateral viscous dissipation is small.

Picking out globally coherent dynamical regimes, the k-means algorithm results are presented in Figure ?? where the numbering on the colorbar is arbitrary. The structure is mainly found in five named regimes summarized as in Table ??, which we now briefly examine. Each is numbered, and a partially descriptive label is attached—no standard nomenclature appears to exist.

Figures ?? and ?? isolate the geographical area and the bar charts represent the balance of the terms in the BV equation illustrated in Figure ??, from the geographical areas determined as distinct by the algorithm. Area-averaging was done for a comparison.

The largest cluster, accounting for 43% of the global ocean (Cluster 1), is depicted in Figure **??**a. The wind stress curl is nearly balanced by the bottom torques. This depth coherent "negative wind curl/bottom torque" region is found primarily in zonal streaks in the tropics, and in a thin ribbon in the Southern Ocean mainly in the Pacific sector. In the Northern Hemisphere, Cluster 1 areas surround the subtropical and subpolar gyres. Large areas of the Arctic Seas are also in this Cluster. Figure **??**b demonstrates that the balance of terms is dominantly between the input of negative vorticity by the wind-stress curl largely balanced by the positive input by the bottom interaction terms.

The next largest dynamical region covers 25% of the ocean area (Cluster 2 Figure **??**c: Interior flow "positive wind curl/beta and bottom torque"), where the wind stress curl inputs positive vorticity, nearly balanced by the beta and bottom interaction terms. In the Northern Hemisphere, this cluster covers the southern region of the subpolar gyres. A zonal streak crosses the equator in both the Atlantic and Pacific, but is absent in the Indian Ocean. The Southern Hemisphere has large Cluster 2 expanses in both the Pacific and Atlantic, but again not in the Indian Ocean.

The 15% of the ocean area selected by Cluster 3 are illustrated in Figure **??**e (Quasi-Sverdrupian "negative wind torque/beta effect"). Dominant areas in the subtropical gyres in the Northern Hemisphere Atlantic and Pacific stand out, together with thin streaks on the Equator. Isolated streaks are seen in the Southern Ocean, and in a large area of the Southern Hemisphere tropical Indian Ocean. This region might be considered also as corresponding to quasi-Sverdrup balance.

The area covered by Cluster 4 (Figure **??**a: Interior flow, vertical, “positive wind torque, beta and bottom stress”), covering 7% of the ocean and which is a complement to Cluster 1. In the Northern Hemisphere, the Cluster largely represents the northern edge of the subpolar gyre. In the Southern Hemisphere, it is found on the eastern edge of the Pacific and Atlantic basins, just to the south, and flaring out westwards of the continental barrier. In the Indian Ocean, this barrier can be seen to be New Zealand or Australia, and the area of this dynamical regime fills the subtropical Indian Ocean down to the border with the Southern Ocean, where this regime is absent. Figure **??**b illustrates that it is an amplified version of the dominant terms seen in Figure **??**d, being an order of magnitude larger, but still having the wind as the major source of barotropic vorticity, with sinks in the Coriolis term and BPT. A small source exists in the non-linear torque.

The Southern Ocean is better represented in the area covering only 2% of the global world ocean seen in Figure **??**c (Cluster 5: “Southern Ocean gyre”), as seen mainly in a series of streaks in the Southern Ocean with negative wind torque, and a complement to Cluster 4. Again, non-linear torque are a small sink.

A summary of the area of the remaining clusters that account for 9% of the world ocean (Figure **??**e: “Dominantly non-linear”). Separate clusters have different colors. Areas of rough bathymetry stand out, such as the Pacific-Antarctic Ridge and the Drake

Passage area. Figure ??f is the overall average, illustrating that the non-linear contribution to the barotropic vorticity dominates, together with the Coriolis term. The different constituents are quite varied, but strong contributions from the non-linear torque are consistently present. Their detailed discussion is the subject of a subsequent study.

4. Discussion and Conclusions

The barotropic vorticity equation closes very accurately in the 20-year time-average ECCOv4 state estimate, and is analyzed for the world ocean using k-means clustering to find regions of common balances. Figure ?? shows that the global ocean has large regions of spatially consistent dynamical-term balances as displayed in Figure ??. Those balances vary among the wind-stress, Coriolis and bottom pressure torque (BPT) terms. Areas where the non-linear torque are small suggest that the linearized BV is a good approximation. Areas where the non-linear torque are important are found in western boundary regions, as well as the Southern Ocean where the Antarctic Circumpolar Current interacts with bathymetric obstacles. The momentum dominated area implies a coherent vertical structure. The subtropical gyre in Cluster 3 is unique in lacking significant contributions by BPT, implying it is shielded from topography. Transition zones have a stronger momentum-driven portion of the BPT, and topographic interactions to become important. Cluster 4 has a stronger baroclinic component to the BPT, feeling topography. The Southern Ocean cluster is like Cluster 2, but with contributions of opposite sign. The remaining ocean has important non-linear contributions, and the linearized barotropic interpretation is not appropriate.

In the North Atlantic Ocean, results are generally consistent with the inferences of **??**. Cluster analysis reveals a shift from a barotropic flow in Cluster 1 and 3, to a strong interior flow (baroclinic meridional, North Atlantic Current, and North Atlantic Deep Water, flow over the Mid Atlantic Ridge) in Cluster 2 and 4. Globally, the Clustering illustrates that strong interior flow is present in vast expanses of the Southern Hemisphere, as well as in the North Pacific. Cluster 4 coincides with regions identified by **???** as areas of water-mass transformation and intermediate pathways in the overturning circulation between surface and deep water. The quasi-Sverdrupian regime in Cluster 3 is not present in the South Atlantic and Pacific. The Southern Ocean mainly has clusters 3 and 5, with significant non-linear contributions.

Five regions cover 91% of the world ocean. Residual areas collected here as "dominantly non-linear" have a small spatial extent, but are dynamically important in the overall circulation. These regions include the Drake Passage region as well as the Antarctic-Pacific Ridge regions where the circulation interacts with topography and cross frontal transport likely takes place. In the Northern Hemisphere, areas in the Labrador Sea and on the continental shelf stand out as non-linear. These nonlinear regions will be the subject of a separate study at higher resolution.

The sign and spatial distribution of the wind-stress term suggests the importance of Ekman pumping (negative) or suction (positive). The equatorial and Southern Ocean regions show Ekman pumping, whereas the subpolar gyre areas have Ekman suction where mode waters are created. The BPT term mirrors the wind-stress term, suggesting it acts as either a source or a sink in opposite complement to the wind-stress. The lack of symmetry in wind driven gyres in the Southern and Northern Hemisphere show that the gyres are not driven solely by the sign of the Ekman pumping. This complex relationship among the terms is the subject of future study.

The use of vertical integrals to describe the circulation is a simplification of what is a three-dimensional problem, as is the use of a model in which the important eddy field is only included through parameterizations such as in CMIP5 and CMIP6 **??**. Future work is intended to apply this and related machinery to fully eddy-resolving ocean states.

5. Acknowledgments

This work was funded by the US National Aeronautics and Space Administration Sea Level Change Team (contract NNX14AJ51G) and through the ECCO Consortium funding via the Jet Propulsion Laboratory. MS acknowledges the advice and support of Anne Reinarz, Roosa Tikkanen and Katherine Rosenfeld.

Appendix A K-Means and influence of Information Criteria

The k-means algorithm is related to methods such as PCA, more traditionally applied to oceanography. Where PCA attempts to represent all data vectors using a low order combination of eigenvectors, minimizing the mean squared reconstruction error, the k-means algorithm represents the data vectors via a small number of clusters. This is also done to minimize the mean squared reconstruction error. In this manner, the k-means algorithm can be interpreted as a very sparse PCA.

Robustness of the regions in terms of the stochastic initialization is highlighted in Figure **??**c, where the k-means clustering was run 100 times. The mean and 2σ are used in Table **??**. The regimes identified are robust, with the extent of the subpolar gyre being the main area where the algorithm shows appreciable variance.

The k-means algorithm is initiated by scattering K first-guesses of where the parameters/clusters could be. This initial guess introduces a stochastic element. The success of the algorithm is sensitive to K, as this determines how the hyperspace given by the dimensions is partitioned. As with regression analysis, adding parameters can increase the accuracy, but over fitting should be avoided. Determining the appropriate value of

K, information criteria (Akaike and Bayesian Information Criteria, AIC and BIC) are used to assess the quality of the statistical model. These measures weight the added accuracy with the cost of adding additional parameters, minimizing the expectation of the prediction error. are used:

AIC=2K-2ln($\mathcal{L}$),

BIC $=$ Kln$(n) - 2\ln(\mathcal{L})$, where n is the number of datapoints and $\mathcal{L}$ is the likelihood:

L=$\Pi_{i=1}^{N}1\sqrt{2\pi\sigma^2}$exp $\left(-(\zeta_i - \hat{\zeta}_i)^2 2\sigma^2\right)$.

ζ_i is the observed, and $\hat{\zeta}_i$ is the prediction, so $(\zeta_i - \hat{\zeta}_i)^2$ are the prediction residuals. In the estimate, the AIC value is minimized, which determines the smallest appropriate order to represent the time-series. As discussed by **?** and **?**, the AIC can overestimate the order. Figure **??**b demonstrates that both the AIC and BIC stabilise at $> 35K$, and the asymptotic nature of the regime.

The Euclidian distance is used, meaning the variance is assumed to be isotropic (meaning round). This leads to the standard practice of normalizing and standardizing data. To elucidate the impact of assumptions the algorithm makes for the classification, a more generalised form of clustering was also tested: Gaussian Mixture Models (GMM). GMM are used to assess the impact of assumptions relating to the covariance structure; spherical, diagonal, tied or full covariance. Using the BIC, the results from the BV data were not seen to be sensitive to this. However, this could be important at higher resolution as the k-means clustering problem is NP-hard and GMM could perform better.

References

Adcroft, A., Hill, C., Campin, J. M., Marshall, J., & Heimbach, P. (2004). Overview of the formulation and numerics of the MIT GCM (pp. 139-150). Presented at the ECMWF Conference Proceedings, Shinfield Park, Reading, UK.

Akaike, H. (1973), ”Information theory and an extension of the maximum likelihood principle”, in Petrov, B.N. ; Cski, F., 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadmiai Kiad, p. 267-281.

Ardyna, M., H. Claustre, J. Sallee, F. DOvidio, B. Gentili, G. van Dijken, F. DOrtenzio, and K. R. Arrigo, 2017: Delineating environmental control of phytoplankton biomass and phenology in the southern ocean. Geophys. Res. Lett., 44, 50165024, doi:10.1002/2016GL072428.

Breuhl S, et al. (1999). Use of clusters analysis to validate IHS diagnostic criteria for migraine and tension-type headache. Headache; 39(3):181-9. A study of validating diagnostic criteria using k-means on symptom patterns.

Church, J.A., et al., 2013: Sea Level Change. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)], Chapter 13, pp. 1137-1216, Cambridge University Press.

ECCO Consortium, 2017a, A Twenty-Year Dynamical Oceanic Climatology: 1994-2013. Part 1: Active Scalar Fields: Temperature, Salinity, Dynamic Topography, Mixed-Layer Depth, Bottom Pressure.

ECCO Consortium, 2017b, A Twenty-Year Dynamical Oceanic Climatology: 1994-2013. Part 2: Velocities, Property Transports, Meteorological Variables, Mixing Coefficients.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, doi:10.5194/gmd-9-1937-2016, 2016.

Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M Ponte, and C. Wunsch, ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation, Geo. Sci. Model Dev., 8, 2015

Fukumori, I., P. Heimbach, R. M. Ponte, C. Wunsch, A dynamically-consistent ocean climatology. Bull. Am. Met. Soc., doi:10.1175/BAMS-D-17-0213.1, in press, 2018

Hauser J and Rybakowski J (1997). Three clusters of male alcoholics. Drug Alcohol Depend; 48(3):243-50. An example of clustering behavior types in addiction research.

Hirschi, J. J.-M., Blaker, A. T., Sinha, B., Coward, A., de Cuevas, B., Alderson, S., and Madec, G.: Chaotic variability of the meridional overturning circulation on subannual to interannual timescales, Ocean Sci., 9, 805-823, doi:10.5194/os-9-805-2013, 2013.

Chris W. Hughes, Simon D. P. Williams, The color of sea level: Importance of spatial variations in spectral shape for assessing the significance of trends, Journal of Geophysical Research, 2010, 115, C10

Vladimir M. Krasnopolsky, Michael S. Fox-Rabinovitz, and Alexei A. Belochitski, "Using Ensemble of Neural Networks to Learn Stochastic Convection Parameterizations for Climate and Numerical Weather Prediction Models from Data Simulated by a Cloud Resolving Model," Advances in Artificial Neural Systems, vol. 2013, Article ID 485913, 13 pages, 2013. doi:10.11552013485913

Kubat, M. (Ed.), 2015, Introduction to Machine Learning, Second Edition. Springer Publishers.

Kulis, B., Jordan, M. I.Revisiting k-means: new algorithms via Bayesian nonparametrics, Proceedings of the 29th International Conference on Machine Learning (ICML '12)July 2012 Edinburgh, UK5135202-s2.0-84867132578

Lumpkin, R., and K. Speer, 2007: Global ocean meridional overturning. J. Phys. Oceanogr., 37, 2550-256.

J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.

Munk, W., 1950: On the wind-driven ocean circulation. J. Meteor., 7, 7993, doi:10.1175/1520-0469(1950)007¡0080:OTWDOC¿2.0.CO;2.

Munk, W.H., and Palmén, E,: Note on the dynamics of the Antarctic Circumpolar Current, Tellus, 3, 53-55, 1940.

Myers, Paul G., Augustus F. Fanning, Andrew J. Weaver, 1996: JEBAR, Bottom Pressure Torque, and Gulf Stream Separation. J. Phys. Oceanogr., 26, 671683.

Perez, Renellys & Garzoli, Silvia and Meinen, Christopher & Matano, Ricardo. (2011). Geostrophic Velocity Measurement Techniques for the Meridional Overturning Circulation and Meridional Heat Transport in the South Atlantic. Journal of Atmospheric and Oceanic Technology. 28. 1504-1521. 10.1175/JTECH-D-11-00058.1.

Priestley, M. B. (1981) Spectral Analysis and Time Series, London: Academic Press.

Reynolds, R.W., D.B. Chelton, J. Roberts-Jones, M.J. Martin, D. Menemenlis, and C.J. Merchant, 2013: Objective Determination of Feature Resolution in Two Sea Surface Temperature Analyses. J. Climate, 26, 2514-2533,

Schoonover, J., Dewar, W., Wienders, N., Gula, J., McWilliams, J.C., Molemaker, M.J., Bates, S.C., Danabasoglu, G. and Yeager, S.: North Atlantic Barotropic Vorticity Balances in Numerical Models, Journal of Physical Oceanography 2016 46:1, 289-303

Lan, S., Schneider, T., Stuart1and, A., and Teixeira, J. (2017). Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations.

Sonnewald, M., C. Wunsch, and P. Heimbach, 2018: Linear Predictability: A Sea Surface Height Case Study. J. Climate, 31, 25992611, https://doi.org/10.1175/JCLI-D-17-0142.1

Speich, S., B. Blanke, and W. Cai, 2007: Atlantic meridional overturning circulation and the Southern Hemisphere supergyre. Geophys.Res.Lett.,34, L23614, doi:10.1029/ 2007GL031583.

Stainforth, D., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D., Kettleborough, J., Knight, S., Martin, A., Murphy, J., Piani, C., Sexton, D., Smith, L., Spicer, R., Thorpe, A. and Allen, M.: 2005, Uncertainty in predictions of the climate response to rising levels of greenhouse gases, Nature 433(7024), 403-406.

D. Stammer, M. Balmaseda, P. Heimbach, A.Koehl, and A. Weaver, 2016: Ocean Data Assimilation in Support of Climate Applications: Status and Perspectives. Ann. Rev. Mar. Sci., 8, 491-518.

Stommel, H. (1948), The westward intensification of wind-driven ocean currents, Eos Trans. AGU, 29(2), 202206, doi:10.1029/TR029i002p00202.

Stouffer, R.J., V. Eyring, G.A. Meehl, S. Bony, C. Senior, B. Stevens, and K.E. Taylor, 2017: CMIP5 Scientific Gaps and Recommendations for CMIP6. Bull. Amer. Meteor. Soc., 98, 95105, https://doi.org/10.1175/BAMS-D-15-00013.1

Yang, Y. (2005), ”Can the strengths of AIC and BIC be shared?”, Biometrika, 92: 937-950,

Yeager, S.: Topographic coupling of the Atlantic overturning and gyre circulations, J. Phys. Oceanogr., 45, 1258-1284.

Wunsch, C., and P. Heimbach, 2007: Practical global oceanic state estimation. Physica D, 230, 197-208.

Wunsch, C. and P. Heimbach, 2013, Dynamically and kinematically consistent global ocean circulation and ice state estimates. In Ocean Circulation and Climate 2nd Edition, Siedler et al., Eds.

Wunsch, C. The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. Bull. Am. Met. Soc. 80, 245-255 (1999).

Wunsch, C. (2013), Covariances and linear predictability of the Atlantic Ocean, Deep Sea Res., Part II, 85, 228-243.

Wunsch, C. (2015), Modern Observational Physical Oceanography: Understanding the Global Ocean. Princeton University Press.

Xu, Y. and L. Fu, 2012: The Effects of Altimeter Instrument Noise on the Estimation of the Wavenumber Spectrum of Sea Surface Height. J. Phys. Oceanogr., 42, 22292233, https://doi.org/10.1175/JPO-D-12-0106.1

Cluster	Area	Leading terms
1	43±3.3%, Depth coherent (Fig. **??**a)	$blue\nabla \times \tau_{sb} + \nabla \times \mathbf{A} black \approx red\nabla p_b \times \nabla H bl$
2	24.8±1.2%, Interior flow (Fig. **??**c)	$red\nabla \times \tau_{sb} black \approx blue\nabla p_b \times \nabla H + \nabla \cdot (f\mathbf{U})$
3	14.6±1%, Quasi-Sverdrupian (Fig. **??**e)	$blue\nabla \times \tau_{sb} black \approx red\nabla \cdot (f\mathbf{U})$ (Fig. **??**f)
4	6.9±2.9%, Interior flow, vertical (Fig. **??**a)	$red\nabla \times \tau_{sb} black \approx blue\nabla \cdot (f\mathbf{U}) + \nabla p_b \times \nabla H$
5	1.9±1%, Interior flow, Southern Ocean (Fig. **??**c)	$blue\nabla \times \tau_{sb} black \approx red\nabla \cdot (f\mathbf{U}) + \nabla p_b \times \nabla H$
6-50	8.9 ± 0.3%, Dominantly non-linear (Fig. **??**e)	$red\nabla \cdot (f\mathbf{U}) black \approx blue\nabla \times \mathbf{A} + \nabla \times \tau_{sb}$ (Fi

Table 1. Percentage of area covered by the area specific balance of the BV equation (**??**) and the corresponding map figure. Leading order terms are sorted by magnitude, colors indicating if barotropic vorticity is added (redredblack) or removed (blueblueblack) by the leading order term, the corresponding bar chart figure shows the full breakdown. The quoted percentage coverage and StD is the mean of 100 runs of the algorithm.

Figure 1. The breakdown of the barotropic vorticity budget (ms^{-1}) over 1992-2013 in the ECCOv4 State Estimate.

Figure 2. Top figure illustrates the area selected by the clusters. Colors represent clusters in arbitrary order. The depth coherent ocean region (43%, Figure **??**a) in dark blue, Interior flow (24.8%, Figure **??**c) in light brown, Quasi-Sverdrupian (14.6%, Figure **??**e) in light green, Interior flow, vertical, (6.9%, Figure **??**a) in dark green, Interior flow, Southern Ocean, (1.9%, Figure **??**c) in lighter blue and the dominantly non-linear torque cover remaining 8.9% (Figure **??**e). Figure b illustrates that the AIC and BIC asymptoteing and we choose a K of 50 for our analysis. Error bars of 2σ capturing the stochastic seed. Figure c demonstrates the robustness of the algorithm with the ocean area, with 100 runs of the classification algorithm finding nearly identical areas (2σ error bars).

Figure 3. Maps of the selected locations (left) and corresponding area averaged histogram (right) of the terms in the BV equation. The colorbar is kept, but the color/ordering of the map are arbitrary. Colors in the barchart indicate if BV is added (redredblack) or removed (bluebllueblack).

Figure 4. Maps of the selected locations (left) and corresponding area averaged histogram (right) of the terms in the BV equation. The colorbar is kept, but the color/ordering of the map are arbitrary. Colors in the barchart indicate if BV is added (redredblack) or removed (blueblueblack).

Figure 5. Schematic of identified regions with names and cluster numbers. The depth coherent area implies a coherent vertical structure in Cluster 1. The quasi-Sverdrupian gyre in Cluster 3 is unique due to lack of BPT. The Inteior flow, vertical, in Cluster 4 has a stronger momentum driven portion of the BPT, and topographic interactions begin to become important. The interior flow in Cluster 2 has a stronger baroclinic component to the BPT and feels topography. The Interior flow, Southern Ocean, in Cluster 4 is like the Interior flow in Cluster 2, but with contributions of opposite sign. The remainder is dominated by non-linear contributions, and the barotropic interpretation is not appropriate.