A hierarchical ensemble manifold methodology for new knowledge on spatial data: An application to ocean physics

Maike Sonnewald[1]

[1]Princeton University

April 4, 2023

Abstract

Algorithms to determine regions of interest in large or highly complex and nonlinear data is becoming increasingly important. Novel methodologies from computer science and dynamical systems are well placed as analysis tools, but are underdeveloped for applications within the Earth sciences, and many produce misleading results. I present a novel and general workflow, the Native Emergent Manifold Interrogation (NEMI) method, which is easy to use and widely applicable. NEMI is able to quantify and leverage the highly complex 'latent' space presented by noisy, nonlinear and unbalanced data common in the Earth sciences. NEMI uses dynamical systems and probability theory to strengthen associations, simplifying covariance structures, within the data with a manifold, or a Riemannian, methodology that uses domain specific charting of the underlying space. On the manifold, an agglomerative clustering methodology is applied to isolate the now observable areas of interest. The construction of the manifold introduces a stochastic component which is beneficial to the analysis as it enables latent space regularization. NEMI uses an ensemble methodology to quantify the sensitivity of the results noise. The areas of interest, or clusters, are sorted within individual ensemble members and co-located across the set. A metric such as a majority vote, entropy, or similar the quantifies if a data point within the original data belongs to a certain cluster. NEMI is clustering method agnostic, but the use of an agglomerative methodology and sorting in the described case study allows a filtering, or nesting, of clusters to tailor to a desired application.

A hierarchical ensemble manifold methodology for new knowledge on spatial data: An application to ocean physics.

Maike Sonnewald[1,2,3]

[1]**Princeton University, Princeton, New Jersey, USA**
[2]**University of Washington, Seattle, Washington, USA**
[3]**NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA**

ABSTRACT

Algorithms to determine regions of interest in large or highly complex and nonlinear data is becoming increasingly important. Novel methodologies from computer science and dynamical systems are well placed as analysis tools, but are underdeveloped for applications within the Earth sciences, and many produce misleading results. I present a novel and general workflow, the Native Emergent Manifold Interrogation (NEMI) method, which is easy to use and widely applicable. NEMI is able to quantify and leverage the highly complex 'latent' space presented by noisy, nonlinear and unbalanced data common in the Earth sciences. NEMI uses dynamical systems and probability theory to strengthen associations, simplifying covariance structures, within the data with a manifold, or a Riemannian, methodology that uses domain specific charting of the underlying space. On the manifold, an agglomerative clustering methodology is applied to isolate the now observable areas of interest. The construction of the manifold introduces a stochastic component which is beneficial to the analysis as it enables latent space regularization. NEMI uses an ensemble methodology to quantify the sensitivity of the results noise. The areas of interest, or clusters, are sorted within individual ensemble members and co-located across the set. A metric such as a majority vote, entropy, or similar the quantifies if a data point within the original data belongs to a certain cluster. NEMI is clustering method agnostic, but the use of an agglomerative methodology and sorting in the described case study allows a filtering, or nesting, of clusters to tailor to a desired application.

Keywords: Data mining, Unsupervised Learning, Model validation

Plain language summary

Within the Earth sciences data is increasingly becoming unmanageably large, noisy and nonlinear. Most methods that are commonly in use employ highly restrictive assumptions regarding the underlying statistics of the data and may even offer misleading results. To enable and accelerate scientific discovery, I drew on tools from computer science, statistics and dynamical systems theory to develop the Native Emergent Manifold Interrogation (NEMI) method. Nemi is intended for wide use within the Earth sciences and applied to an oceanographic example here. Using domain specific theory, manifold representation of the data, clustering and sophisticated ensembling, NEMI is able to highlight particularly interesting areas within the data. In the paper, I stresses the underlying philosophy and appreciation of methods to facilitate understanding of data mining; a tool to gain new knowledge.

Key points:

1: Few tools for data mining within the Earth Sciences use recent advances in methodologies despite data available becoming unwieldly

2: The method Native Emergent Manifold Interrogation (NEMI) is presented. NEMI scales and performs well on very complex and nonlinear data

3: I stresses the underlying philosophy and appreciation of methods to facilitate understanding of data mining; a tool to gain new knowledge

1 INTRODUCTION AND PROBLEM STATEMENT

In this manuscript I introduce a generic methodology to determine areas of interest in a dataset that can have arbitrarily complex and nonlinear covariance structures. For simplicity, the method is given a name: Native Emergent Manifold Interrogation (NEMI). Nemi was developed to address the need to identify patterns and perform 'data mining' in the increasingly large, highly complex and complicated data that is becoming common within the Earth sciences. Due to the challenges posed by modern data, traditional methods of analysis are often inadequate, meaning that they fail to converge or offer little insight. NEMI blends dynamical systems theory with clustering, but importantly invites room at key areas for domain specific input 'native' to the research problem NEMI is applied to. With NEMI, I address the issue of mismatching 'data science' methods and data, where the practitioners of Earth science or more computational sciences often suffer under the difficulty of interdisciplinary communication. NEMI is a generalisation of the methodology in Sonnewald et al. (2020) that targeted plankton ecosystems, in that is is designed to scale to larger datasets. Scaling is one of the true bottlenecks in data mining for scientific applications. NEMI is generalised to work with any data, where the particular example application used here is geospatial data. I have used an explicitly hierarchical approach, making NEMI less parametric (fewer parameters to tune and less danger of noise interference) and intuitively useful both for global (for example the whole Earth in the present example) or more local applications (for example a basin or more regional assessment). Another novelty in NEMI is the lack of a fixed field-specific benchmark criteria (used in Sonnewald et al. (2020)), where I have generalised so a field agnostic option is available. Lastly, NEMI invites the use of a range of uncertainty quantification options in the final cluster evaluation. The intended readership of this manuscript are interested practitioners from the Earth sciences, meaning scientists interested in applying NEMI, with an interest in understanding the underlying philosophy and rationale beneath the architecture of the pipeline. I have attempted to describe concepts in detail and refer the interested reader to further materials. Here there are two main actors; the data and the methods of analysis. Oceanographic examples are used and an ocean numerical model dataset used as an example. Note that the present manuscript focuses on NEMI, and I do not include an general overview of data-mining within the Earth sciences (see Götz et al. (2015)), or provide a general overview of machine learning within the Earth sciences (see Fleming et al. (2021); Sonnewald et al. (2021); Beucler et al. (2021)).

The paper is structured as follows: to give NEMI context, I initially move through explaining the problems related to exploring data using machine learning methodologies, these being the data (section 2) and the methodologies (section 2.2) in very general terms. I move through a synthetic example of a simpler method to illustrate how, and why, this fails on more complex data (section 2.2.2). Then, in section 3, I move through the manifold-based projection for data cleaning, visualizing, and strengthening the associations between different components of the data. In section 2.2 I explain the actual clustering, and how this is chosen based on observations of the manifold. Sections 3 and 2.2 are part a) of NEMI as shown in Fig. 1. Section 4 illustrates the important step regarding how to treat and sort the resultant clusters for enhanced utility. Then, in section 5, I return to the issue of stochasticity, and demonstrate simple and more advanced methods to utilize this aspect of NEMI. Sections 4 and 5 are part b) of NEMI as shown in Fig. 1. Finally, section 6 provides an outlook on potential application and implications. NEMI is a method that can not only be applied to complex data, but is also flexible and verifiable. I refer to NEMI as the full workflow, but separate parts can be used and adapted as appropriate for the practitioner. The following provided code and examples uses the python programming language and key parameters are highlihgted. Note that parameters not discussed could be significant depending on the application, and the documentation should be read. The manuscript intends to give a thorough explanation of the reasoning behind NEMI, and aims to empower practitioners with the rationale behind different method choices so it can be applied to different data. I do not intend for the manuscript to stand entirely on its own as many methods NEMI draws from span a wide array of fields and, only brief explanations are within scope and should be seen as starting points for further reading. The code for NEMI is available on GitHub and also as a PyPi package: https://github.com/maikejulie/NEMI

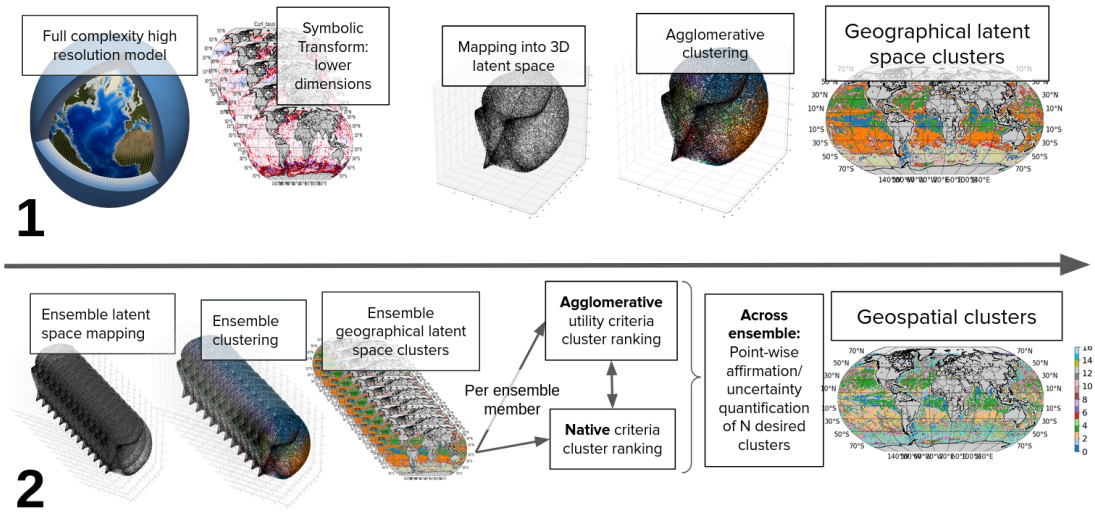

Figure 1. Sketch of workflow in NEMI. Sketch of NEMI workflow. Part 1 (top rwo) illustrates moving from the data in its rew form, through initial symbolic renditioning, manifold transformation and clustering. Part 2 (bottom row) shows the ensembling, agglomerative utility ranking and native (field specific) utility ranking within each ensemble member. Finally, the cluster for each location is determined looking across the ensemble. (Top left image of model adapted from encyclopedie-environnement.org).

2 DATA

Data is at the core of most discoveries within the Earth sciences and can come from numerical models or in situ or remote observations. However, what data to choose can be the most crucial step of any scientific endeavor.

2.1 The ocean data

In the illustration of a NEMI application in the present manuscript, I take data from an ocean model (MOM6, Griffies et al., 2023). Working on the full set of fields would have many parameters. The ocean model is discretized in latitude and longitude, as well as in depth, meaning that the model equations are solved on a grid that subdivides the ocean area and depth. The area covered within each grid point varies widely. Each data point approached naively would consist of one point in the depth, latitude, and longitude, where the model has 75 depth levels. We are interested in how the ocean is moving (as is described by the model equations in terms of momentum), and for each data point this amounts to 39 different fields for each depth level, where each field is one term in the equations that the model is solving, along with three additional ones at the sea floor. The equation terms can be thought of as our 'features' or 'dimensions'. As such, each location in latitude and longitude amounts to a vector of length 2989 entries or dimensions (39*75+3). See Khatri et al., 2023 for further details on the momentum budget closure in MOM6.

Working in a space consisting of 2928 dimensions is difficult, so we initially make the data more manageable using oceanographic theory. This can be thought of as simplifying the latent space within the data and is highly field-specific. This was described in detail in Sonnewald et al. (2019). For this simplification I use the barotropic vorticity (BV) equation terms as the data for NEMI, reducing the 2928 dimensions to five (Fig. 2). Although generally applicable to any data, NEMI was developed around the BV data as output from a fully realistic numerical ocean model. As such, the data that is used here is a parameter x that is a vector field defined at every grid cell (lon, lat) on the discretized MOM6 ocean sphere, with each element x_i representing a five-dimensional vector on the horizontal grid of the model. The index i uniquely identifies a grid point on the sphere, with (lon,lat) = (ϕ_i, θ_i). The features (dimensions) of each vector x_i correspond to the five terms in the BV budget. For the interested reader a description of the BV equation and how it relates to the numerical model follows. Skip ahead for further discussion of NEMI.

Early works (Sverdrup, 1947; Munk, 1950; Stommel, 1948) recast the intractably complicated full equations to describe how meridional ocean flows develop by taking the curl of the depth-integrated

momentum equations, and arriving at the barotropic vorticity (BV) equation. The steady BV balance under incompressibility is expressed as:

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

where $\beta = \partial f / \partial y$ is the northward derivative of the Coriolis parameter (f), $V = \int \rho v \mathrm{d}z$ is the depth-integrated northward mass transport from density ρ and meridional velocity v, ∇ is the horizontal gradient operator, p_b is the pressure at the bottom, and $H = h + \eta$ is the bottom depth. H is the water column thickness, an h is the distance from the resting ocean surface to the bottom topography and η the sea surface height anomaly. The stress produced by wind and bottom friction (external) is denoted by τ, and $\mathbf{A}$ and $\mathbf{B}$ are the depth integrals of the nonlinear and the horizontal viscous terms, respectively (Hughes and de Cuevas, 2001).

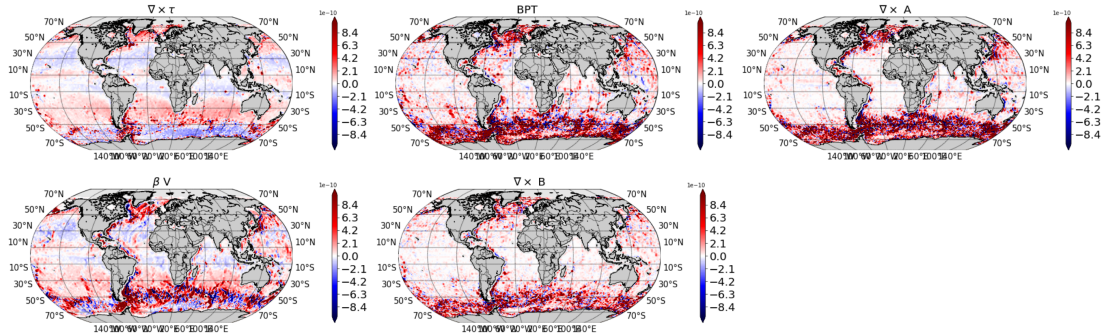

Figure 2. The terms of the barotropic vorticity equation. Each term is in ms^{-1}. Note how certain areas have clear large spatial patterns, while others can be highly variable. Top from left: $\nabla \times \tau$, $\nabla \times (p_b \nabla H)$ and $\nabla \times \mathbf{A}$. Bottom from left: $-\beta V$ and $\nabla \times \mathbf{B}$. See Fig. 3 for close-ups illustrating the complexity of the data further.

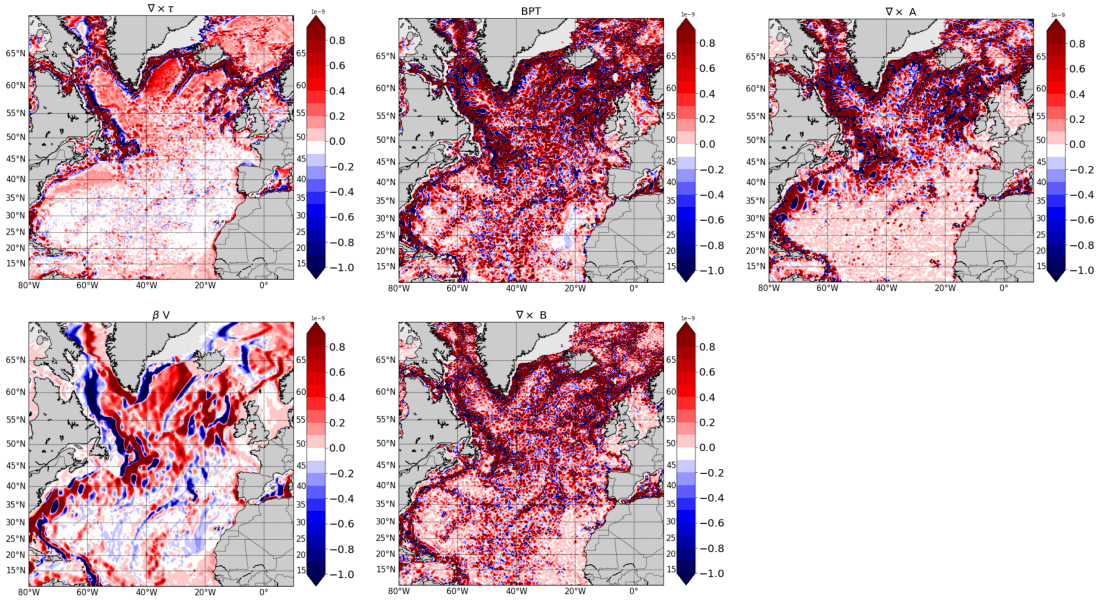

Figure 3. The terms of the barotropic vorticity equation, North Atlantic section. Each term is in ms^{-1}. Terms labeled as in Fig. 2.

2.1.1 Common problems with realistic data

In the example featuring the BV data from the numerical ocean model MOM6, NEMI is applied to a very complex problem. In general terms, data can suffer from issues such as: 1) noise, which is meaningless data that mask components of interest and sources can include instrument error or numerical artifacts; 2) sparseness, where only part of the desired data is available and examples include the wealth of data available at the ocean surface, but difficulty in acquiring subsurface data; 3) unbalance, which refers to data that has a wide range and only a small proportion of information of interest, for example a global dataset where one wishes to detect episodic ocean convection. The extent to which data is afflicted by these issues is often unknown, and checking the nature of the data extensively before starting analysis is always advisable. For the BV data, issues 1 and 3 are of note. The raw data, here only a smoother with a Gaussian kernel with a standard deviation of 1, is presented in Fig. 4 as a 'pairplot' (terminology from the 'seaborn' python library). The pairplot shows each dimension (here each term in the BV equation) as a scatterplot of each other term, with the associated probability distribution function of the dimension as a barplot. Such a pairplot is a nice way of initially assessing what issues are present within the data. In Fig. 4, the points are unreasonably centered around zero, and it suffers from outliers.

For NEMI, as is generally advisable, the data must by appropriately cleaned and pre-processed. Standardizing and normalizing are standard; for example, one can scale as $z = (x-u)/s$, where z is the scaled data, x is the original data, u is the mean and s is the standard deviation. This is done separately for each dimension, or equation term. Applied to the BV data, we arrive at the pairplot shown in Fig. 5 that reveals different structures. Note that the individual distributions only give a vague representation of data density. Many other methods for data-scaling exist that are suited for e.g., log distributed data. Experimenting with the initial scaling can be highly beneficial. The rationale behind scaling and normalizing is that the covariance between variables is much more interesting than their individual magnitudes. For example, consider the global data of ocean temperature and fish stock abundance, where the magnitude of variability in temperature is small compared to the magnitude of variability in fish stock abundance. Without scaling, the temperature variable would appear meaningless for fish stock abundance, even though we expect a difference between Arctic and tropical regions. After getting to know the data through initial inspection and scaling, we are ready to consider methodologies for further exploration.

2.2 Methods for data mining from unsupervised learning

Novel methods from 'data science' are increasingly being used to great advantage. In Sonnewald et al. (2021) a review of current progress and a brief introduction of methods can be found focused on physical oceanography. However, matching methods to data and robustly verifying their results requires knowledge both of the algorithm and the application. A computer scientist may believe she has arrived at a significant and interesting answer, but this may not be useful to an earth scientist if, for example, the uncertainty related to the spatial position of an identified region is too great. Along with a method's power must also come an appropriate level of skepticism and emphasis on validation, statistical or otherwise. NEMI is an answer to the issue of not having satisfying metrics to use to determine statistical significance of clustering results. Clustering is the task of dividing data into sub-groups so that data points within each group are similar and dissimilar to the data points in other groups. Clustering is largely regarded to be an 'unsupervised' machine learning methodology, meaning that the data is given to the method without explicit 'labels'. Clustering can be seen as the act of determining labels that can then be interpreted and offer insight to the practitioner.

There is a large and growing number of clustering algorithms available. It is beyond the scope of this article to give an overview of these. In general terms, common to all clustering methodologies is that the act of seeking to determine sub-groups within the data means that the differences between the overall data and the sub-sets that the clustering method has chosen should be determined. Put differently, every methodology should evaluate the differences between the overall data and clustered sub-sets. Note that such partitioning of the data's covariance space is also the backbone of, for example, neural networks. When applying a clustering algorithm, the resultant 'model' is the algorithm and the chosen values for any parameters or similar, referred to as clustering model hereafter. As such, it is critical to quantify how well the clustering model is able to represent the data. This is also seen in simple regression. Within clustering and regression, we search for an underlying and general model, or formula, to describe the data. To illustrate, the left column in 6 shows an underfitted model, attempting to partition the data with a straight line. This misclassifies large amounts of data. The rightmost column illustrates a model fit

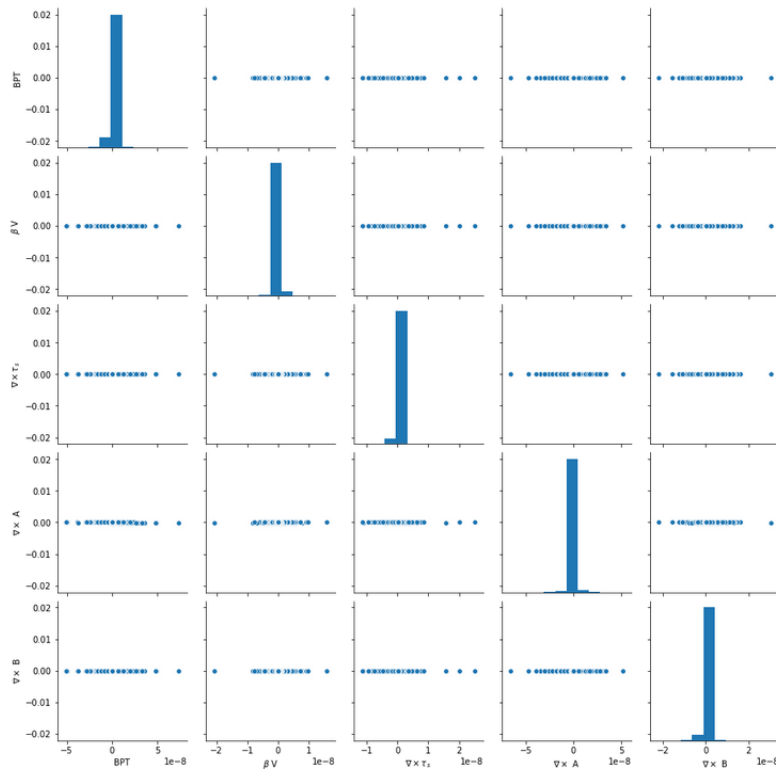

Figure 4. A basic look at the data. Each variable in the BV data is plotted against the other. Note that the data is not scaled, and the odd lack of structure indicates that the data is heavily skewed towards smaller numbers. See Fig. 5 for the scaled example.

where every point is accounted for, which also fails to reasonably approximate the underlying model. The middle column illustrates a general fit that represents a model that closely approximates the underlying model from which the data was drawn.

2.2.1 Validation

To validate a clustering application, showing that we have successfully discovered a reasonable representation of the underlying model, there are two main techniques: 1) external, and 2) internal validation. External validation requires a subset of the data to have known labels to compare to. Internal validation revolves around cohesion within a cluster and the degree of separation between different clusters. If the cohesion within a cluster is bigger than the degree of separation between clusters, then the clustering method is successful. However, recall the problem of overfitting. Many methods for verification of model skill exist, including the Silhouette coefficient, the Calisnki-Harabasz coefficient, the Dunn index, the Xie-Beni score, the Hartigan index, and the use of information criteria. It is beyond the scope of this article to go through all the above, but the example below will briefly introduce information criteria.

2.2.2 Practical example: k-means on idealised and BV data

A very popular method for clustering is called k-means (MacQueen, 1965). It is fast and conceptually simple, making it an excellent first choice for data exploration. 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. To initialize, the k-means algorithm makes a stochastic guess. This means that points are initially scattered across the data, and the algorithm iterates until a "maximum" in found. This maximum is determined by minimizing the objective function J:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||\mathbf{x}_i^j - \mathbf{c}_j||^2,$$

where k is the number of clusters, n is the number of data points, the vector $\mathbf{x}_i$ correspond to the five

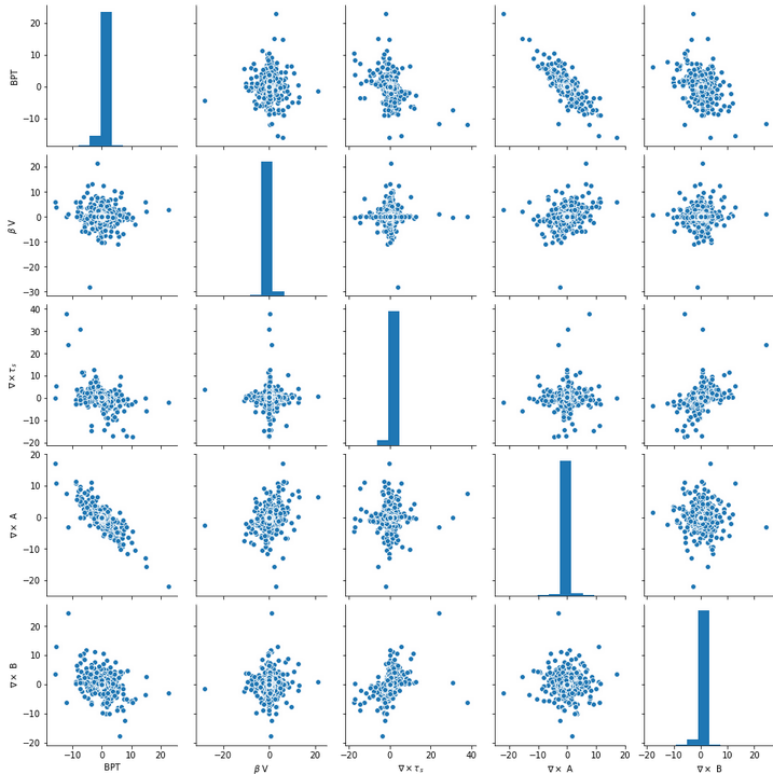

Figure 5. The scaled BV data. Each variable in the BV data is plotted against the other. Note that compared to Fig. 4 much more structure is visible.

terms in the BV budget, and $\mathbf{c}_j$ is the estimated location of cluster j. The number of k clusters is a free parameter that is chosen before the algorithm is applied. Initially, the cluster centers have random values scattered throughout the parameter space. Each cluster $j = 1, \ldots, k$ is represented by the five-dimensional characterizing vector $\mathbf{c}_j$, and the k-means classification attributes each vector $\mathbf{x}_i$ to a unique cluster c_j, so $\mathbf{x}_i = \mathbf{x}_i^j$. The distance between a data point is given by $\mathbf{x}_i^j$ and the cluster center $\mathbf{c}_j$ is determined as: $||\mathbf{x}_i^j - \mathbf{c}_j||^2$. In this way, each data point in $\mathbf{x}$ is associated with the closest k-cluster. Then, the position of $\mathbf{c}_j$ is calculated again, and the association is reassessed until the solution converges.

Note that k-means uses only one parameter (the number of clusters) and an initial stochastic guess for the cluster centers. Effectively, k-means clustering minimizes within-cluster variances (squared Euclidean distances), which also entails that k-means would work *perfectly* if the data were separated into tidy clumps with Gaussian distributions (round). Unfortunately, very few data have this type of covariance space and suffer from interconnected and decidedly non-Gaussian (and nonlinear) statistics. Put differently, the strength but also the weakness of this clustering method is that it works by partitioning the data into Voronoi cells. Effectively, the algorithm can only draw straight lines to partition the data and cannot isolate any more complex covariance structures.

In contrast to the BV example, Fig. 7 shows an idealized scenario that illustrates how k-means can be successfully used. The top left panel shows a dataset of tightly clustered points that are well-separated from neighbouring clusters and each has a Gaussian distribution (round). The top middle panel illustrates how k-means is successfully applied to discover this correct underlying structure in the data (here called 'true'). The top right panel shows the *same* data but with noise added. Using the labels discovered from the 'true' underlying model in the top middle panel, the bottom left panel shows where the data *should* be classified. The middle lower panel contain the classification results. The colors are arbitrary, but note that while there is some misclassification, the performance of the model determined using k-means is reasonable.

Each run of the k-means algorithm on the BV data results in one statistical 'model'. As such, we want to assess how well different models fit the data, where we can vary the number of clusters (k) and

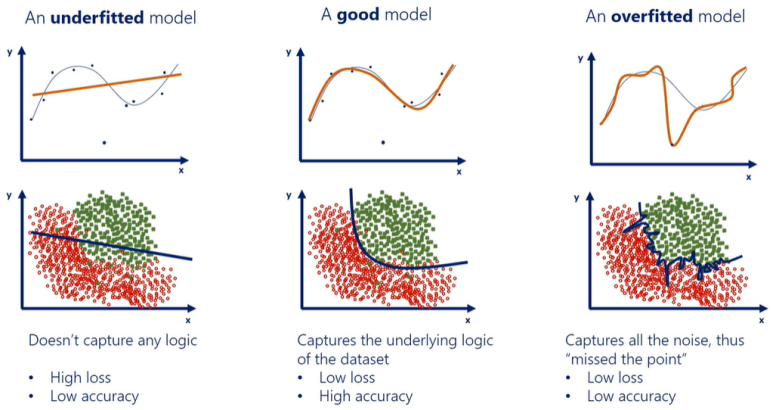

Figure 6. Figure illustrating a model fitting exercise. Adapted from 365datascience.com [1].

different model initializations, which give different results due to the initial stochastic guess (note that some implementations use a fixed seed which masks this potential source of model error). To assess the fit of the model, we apply validation methods. Here, the 'fit' of the model is assessed using information criteria (IC). The IC can tell us how 'complex' our k-means model should be, where increasing the number of k is equivalent to increasing the complexity. The IC illustrates if we are capturing **more information** if we add another k, or if the maximum is reached and the model is complex enough. Recalling the example in Fig. 6 where accounting for every data-point is not desirable, and this statistical model of the data is too complex, meaning it will not generalise well.

Different methods exist for estimating the IC, and here I will discuss the Akaike IC and Bayesian IC (AIC and BIC respectively). The AIC and BIC have been very well studied, and are therefore preferred. The maximum likelihood function is the basis of both and is the primary tool for estimating the parameters of an assumed probability distribution given data (here the data we cluster on). The likelihood ($\mathscr{L}$) is defined as:

$$\mathscr{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x),$$

Here, X is a discrete random variable with probability mass function p depending on a parameter θ. If thought of as a function of θ, it is the likelihood function, given the outcome x of the random variable X. Suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model (for example the number of cluster guesses from k-means). Then, $\hat{L}$ is the maximum value of the likelihood function for the model. Then the AIC value is estimated as follows:

$$\text{AIC} = 2k - 2\ln(\hat{L}).$$

With a set of different candidate models (for example, comparing models determined using different numbers of k clusters), the AIC with the lowest number will be the one that fits the data best. The goodness of fit is assessed by the likelihood function. To discourage overfitting, the penalty term ($2k$) increases as the complexity (number of k clusters) increases. As such, the AIC will in general asymptote, and a good model is determined when this happens.

The BIC also uses the likelihood function to determine the goodness of fit, but uses a different penalisation to determine if the model is overfitted:

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L}),$$

where n is the sample size. As discussed by Yang (2005); Harvey (1982), the AIC can overestimate the order, where the BIC penalisation term discourages this more strongly. See figures in section 2.2 for an example of how a model can fail to find an underlying model. In short, the AIC should asymptote, while the BIC should start increasing. A number of k somewhere between these two (if they both occur) could offer a good fit.

It follows that the AIC and BIC are inappropriate if the number of k is unmanageably large, or is close to the number of data points when we have no reason to suspect it should. The relative simplicity of

the AIC and BIC compared to many other model validation methods demonstrates the difficult nature of assessing if a 'good' approximation of the underlying model has been found, and stresses the importance of applying common sense, additional checks, and caution. Note the AIC and BIC are useful in many applications of model selection, for example auto-regressive model estimation (Sonnewald et al., 2018) as is commonly used without validation of the chosen order. Use of a statistical model without assessment of how well the model approximates the data can be highly unfortunate, including that a model that is needlessly complex is chosen or vise versa as discussed in Sonnewald et al. (2018).

Returning to the idealised example in Fig. 7, the bottom right panel illustrates the use of the AIC and BIC, where the 'correct' number of k is 5. In Fig. 7 the AIC appears to have asymptoted, and the BIC to reach its lowest point before going upwards again. As such, 5 clusters are correctly identified as the optimal number.

In Sonnewald et al. (2019), using BV data on a 1° horizontal resolution ocean model (See Sonnewald et al. (2019) for details on the model), k-means was used to good effect. This was astonishing, as the success of k-means suggested that large proportions of the ocean had an underlying linear distribution. Sonnewald et al. (2019) both illustrated that there were dominant partitioning within the BV data at this relatively low resolution, but also that the data was relatively normally distributed. This was unexpected in an oceanographic context using data from a realistic model. The partitioning of the data has led to scientific insight as the ML effectively performs an empirical leading order analysis that can subsequently be explored.

Running an AIC/BIC check on BV data from MOM6 at 1/4° (Fig. 9) illustrates that k-means is an inappropriate method for exploring this data. This is evident in that the AIC has not asymptoted after adding even 350 k, and while the BIC has started turning upwards the standard deviation (shown in the blue shading) is fairly large. To illustrate spatially on the ocean, Fig. 8 shows the spatial patterns associated with k set to 50 and 200. Neither are helpful, and section 2.2 illustrates that k-means is actually doing with its inability to work with nonlinear data.

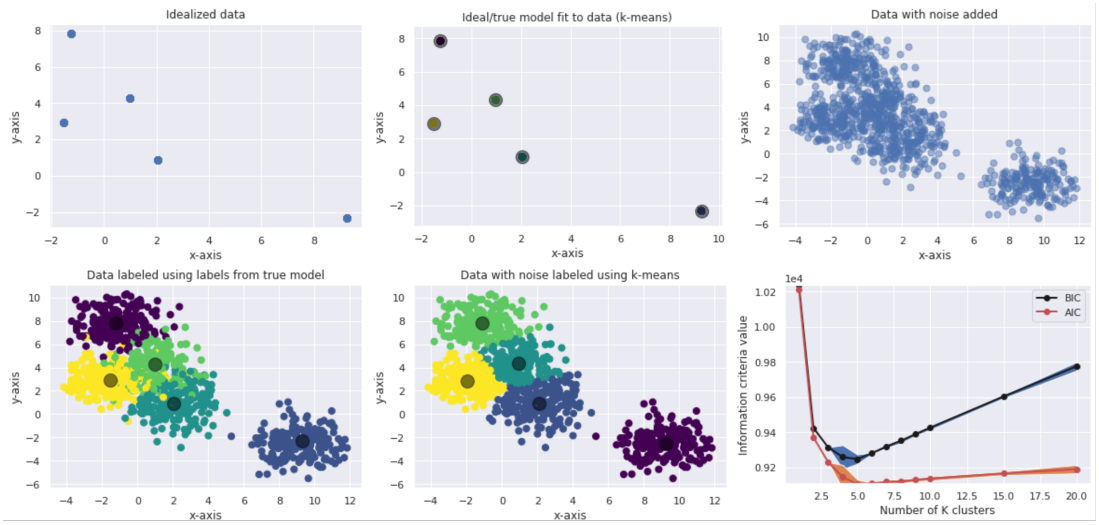

Figure 7. An illustration of concepts on idealised data.

3 MANIFOLD APPROXIMATION OF THE UNDERLYING COVARIANCE STRUCTURE

The NEMI methodology takes the approach that validation is of the utmost importance. As discussed above, validation can take multiple forms. The data-mining challenge that NEMI addresses uses a symbolic methodology to characterize the 'latent space'. The latent space is the covariance structures in the data hidden to our human perception. This can be imagined as how variables relate and change according to one another. The 'symbolic' methodology refers to reducing the size of the latent space from the original dimensions (over 270 for the closed momentum budget of the ocean model) down to a few

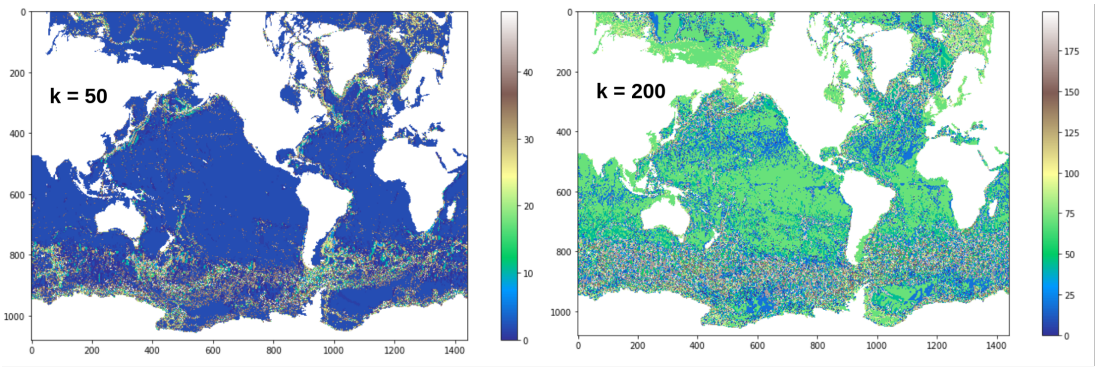

Figure 8. An illustration of running k-means on the BV data. To the left a k of 50 is chosen. To the right a k of 200 is used.

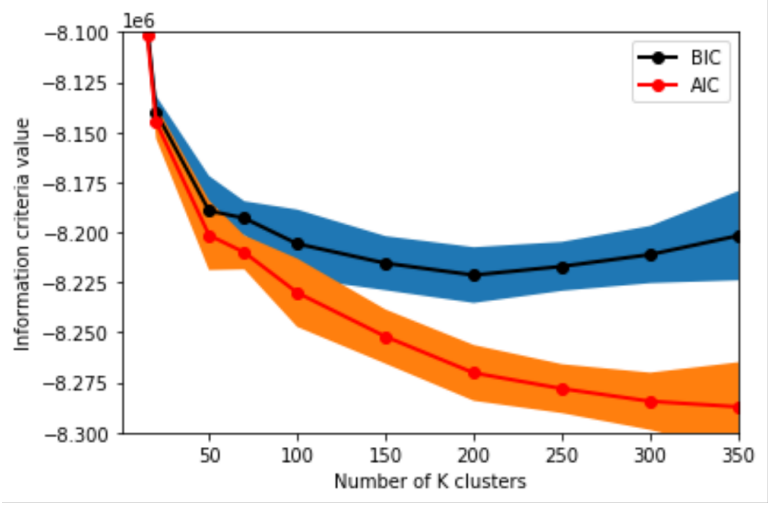

Figure 9. An illustration of the AIC and BIC criteria run on the BV data. Note that the AIC fails to converge and the BIC stays fairly flat. The AIC and BIC indicate that the k-means algorithm is not converging.

(i.e., five) using oceanographic theory. However, five is still too many for human perception. As such, we further characterize the latent space using a manifold methodology.

A mathematical manifold is a construct from topology: any local point resembles the Euclidean space near each point. Effectively, the 'distances' between different datapoints are used to determine relations. This has the convenient property, which makes it useful in NEMI, that the space is homeomorphic. This homeomorphism means that one shape can be transformed to another, without violating the relationships between the datapoints. One common example is that a doughnut (torus) can be transformed into a coffee mug, as both have one hole. It is beyond the scope of this article to give a thorough introduction to topology, but the key utility is that a confusing ball of covariances can efficiently be untangled **without losing the nonlinear structures**. For a visual example, imagine a scarf tangled on a table. The scarf has various patterns that may look oddly disjointed and tangled together. If you spread out the scarf, the complicated 3D structure becomes a smooth and fully visible (approximately) two-dimensional object. Any patterns on the scarf can be seen. In NEMI, the threads making up the scarf and its pattern are the barotropic vorticity equation terms.

Manifold methodologies have two further convenient properties: they can be used to reduce the dimensionality for visualization and 'strengthen' associations between different areas of the data, allowing patterns to emerge more clearly. NEMI employs the manifold methodology UMAP (Uniform Manifold Approximation and Projection McInnes et al. (2018)) as a processing step with considerable advantages. First, the UMAP methodology projects the data into three dimensions, meaning that the data can be

visualized. This allows an additional external validation step that will be discussed later. Second, the UMAP methodology works by assessing the connectedness of the data as described above. Third, through the UMAP application, the noise that posed an issue in the k-means application is lessened.

So what is a manifold in relation to actual data? We start with simple combinatorial building blocks (called simplices) of the distances between the data points. One data point is a 0 simplex, two connected points is a 1 simplex, three connected points is a 2 simplex (a triangle), a 3 simplex has four connected points (a pyramid), and we can continue upwards adding dimensions. We can construct different simplexes and combine these together, and in practice the simplexes do not need to have very high order to cover their local space. If this sounds similar to a k nearest neighbour graph (distinct from k-means) note that there the choice of the radii can have immediate detrimental impact on a graphs ability to approximate the underlying space, which is amplified if a dimensionality reduction is attempted (if the space were uniformly sampled it would work). The problem of having non-ideal data remains, and we can only assume that we are not uniformly distributed. Using Riemannian geometry the non-uniformness can be leveraged as fortuitous. In UMAP we **assume** that we have uniformly distributed data, and then use the actual distances between the data-points to create a map of the underlying manifold. Effectively, to map out the manifold we choose a unit 'ball' about a point stretches to the k-th nearest neighbor of the point, where k is the sample size we are using to approximate the local sense of distance (I use 'k' to conform with the overall machine learning literature, but note that this is distinct from the k in k-means). In UMAP, each point is given its own unique distance function. This lets us choose a number of 'neighbouring' data-points to use, rather than needing to determine the distance as k nearest neighbours would have required.

We now add to the concept of the manifold that it is locally connected, meaning that it describes one space, rather than a set of disconnected spaces. However, in a simplified sense, because we looked at the neighbouring points to assess the distances, two neighbouring points may individually have different values describing the same distance. As such, a useful mental construct with which to envision this set of UMAP is to think of it as a weighted graph, where the weights describe the distances. If there are conflicting weights associated with the simplices we will interpret the weights as the probability of the simplex existing.

Embedding the manifold into a lower-dimensional space can now happen based on the notion that we have the information about the manifold approximated by the data points, and we wish to conserve the associated probabilities between the data points in the lower dimensional space. In comparing the original topological structure of the manifold with a lower dimensional candidate. Both would share the same 0 simplices, and we can imagine that we are comparing the two vectors of probabilities indexed by the 1-simplices. For this we use the cross-entropy. In Information theory, the cross-entropy is a concept describing if two probability distributions are drawn from the same set.

To estimate the cross-entropy, say the set of all 1-simplices is E, and we have arrived at weight functions so the weight of the 1-simplex e is $w_h(e)$ in the *high* dimensional case. Now $w_l(e)$ is the weight of e in the *low* dimensional case, and the cross entropy will be:

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right).$$

Now, we minimize the cross-entropy to arrive at our low dimensional embedding of the migh dimensional manifold. Here, the first term $w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right)$ can be thought of as a force that attracts the points whenever there is a large weight associated in the high dimensional case. If $w_l(e)$ is as large as possible, the term will be minimized. This occurrs when the distance between the points is as small as possible, and effectively when the UMAP algorithm is focusing on the very local structure. In contrast, a repulsive force is found in the $(1 - w_h(e)) \log\left(\frac{1-w_h(e)}{1-w_l(e)}\right)$ term between the ends of e when $w_h(e)$ is small.

Key concepts to understand the limitation of such 'manifold' based methods is that we are assuming that our data points populate the manifold of the underlying model well enough. For example, if we think about a landscape with mountains, we may have less data points among the mountains than in the surrounding areas that are also 'smoother'. A manifold representation of the landscape based on this data would only roughly describe the true landscape in the mountaineous regions.

Now, what does a UMAP rendition of the highly complex and **complicated** BV data look like? In Fig. 10 a three dimensional rendition is demonstrated from different angles. The shape can vary depending

on the parameters chosen, as stressed above. In Fig. 10 we can see that there are clear areas that, from all angles, are more dense and some that are more sparsely populated. We will use the visualization for choosing a clustering algorithm below. The sensitivity to parameters (or how 'brittle' the method is) is highly dependent on how the data's complexity. In this example, a large ensemble sweeping through the parameters (described above) was needed to arrive at a reproducible manifold representation. The concept of a reproducible manifold means that one should be able to run the algorithm on the data and recover the same (or sufficiently similar) structure. Here, small differences can have large impact, and they can be difficult to pick up by eye. The importance of small differences is part of the reason why NEMI employs the additional checks and leverages the associated uncertainty. In Fig. 11 three renditions of running the UMAP algorithm on the processed BV data. The plots in panels a-c in Fig. 11 may look very similar to the human eye, but note the differences in the arrays. For example, the first number in the array goes from 7.895877 in the manifold in Fig. 11a, to 7.892489 and 7.875971 in b and c respectively. These differences may appear small, but they are present and can skew results. Determining the acceptable and appropriate level of difference is critical to the success of NEMI.

UMAP is similar to other methods such as t-SNE (t-Distributed Stochastic Neighbor Embedding, van der Maaten and Hinton (2008)). NEMI as presented here uses UMAP, but note that the t-SNE method was used in Sonnewald et al. (2020). Both UMAP and t-SNE have drawbacks, and one should weight carefully if these are appropriate for the data. These include that t-SNE, like UMAP, does not completely preserve density. UMAP, like t-SNE, can also create tears in clusters that should not be there, resulting in a finer clustering than is necessarily present in the data. Overall, such issues are exactly why NEMI was developed with additional validation steps. As such, NEMI uses both external and internal validation.

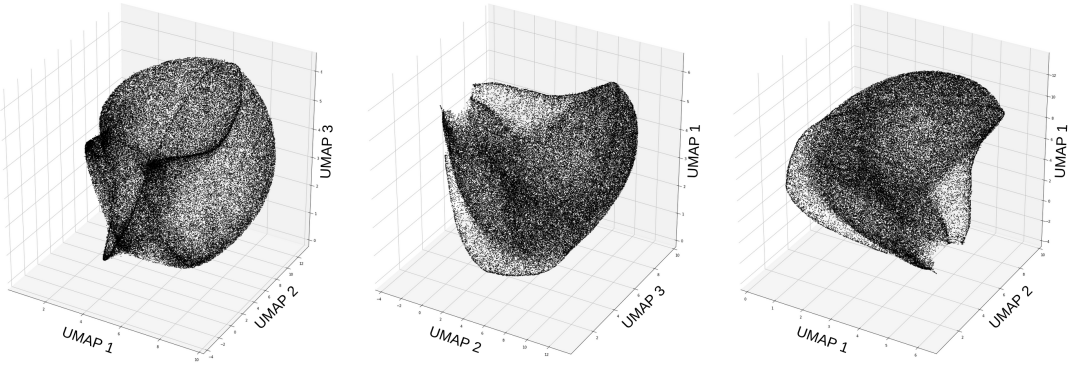

Figure 10. One UMAP manifold from different angles.

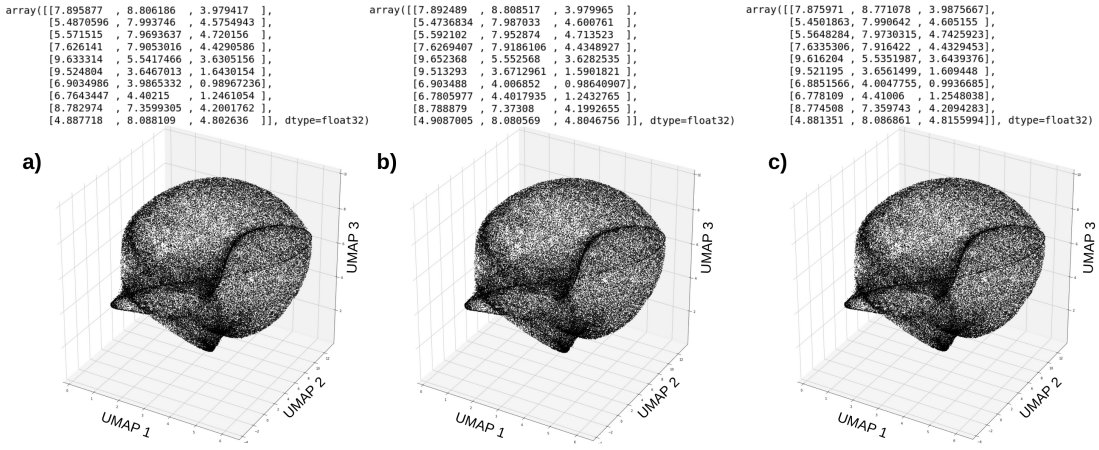

Figure 11. Three different ensemble members, with a part of the associated data. Note that while the manifold renditions look very similar, the data associated highlights the slight differences.

4 CLUSTERING: LEVERAGING MANIFOLD APPROXIMATION

The use of manifold and dimensionality reduction methodologies in NEMI leads to a very convenient three-dimensionsl visualization of the data (Figs. 10 and 11). In terms of clustering (or data-mining) this makes it visually apparent that an algorithm needs to be able to deal with data that is: 1) not well-separated (e.g., one continuous-seeming structure), 2) highly nonlinear, and 3) of varying densities meaning that the points are more likely to be found in certain areas.

There is growing number of different clustering algorithms available to the practitioner. For validation NEMI uses a hierarchical cluster analysis (HCA, the clusters found by the ML method will hereafter be refrered to as 'HCA clusters'), specifically an agglomerative methodology. Here, an agglomerative algorithm initially assumes that each data-point is its own cluster, and pairs of clusters are merged as one moves up the hierarchy. This is a "bottom-up" approach, where a "divisive" approach would be the opposite ("top-down") and assume that the initial step is to have one cluster represent the whole dataset and proceed to divide the data. Note that the agglomerative clustering methodology is not stochastic. The hierarchical element is useful as it means that running the algorithm on the same data will not introduce uncertainty in what clusters are found. In NEMI this refers to the same manifold rendition of the BV data. Using a hierarchical method is intuitively useful both for global (for example the whole Earth in the present example) or more local applications (for example a basin or more regional assessment).

The agglomerative hierarchical clustering methodology is presented as a cartoon in Fig. 12. Here, Fig. 12a shows the data points 1 to 6 in a two dimensional space (here 'UMAP 1' and 'UMAP 2' for simplicity in relating to the present section, although this is strictly a cartoon and UMAP was not applied). The data points have a certain distance to each other within this space. In Fig. 12b, initially each data point is progressively clumped together in relation to the distance between the points in panel a. As such points 4 and 5 and initially grouped, as are 3&2, while 6 and 1 remain isolated. In the next aggregation level, 6 is brought into the 5&4 cluster, becoming 6&5&4. The other points remain disaggregated as the distance between them is still too large (see Fig. 12a)). At the next level, the 3&2 and 6&5&4 clusters are merged into 6&5&4&3&2. Finally, on the next level the data point 1 is brought into the cluster, comprising now of the entire data set.

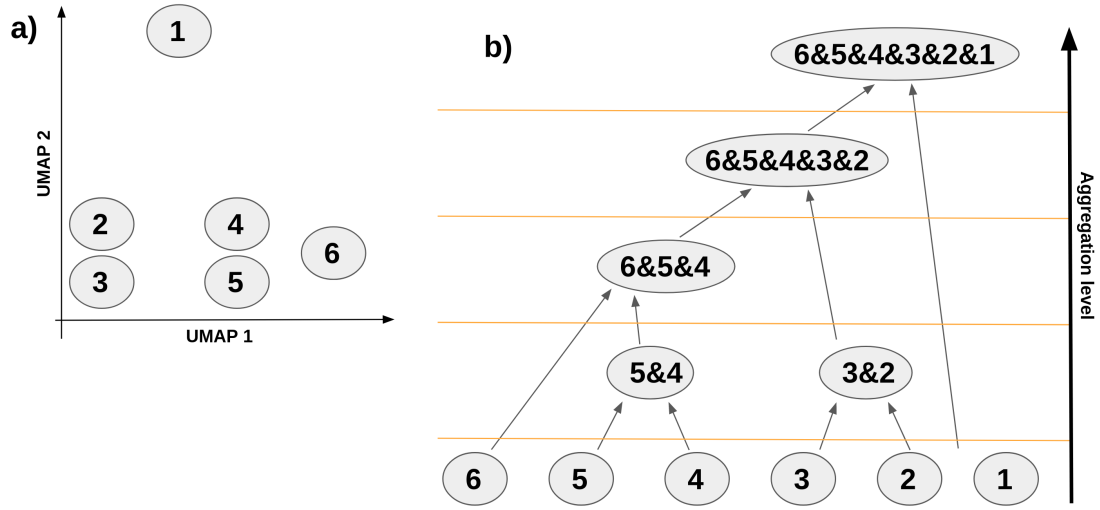

Figure 12. Sketch of the agglomerative clustering functions.

Having chosen an agglomerative methodology, I will highlight two hyperparameters that are of greatest relevance to the practitioner. The choices are that of the distance metric and linkage method. The distance metric is an expression of how the separation of the points is quantified. To illustrate, imagine a room of people, such as an auditorium with a lecturer on a podium and students sitting at a distance. If grouping the people using physical distance, the students would be clustered together because the gap between the students would be smaller than the distance to the lecturer. That is, the gap between the teacher and the closest students would be a defining feature of the data. However, if one used a metric such as how well the students know each other (e.g., how many friends they have in common),

there would likely be clear groupings within the students. As such, the distance metric chosen should be considered carefully. In NEMI, the use of the manifold methodology and a *closed budget* means that we can directly link the distance in UMAP space (seen in Figs. 11 and 10 as the distance between points) to the clustering. The use of budgets implies that a Euclidean methodology is appropriate. A Euclidean metric effectively uses Pythagoras's theorem in Cartesian coordinates. Next, the choice of the linkage method is often dictated by computational capacity. Note that methods scale differently with the size of the dataset. Here, the simplest method (single linkage) scales as $\mathcal{O}(n^3)$ where n is the number of points and should be avoided unless the dataset is very small. The second hyperparameter of interest, the linkage method, groups points. Recall the validation methodologies that look at internal versus external validation. Here, the linkage method can be seen in relation to the internal method. In NEMI we have used the Ward linkage method Ward (1963). Ward's method uses a minimum variance criterion that minimizes the total within-cluster variance. Let d be the distance between points i and j in data vector $\mathbf{x}$. The initial distances in Ward's method are Euclidean distances between points:

$$d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||^2.$$

Note that in NEMI the use of the 'cut' is equivalent to the direct number of clusters that are returned (the HCA clusters). So why not just use these? The reason for this was illustrated in Fig. 12, where in our BV data example the more 'extreme' outliers would be immediately focused on, and the wide swaths of the open ocean that are dynamically highly interesting would not be identified. For example constitute term balances that are opposite. The HCA clusters, approached naively, therefore have limited utility.

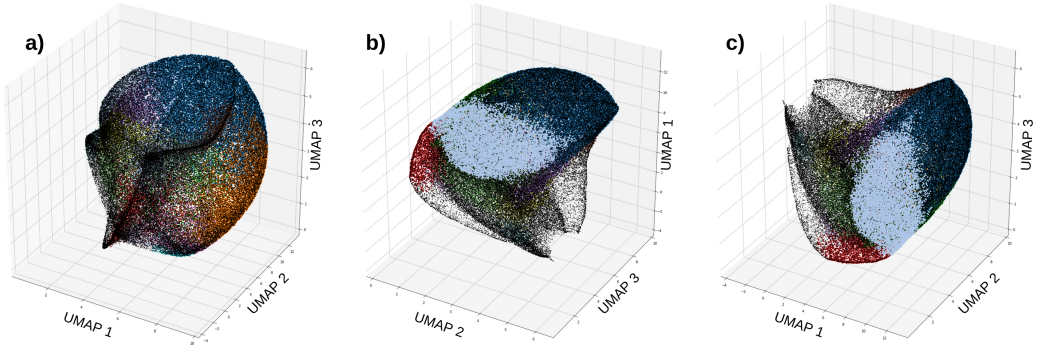

Figure 13. The agglomerative clustering on UMAP with 17 clusters. Panels a, b, and c show the same manifold from different angles. See sub-sampled version of a) in Fig. 14 to highlight shapes that are picked up by NEMI.

In Fig. 4 the application of the hierarchical clustering to a UMAP rendition is illustrated from a few angles (panels a-b are from the same manifold with the same clusters). The colors indicate the 17 different clusters (more detail on this below) and show how the clusters successfully isolate the ridges running along the sides of the data (see Fig. 14 for a sub-sampled version of panel a from Fig. 14 where details are highlighted). Note also that Fig. a displays one arbitrary iteration (i.e., ensemble member (i.e., ensemble member) of UMAP, with clusters determined on another UMAP ensemble member. In Fig. 15, a k-means rendition with 200 k (as looked visually reasonable in section 2.2.2) is displayed on the manifold used in Fig. 4 and 14 (the pale and translucent colours were chosen to enhance the readability due to the large number of colors). Note that the clustering was performed on the BV data before the UMAP algorithm and only subsequently projected onto the UMAP manifold. Each data-point is projected onto three dimensions from a five-dimensional space the locations are retained, the number of data-points remain the same, but the number of dimensions change. In Fig. 15 colors do not delineate the areas that are observed to be grouped together; this is a visual demonstration of how k-means fails to identify key regions. The figures illustrates what k-means does: the algorithm is applied to the manifold rendition in Fig. 16 and is forced to artificially separate the data coarsely using 'straight lines' across the entire data volume. Remembering that the UMAP rendition of the BV data is used to 'simplify' and 'clean' the data, it becomes apparent how difficult it would be to apply k-means to the non-transformed data. In supplement to the information criteria, this additional visual appraisal of the performance of the algorithm

underscores that the k-means algorithm is a poor choice. This method of validation can be applied widely beyond the examples used here.

As with most clustering and machine learning applications, there is no guarantee of finding the optimum solution. There might not even be one. However, if an optimum does exist for the agglomerated clusters, it is guaranteed to be found via single-linkage. Due to computational costs the application of single-linkage application is largely impractical. Other methods, such as the Density-based spatial clustering of applications with noise (DBSCAN Ester et al. (1996)) used in Sonnewald et al. (2020) can be useful, especially if the data is more separated. However, in this example arriving at a robust set of clusters was difficult using DBSCAN. Note that DBSCAN performs **considerably** better, in terms of scaling to larger datasets, so if possible this method is recommended.

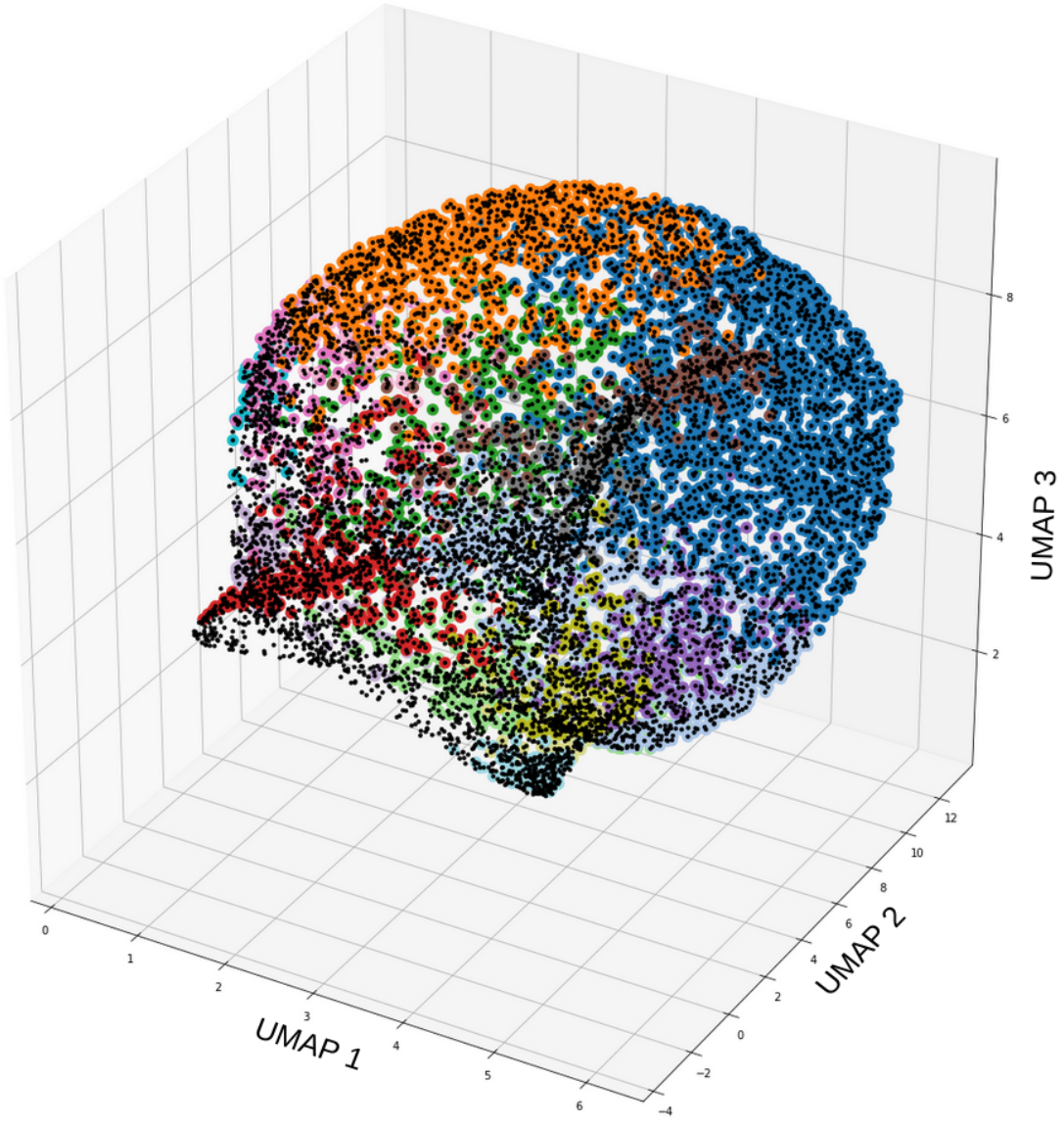

Figure 14. The agglomerative clustering on UMAP with 17 clusters, heavily sub-sampled. Illustration to supplement Fig. 4. Note I use an arbitrary ensemble member for the manifold and a different ensemble member for the clusters.

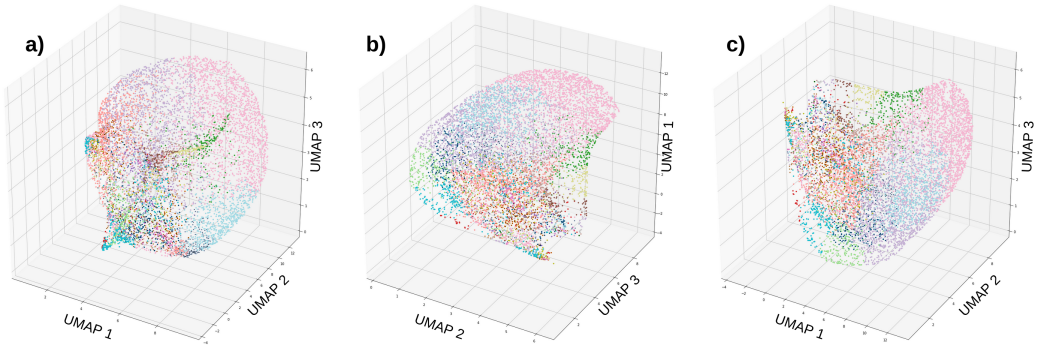

Figure 15. The k-means algorithm with k = 200 result projected onto a UMAP manifold. Panels a, b, and c show the same manifold from different angles. Note that the clusters should be coherent on the manifold if the method is successful. Note there is poor coherence and the clusters are somewhat arbitrarily separating chunks of the space. This confirms earlier suspicions that the k-means algorithm was not succeeding in arriving at a good model representation.

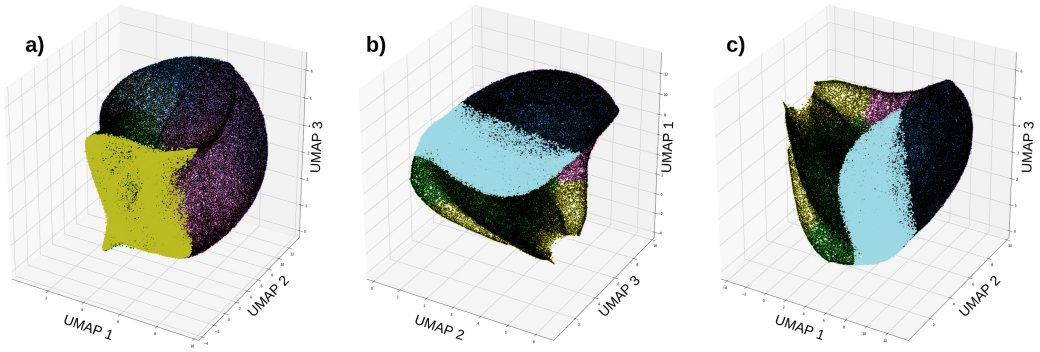

Figure 16. The k-means algorithm applied to a UMAP manifold. Panels a, b, and c show the same manifold from different angles. Here the impact of k-means is illustrated. Note how the manifold is artificially 'chopped' up in ways that clearly do not respect the data.

SORTING FOR DESIRED TRAITS: UTILITY AND VALIDATION

The visual check of the clustering algorithm chosen in NEMI as discussed in section 4 is the first step towards validation. However, the second validation step also builds on the hierarchical aspect of the clustering algorithm. For additional external validation in NEMI, we turn to established oceanographic theory and leverage that the data we are using is the BV budget. While this may appear specific to the BV data, it is generalizable such as in Sonnewald et al. (2020), who used the idea of 'provinces' in ecology and how they compared to established notions.

For validation and utility, let us return to a concept introduced in **?** in relation to cluster validation and assessment. The concept of using what is **useful** in an oceanographic context. Put differently, having a model that is a good fit to the data can be completely useless and misleading, for example, if a key parameter was missing from the data (think of the hydrodynamic paradox where missing boundary layer friction stood in the way of progress for over 100 years). A focus on the scientific problem at hand can be very powerful (in the hydrodynamic paradox this would be working on the equation terms **?**). Here, as in **?**, it is critical that the algorithm can robustly recover and reproduce geographical sub-regions. Namely, if the algorithm does not repeatedly recover the same geographical areas, the identified clusters, however reasonable it may look given statistical checks or other validation, have no utility. Ultimately, a criteria, defined here by the practitioner as finding the same spatial area, is the final objective. From Figs. 4, 14 and 11, it may seem surprising that the same area is not recovered precisely after each iteration of this component of NEMI. However, despite the precision apparent in the Figs. 4, 14 and 11, there is geographical variability. This variability, as discussed in the next section, is intrinsically useful.

The next step in NEMI is to sort the clusters for each UMAP rendition by spatial similarity. For this sorting, we weight by geographical extent is used as a weighting because large areal extents are seen as a relevant feature to favor. As such, the next component of NEMI sorts the clusters from each UMAP and agglomerative clustering iteration and then assesses which clusters are most similar in the geographical region covered (scaled by area covered which varies widely across the model grid) across the ensemble members.

In addition to sorting via coherent spatial cover across the ensemble of UMAP and agglomerative clustering repeats, the agglomerative methodology allows the selection of different aggregation levels, with NEMI having these be the number of HCA clusters. As such, NEMI is designed to be appropriate both for global and regional applications. Specifically, a practitioner in need of a globally representative set of clusters would select a small level of aggregation, while a regional application should choose a higher one.

In combination, the choice of aggregation level, as well as sorting by area size, allows one to select the **number of clusters**, together with the **spatial level** one is wishing to focus on. Note that it is up to the practitioner to determine a reasonable level and effectively number of clusters, as well as acceptable uncertainty/entropy (discussed below). Overall, note that this feature is of specific concern if working in the equatorial region compared to high latitude regions. Mid-latitudes see much less impact, as is expected.

The level of aggregation as well as the number of clusters is illustrated in Fig. 17. Three different ensemble members are shown separately (rows), with an aggregation level of 350 with 6 clusters in the first two columns, and an aggregation level of 350 with 20 clusters in the third and fourth columns. Columns one and three show the global ocean, and columns two and four show the North Atlantic. Note that the three members look very similar, particularly in their global distributions. I omit plotting the 350 clusters as this offers limited insight due the colour scale.

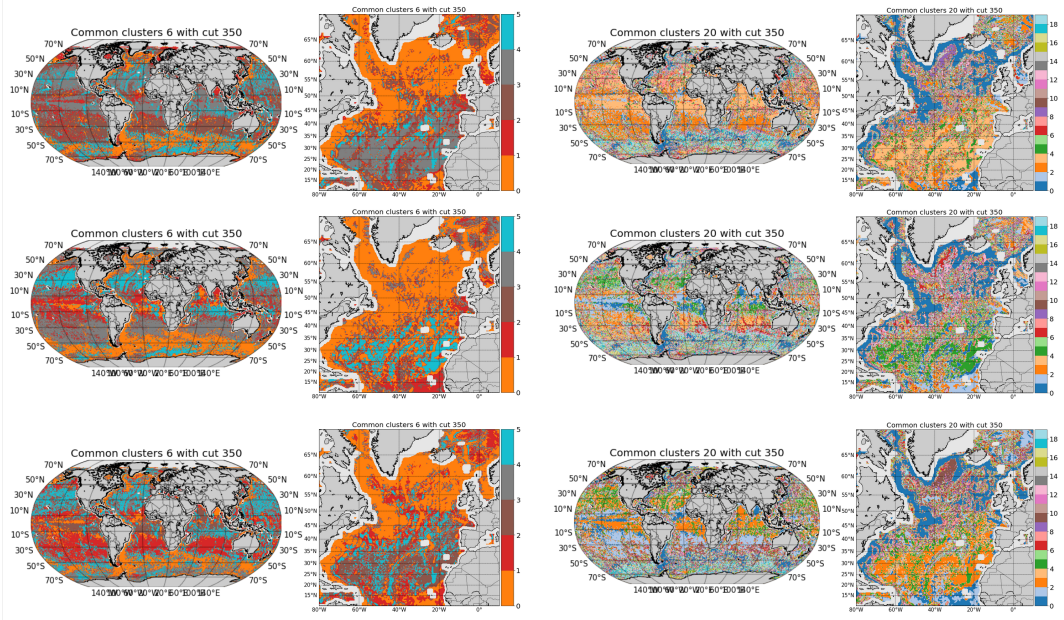

Figure 17. Demonstration of the changes in cluster locations within the ensemble. Three arbitrarily chosen different ensemble members are shown separately (rows), with an aggregation level of 350 with 6 clusters in the first two columns, and an aggregation level of 350 with 20 clusters in the third and fourth columns. Columns one and three show the global ocean, and columns two and four show the North Atlantic. Note plotting the 350 clusters offers limited insight due the colour scale.

Having determined the desired level of aggregation as well as number of clusters, validation via theory, or field-specific intuition should also occur. For example, within the BV budget certain balances are known and expected in certain regions. Specifically, a canonical balance between the windstress curl and advective component (see Sonnewald et al. (2019); Sonnewald and Lguensat (2021) for extensive

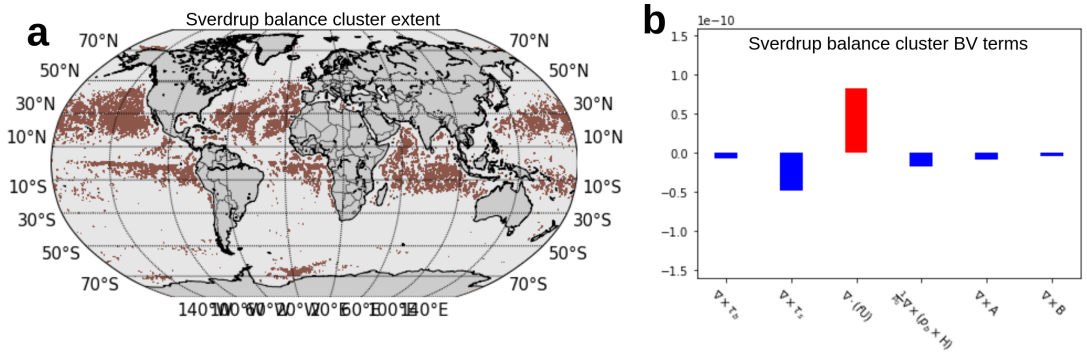

Figure 18. Figure showing canonical and expected balance for validation via expert judgement. The expected balance (meaning that the terms add up to zero) between the windstress curl and advective component can be seen. Fig. 18a shows its geographical extent, and Fig. 18b shows the area averaged terms balances in the locations NEMI highlighted.

description). If this balance between dominant terms is **not seen**, this does not necessarily invalidate the results, but it should mean that the results are treated with increased caution. For example, NEMI here is applied to a realistic coupled model **?**, where intuition and experience strongly suggests the balance between the windstress curl and advective component should emerge in the subtropics in the Northern Hemisphere (Munk, 1950; Sverdrup, 1947). In Fig. 18, just this balance (meaning that the terms add up to zero) between the windstress curl and advective component can be seen, where Fig. 18a shows its geographical extent, and Fig. 18b shows the area averaged terms balances in the locations NEMI highlighted. The exact locations where the balance does not hold (where there are other clusters mixed in, for example, over ridges) can lead to new studies and new scientific insight (For example **?**). As such, NEMI is an avenue towards generating new knowledge with machine learning. However, if this were a BV balance in an idealized channel-model set-up one would not necessarily flag the absence of this balance as suspicious. As a general tool, this step of NEMI requires field specific intuition, where the machine learning and scientist should interact to forge and identify new avenues of discovery.

5 LEVERAGING AND MANAGING NOISE

The issue of noise and stochasticity within data and methods may at first appear to be a challenge that only increases the difficulty of building applications interpeting them. In this section I will describe the final notion and step of NEMI and make a case that a stochastic-friendly methods are needed for crafting methodologies applied to 'real' data.

No data is perfect, and methods, like most from machine learning, must find optimal ways of approximating the 'underlying' model. However, as demonstrated in Fig. 6, being able to account for the slight variations, for example in the sine curve in the top middle panel, can improve a model's utility. Having a methodology that is able to reflect the uncertainty of the model fit can be highly beneficial. The two-dimensional examples in Fig. 6 are simple cases, but the highly nonlinear BV data poses a more difficult problem. In NEMI, as with any neural network application or optimization algorithm, the method application will determine the best fit given its initial conditions (i.e., parameters), including a stochastic or random seed. In many cases, a slight perturbation in initial conditions can lead to a different result, meaning a different model representation. In NEMI, this would be a different manifold, as was demonstrated in Fig. 11. What this sensitivity to initial conditions means in practice is that there are multiple landscape of possible solutions that the model can converge to and that these different states can be reached given just a small difference in parameters.

The sensitivity to parameters may appear to be a weakness in a methodology, and will be if a model of sufficient utility is not arrived. However, in the application of NEMI to the BV data the slight sensitivity to parameters allows the exploration of the complex covariance space of the BV data. Consequently, NEMI allows an estimation of the uncertainty. Thought of in the framework of bias versus variance, having a good approximation of the variance within the covariance space of the data a methodology describes is highly beneficial. The application of a manifold methodology facilitates this.

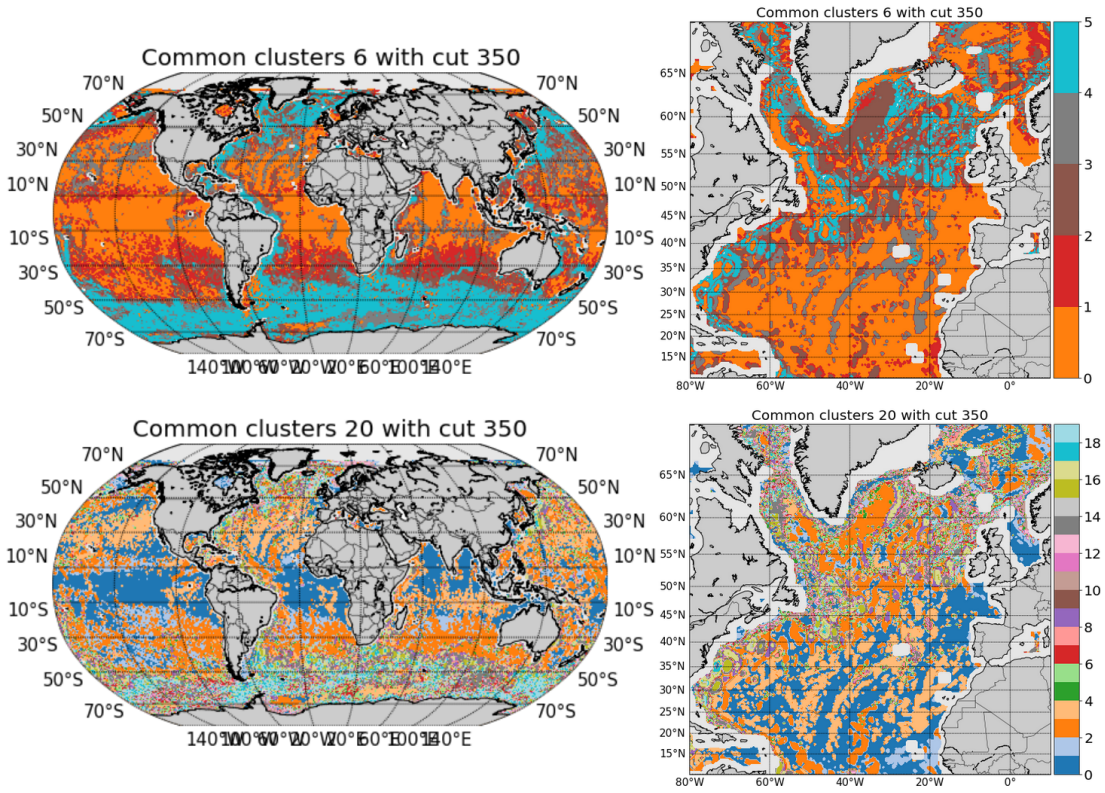

Figure 19. The clusters (BV dynamical regimes). For aggregation level 350, cluster numbers of 6 and 20 are shown. Note that in comparison to Fig. 17 the clusters here are much smoother.

A number of methods and approached can be used to estimate the uncertainty. In the BV example a geographic majority vote was used. This means that for each geographical location, the cluster that was most often flagged throughout the ensemble was the one chosen. Here, this was done largely for simplicity. Note that other methods, such as entropy are highly suitable as described in appendix A. Using entropy would allow the assessment of how *many* different clusters were chosen. If the majority was between two very different ones, this could be important information or If an area were highly contested then our confidence in that area would be lowered. Naturally, having this information is highly valuable in an of itself, and for the interested practitioner I recommend exploring this avenue and feature of NEMI.

5.1 Oceanographic interpretation of regimes

In Fig. 19 the product of applying NEMI to the BV data is shown. An HCA cluster aggregation level of 350 is chosen, and six (top row) and twenty (bottom row) clusters are demonstrated. Comparing to Fig. 19, we can see that while the figures look somewhat similar, the 'static' (the result of the clusters changing ever so slightly) has been greatly suppressed. Overall, the clusters are much smoother and crucially *reproducible*. To illustrate the utility of these choices of cluster numbers, I will briefly give two examples of the utility. Note however, that the number of clusters (here dynamical regimes) is entirely up to the practitioner and will likely depend on the research question at hand. To illustrate the now oceanographic context I will refer to the final cluster products as 'dynamical regimes' as these illustrate an objective empirical leading order analysis of the closed BV equation.

In Fig. 19, the top row shows the large overall dynamical regimes that are very interesting when assessing the global structures. Note that coherent areas in the areas where the wind stress curl ($\nabla \times \tau$) are largely coherent have been grouped together, despite having opposite signs. Note how we know from Fig. 18 and from oceanographic intuition that these areas should be similar but have opposite main drivers. For example, in the Northern Hemisphere in the large wind gyre areas (see Sonnewald et al., 2023 for a detailed theoretical description) the wind stress curl is negative and balanced largely by positive planetary advection. In similar areas in the Southern Hemisphere this effect is opposite, which is also

intuitive due to the symmetry of the Earth around the equator. From a clustering perspective, the fact that the terms are similar allows them to be grouped together, and different areas to stand out more. See for example the grey streaks through the North Atlantic running approximately latitudinally from 43 to 17°N. These are colocated with where significant areas of variability in the sea floor are found (for example the mid-Atlantic ridge). The clustering here illustrates what an important feature this is, and that this should be paid close attention to.

In the bottom row of Fig. 17 an example where there are 20 dynamical regimes is shown. Here, the area where the 'classical' wind gyres are found are seen and the Northern Hemisphere and Southern Hemisphere dynamical regimes are distinguished. Note the increased detail around, for example, the coast. As a thought example, imagine that a current is flowing along the coast. The coasts have large features such as canyons. Moving south to north, a current moving into a canyon would suddenly have more room, and the vorticity contributed by the bottom pressure torque would decrease significantly. As the current moves further north the other side of the canyon would be reached and the current would become more constricted again, where the bottom pressure torque term would increase. These would emerge as separate dynamical regimes in a study where a larger number of dynamical regimes is chosen, but most likely not appear in a study choosing a lower number.

Note that the two examples above use examples where one as 'equal but opposite' scenarios being grouped together. This was chosen as an accessible example but should by no means be seen as the only possible cancelling effect. Recall the complicated covariance space being queried and the highly nonlinear data. Further investigation of the dynamical regimes in the BV equation in MOM6 is the topic of another study.

6 CONCLUSION

Here, I presented the method Native Emergent Manifold Interrogation (NEMI), which is a generalisation of the methodology presented in Sonnewald et al. (2020). NEMI is designed for 'data mining', or put differently, to find underlying patterns within data. Nemi is a generalisation of the methodology in Sonnewald et al. (2020) that targeted plankton ecosystems, in that is is designed to scale to larger datasets. Scaling is a formidable bottlenecks in data mining for scientific applications. In NEMI I have generalised a workflow that can accommodate a wide array of data, where the particular example application used here is geospatial data. An explicitly hierarchical approach is used, making NEMI less parametric (fewer parameters to tune and less danger of noise interference) and intuitively useful both for global (for example the whole Earth in the present example) or more local applications (for example a basin or more regional assessment). NEMI does not use a fixed field-specific benchmark criteria (used in Sonnewald et al. (2020)), but is generalised so a field agnostic option is available. Lastly, NEMI invites the use of a range of uncertainty quantification options in the final cluster evaluation, from a majority vote to entropy. I demonstrate NEMI's application to a numerical ocean model, namely MOM6 (Griffies et al., 2023), and represent the barotropic vorticity balance of a time-mean of a model run. Here, the data serves as an example of a highly nonlinear and complicated covariance structure, within which reside highly valuable oceanographic patterns. NEMI is used to extract these patterns and facilitate further scientific discovery. However, NEMI is entirely general, and can be used on a range of data from the Earth sciences and beyond.

OPEN RESEARCH

The code for the Native Emergent Manifold Interrogation (NEMI) method is available here:
https://github.com/maikejulie/NEMI
DOI: 10.5281/zenodo.7764719

ACKNOWLEDGMENTS

I gratefully acknowledge students and colleagues who have requested details regarding this methodology to the extent that writing it down seemed appropriate.

Funding: Cooperative Institute for Modeling the Earth System, Princeton University, under Award NA18OAR4320123 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and

do not necessarily reflect the views of Princeton University, the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce.

APPENDIX

A: Entropy for uncertainty estimation

Entropy (H) can be used as a measure of uncertainty. As discussed in Clare et al. (2022): In information theory, entropy is the expected information of a random variable, and for each sample i is given by

$$H_i = -\sum_{j=1}^{N_l} p_{ij} \log(p_{ij}), \tag{2}$$

here N_l is the number of possible outcomes for each location and p_{ij} is the probability of each outcome j for sample i (Goodfellow et al., 2016). The larger the entropy, the less skewed the distribution will be and the more uncertain the outcome. The concept of entropy can be directly applied to manage the potentially different results from NEMI for each geographic location within the ensemble. If this is better than a simpler method, such as a majority vote, depends entirely on the application.

REFERENCES

Beucler, T., Ebert-Uphoff, I., Rasp, S., Pritchard, M. S., and Gentine, P. (2021). Machine learning for clouds and climate (invited chapter for the agu geophysical monograph series "clouds and climate").

Clare, M. C., Sonnewald, M., Lguensat, R., and Deshayes, J. (2022). Explainable artificial intelligence for bayesian neural networks: toward trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, 14(11):e2022MS003162.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise.

Fleming, S., Watson, J., Ellenson, A., Cannon, A., and Vesselinov, V. (2021). Machine learning in earth and environmental science requires education and research policy reforms. *Nature Geoscience*, 14.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Götz, M., Richerzhagen, M., Bodenstein, C., Cavallaro, G., Glock, P., Riedel, M., and Benediktsson, J. A. (2015). On scalable data mining techniques for earth science. *Procedia Computer Science*, 51:2188–2197. International Conference On Computational Science, ICCS 2015.

Harvey, A. C. (1982). Spectral analysis and time series, m. b. priestly. two volumes, 890 pages plus preface, indexes, references and appendices, london: Academic press, 1981. price in the uk: Vol. i, £49-60: Vol. ii, £20-60. *Journal of Forecasting*, 1(4):422–423.

Hughes, C. W. and de Cuevas, B. A. (2001). Why Western Boundary Currents in Realistic Oceans are Inviscid: A Link between Form Stress and Bottom Pressure Torques. *Journal of Physical Oceanography*, 31(10):2871–2885.

MacQueen, J. (1965). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Math., Stat., and Prob*, page 281.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861.

Munk, W. H. (1950). ON THE WIND-DRIVEN OCEAN CIRCULATION. *Journal of Meteorology*, 7(2):80–93.

Sonnewald, M., Dutkiewicz, S., Hill, C., and Forget, G. (2020). Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science advances*, 6(22):eaay4740.

Sonnewald, M. and Lguensat, R. (2021). Revealing the impact of global heating on north atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, 13(8):e2021MS002496. e2021MS002496 2021MS002496.

Sonnewald, M., Lguensat, R., Jones, D., Düben, P., Brajard, J., and Balaji, V. (2021). Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, 16.

Sonnewald, M., Wunsch, C., and Heimbach, P. (2018). Linear predictability: A sea surface height case study. *Journal of Climate*, 31:2599–2611.

Sonnewald, M., Wunsch, C., and Heimbach, P. (2019). Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, 6(5):784–794.

Stommel, H. (1948). The westward intensification of wind-driven ocean currents. *Transactions, American Geophysical Union*, 29(2):202–206.

Sverdrup, H. U. (1947). Wind-driven currents in a baroclinic ocean; with application to the equatorial currents of the eastern pacific. *Proceedings of the National Academy of Sciences*, 33(11):318–326.

van der Maaten, L. and Hinton, G. (2008). Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

A hierarchical ensemble manifold methodology for new knowledge on spatial data: An application to ocean physics.

Maike Sonnewald[1,2,3]

[1]**Princeton University, Princeton, New Jersey, USA**
[2]**University of Washington, Seattle, Washington, USA**
[3]**NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA**

ABSTRACT

Algorithms to determine regions of interest in large or highly complex and nonlinear data is becoming increasingly important. Novel methodologies from computer science and dynamical systems are well placed as analysis tools, but are underdeveloped for applications within the Earth sciences, and many produce misleading results. I present a novel and general workflow, the Native Emergent Manifold Interrogation (NEMI) method, which is easy to use and widely applicable. NEMI is able to quantify and leverage the highly complex 'latent' space presented by noisy, nonlinear and unbalanced data common in the Earth sciences. NEMI uses dynamical systems and probability theory to strengthen associations, simplifying covariance structures, within the data with a manifold, or a Riemannian, methodology that uses domain specific charting of the underlying space. On the manifold, an agglomerative clustering methodology is applied to isolate the now observable areas of interest. The construction of the manifold introduces a stochastic component which is beneficial to the analysis as it enables latent space regularization. NEMI uses an ensemble methodology to quantify the sensitivity of the results noise. The areas of interest, or clusters, are sorted within individual ensemble members and co-located across the set. A metric such as a majority vote, entropy, or similar the quantifies if a data point within the original data belongs to a certain cluster. NEMI is clustering method agnostic, but the use of an agglomerative methodology and sorting in the described case study allows a filtering, or nesting, of clusters to tailor to a desired application.

Keywords: Data mining, Unsupervised Learning, Model validation

Plain language summary

Within the Earth sciences data is increasingly becoming unmanageably large, noisy and nonlinear. Most methods that are commonly in use employ highly restrictive assumptions regarding the underlying statistics of the data and may even offer misleading results. To enable and accelerate scientific discovery, I drew on tools from computer science, statistics and dynamical systems theory to develop the Native Emergent Manifold Interrogation (NEMI) method. Nemi is intended for wide use within the Earth sciences and applied to an oceanographic example here. Using domain specific theory, manifold representation of the data, clustering and sophisticated ensembling, NEMI is able to highlight particularly interesting areas within the data. In the paper, I stresses the underlying philosophy and appreciation of methods to facilitate understanding of data mining; a tool to gain new knowledge.

Key points:

1: Few tools for data mining within the Earth Sciences use recent advances in methodologies despite data available becoming unwieldly

2: The method Native Emergent Manifold Interrogation (NEMI) is presented. NEMI scales and performs well on very complex and nonlinear data

3: I stresses the underlying philosophy and appreciation of methods to facilitate understanding of data mining; a tool to gain new knowledge

1 INTRODUCTION AND PROBLEM STATEMENT

In this manuscript I introduce a generic methodology to determine areas of interest in a dataset that can have arbitrarily complex and nonlinear covariance structures. For simplicity, the method is given a name: Native Emergent Manifold Interrogation (NEMI). Nemi was developed to address the need to identify patterns and perform 'data mining' in the increasingly large, highly complex and complicated data that is becoming common within the Earth sciences. Due to the challenges posed by modern data, traditional methods of analysis are often inadequate, meaning that they fail to converge or offer little insight. NEMI blends dynamical systems theory with clustering, but importantly invites room at key areas for domain specific input 'native' to the research problem NEMI is applied to. With NEMI, I address the issue of mismatching 'data science' methods and data, where the practitioners of Earth science or more computational sciences often suffer under the difficulty of interdisciplinary communication. NEMI is a generalisation of the methodology in Sonnewald et al. (2020) that targeted plankton ecosystems, in that is is designed to scale to larger datasets. Scaling is one of the true bottlenecks in data mining for scientific applications. NEMI is generalised to work with any data, where the particular example application used here is geospatial data. I have used an explicitly hierarchical approach, making NEMI less parametric (fewer parameters to tune and less danger of noise interference) and intuitively useful both for global (for example the whole Earth in the present example) or more local applications (for example a basin or more regional assessment). Another novelty in NEMI is the lack of a fixed field-specific benchmark criteria (used in Sonnewald et al. (2020)), where I have generalised so a field agnostic option is available. Lastly, NEMI invites the use of a range of uncertainty quantification options in the final cluster evaluation. The intended readership of this manuscript are interested practitioners from the Earth sciences, meaning scientists interested in applying NEMI, with an interest in understanding the underlying philosophy and rationale beneath the architecture of the pipeline. I have attempted to describe concepts in detail and refer the interested reader to further materials. Here there are two main actors; the data and the methods of analysis. Oceanographic examples are used and an ocean numerical model dataset used as an example. Note that the present manuscript focuses on NEMI, and I do not include an general overview of data-mining within the Earth sciences (see Götz et al. (2015)), or provide a general overview of machine learning within the Earth sciences (see Fleming et al. (2021); Sonnewald et al. (2021); Beucler et al. (2021)).

The paper is structured as follows: to give NEMI context, I initially move through explaining the problems related to exploring data using machine learning methodologies, these being the data (section 2) and the methodologies (section 2.2) in very general terms. I move through a synthetic example of a simpler method to illustrate how, and why, this fails on more complex data (section 2.2.2). Then, in section 3, I move through the manifold-based projection for data cleaning, visualizing, and strengthening the associations between different components of the data. In section 2.2 I explain the actual clustering, and how this is chosen based on observations of the manifold. Sections 3 and 2.2 are part a) of NEMI as shown in Fig. 1. Section 4 illustrates the important step regarding how to treat and sort the resultant clusters for enhanced utility. Then, in section 5, I return to the issue of stochasticity, and demonstrate simple and more advanced methods to utilize this aspect of NEMI. Sections 4 and 5 are part b) of NEMI as shown in Fig. 1. Finally, section 6 provides an outlook on potential application and implications. NEMI is a method that can not only be applied to complex data, but is also flexible and verifiable. I refer to NEMI as the full workflow, but separate parts can be used and adapted as appropriate for the practitioner. The following provided code and examples uses the python programming language and key parameters are highlihgted. Note that parameters not discussed could be significant depending on the application, and the documentation should be read. The manuscript intends to give a thorough explanation of the reasoning behind NEMI, and aims to empower practitioners with the rationale behind different method choices so it can be applied to different data. I do not intend for the manuscript to stand entirely on its own as many methods NEMI draws from span a wide array of fields and, only brief explanations are within scope and should be seen as starting points for further reading. The code for NEMI is available on GitHub and also as a PyPi package: https://github.com/maikejulie/NEMI

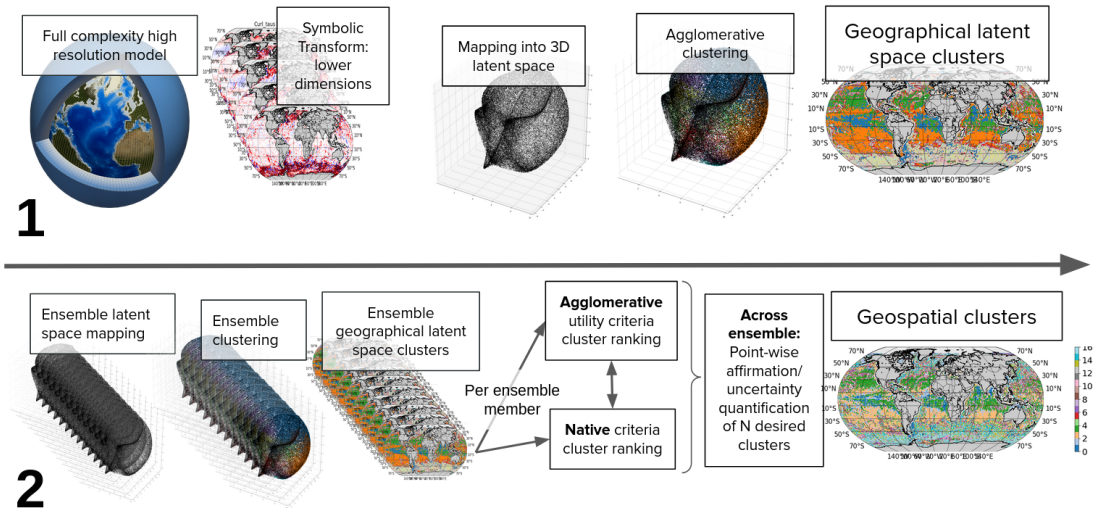

Figure 1. Sketch of workflow in NEMI. Sketch of NEMI workflow. Part 1 (top rwo) illustrates moving from the data in its rew form, through initial symbolic renditioning, manifold transformation and clustering. Part 2 (bottom row) shows the ensembling, agglomerative utility ranking and native (field specific) utility ranking within each ensemble member. Finally, the cluster for each location is determined looking across the ensemble. (Top left image of model adapted from encyclopedie-environnement.org).

2 DATA

Data is at the core of most discoveries within the Earth sciences and can come from numerical models or in situ or remote observations. However, what data to choose can be the most crucial step of any scientific endeavor.

2.1 The ocean data

In the illustration of a NEMI application in the present manuscript, I take data from an ocean model (MOM6, Griffies et al., 2023). Working on the full set of fields would have many parameters. The ocean model is discretized in latitude and longitude, as well as in depth, meaning that the model equations are solved on a grid that subdivides the ocean area and depth. The area covered within each grid point varies widely. Each data point approached naively would consist of one point in the depth, latitude, and longitude, where the model has 75 depth levels. We are interested in how the ocean is moving (as is described by the model equations in terms of momentum), and for each data point this amounts to 39 different fields for each depth level, where each field is one term in the equations that the model is solving, along with three additional ones at the sea floor. The equation terms can be thought of as our 'features' or 'dimensions'. As such, each location in latitude and longitude amounts to a vector of length 2989 entries or dimensions (39*75+3). See Khatri et al., 2023 for further details on the momentum budget closure in MOM6.

Working in a space consisting of 2928 dimensions is difficult, so we initially make the data more manageable using oceanographic theory. This can be thought of as simplifying the latent space within the data and is highly field-specific. This was described in detail in Sonnewald et al. (2019). For this simplification I use the barotropic vorticity (BV) equation terms as the data for NEMI, reducing the 2928 dimensions to five (Fig. 2). Although generally applicable to any data, NEMI was developed around the BV data as output from a fully realistic numerical ocean model. As such, the data that is used here is a parameter x that is a vector field defined at every grid cell (lon, lat) on the discretized MOM6 ocean sphere, with each element x_i representing a five-dimensional vector on the horizontal grid of the model. The index i uniquely identifies a grid point on the sphere, with (lon,lat) = (ϕ_i, θ_i). The features (dimensions) of each vector x_i correspond to the five terms in the BV budget. For the interested reader a description of the BV equation and how it relates to the numerical model follows. Skip ahead for further discussion of NEMI.

Early works (Sverdrup, 1947; Munk, 1950; Stommel, 1948) recast the intractably complicated full equations to describe how meridional ocean flows develop by taking the curl of the depth-integrated

momentum equations, and arriving at the barotropic vorticity (BV) equation. The steady BV balance under incompressibility is expressed as:

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

where $\beta = \partial f / \partial y$ is the northward derivative of the Coriolis parameter (f), $V = \int \rho v \mathrm{d}z$ is the depth-integrated northward mass transport from density ρ and meridional velocity v, ∇ is the horizontal gradient operator, p_b is the pressure at the bottom, and $H = h + \eta$ is the bottom depth. H is the water column thickness, an h is the distance from the resting ocean surface to the bottom topography and η the sea surface height anomaly. The stress produced by wind and bottom friction (external) is denoted by τ, and $\mathbf{A}$ and $\mathbf{B}$ are the depth integrals of the nonlinear and the horizontal viscous terms, respectively (Hughes and de Cuevas, 2001).

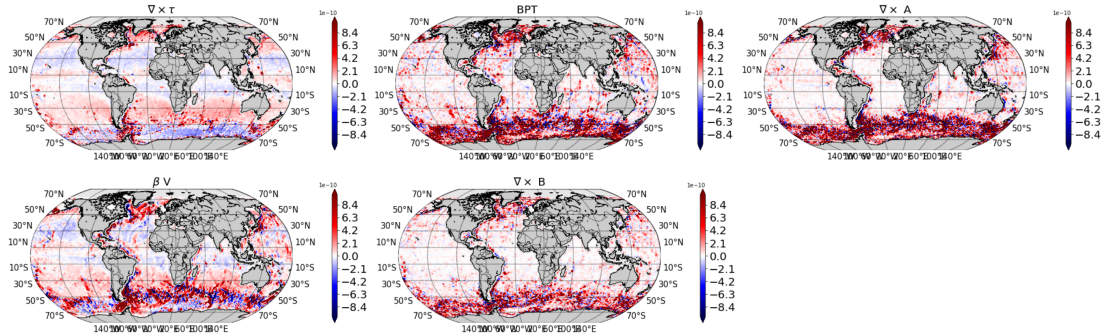

Figure 2. The terms of the barotropic vorticity equation. Each term is in ms^{-1}. Note how certain areas have clear large spatial patterns, while others can be highly variable. Top from left: $\nabla \times \tau$, $\nabla \times (p_b \nabla H)$ and $\nabla \times \mathbf{A}$. Bottom from left: $-\beta V$ and $\nabla \times \mathbf{B}$. See Fig. 3 for close-ups illustrating the complexity of the data further.

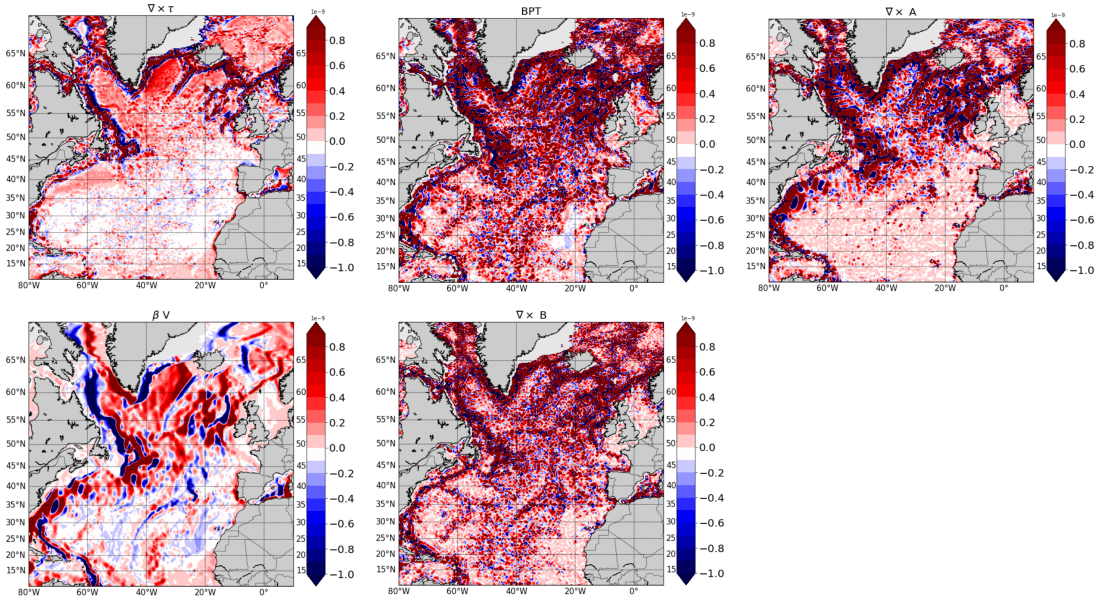

Figure 3. The terms of the barotropic vorticity equation, North Atlantic section. Each term is in ms^{-1}. Terms labeled as in Fig. 2.

2.1.1 Common problems with realistic data

In the example featuring the BV data from the numerical ocean model MOM6, NEMI is applied to a very complex problem. In general terms, data can suffer from issues such as: 1) noise, which is meaningless data that mask components of interest and sources can include instrument error or numerical artifacts; 2) sparseness, where only part of the desired data is available and examples include the wealth of data available at the ocean surface, but difficulty in acquiring subsurface data; 3) unbalance, which refers to data that has a wide range and only a small proportion of information of interest, for example a global dataset where one wishes to detect episodic ocean convection. The extent to which data is afflicted by these issues is often unknown, and checking the nature of the data extensively before starting analysis is always advisable. For the BV data, issues 1 and 3 are of note. The raw data, here only a smoother with a Gaussian kernel with a standard deviation of 1, is presented in Fig. 4 as a 'pairplot' (terminology from the 'seaborn' python library). The pairplot shows each dimension (here each term in the BV equation) as a scatterplot of each other term, with the associated probability distribution function of the dimension as a barplot. Such a pairplot is a nice way of initially assessing what issues are present within the data. In Fig. 4, the points are unreasonably centered around zero, and it suffers from outliers.

For NEMI, as is generally advisable, the data must by appropriately cleaned and pre-processed. Standardizing and normalizing are standard; for example, one can scale as $z = (x-u)/s$, where z is the scaled data, x is the original data, u is the mean and s is the standard deviation. This is done separately for each dimension, or equation term. Applied to the BV data, we arrive at the pairplot shown in Fig. 5 that reveals different structures. Note that the individual distributions only give a vague representation of data density. Many other methods for data-scaling exist that are suited for e.g., log distributed data. Experimenting with the initial scaling can be highly beneficial. The rationale behind scaling and normalizing is that the covariance between variables is much more interesting than their individual magnitudes. For example, consider the global data of ocean temperature and fish stock abundance, where the magnitude of variability in temperature is small compared to the magnitude of variability in fish stock abundance. Without scaling, the temperature variable would appear meaningless for fish stock abundance, even though we expect a difference between Arctic and tropical regions. After getting to know the data through initial inspection and scaling, we are ready to consider methodologies for further exploration.

2.2 Methods for data mining from unsupervised learning

Novel methods from 'data science' are increasingly being used to great advantage. In Sonnewald et al. (2021) a review of current progress and a brief introduction of methods can be found focused on physical oceanography. However, matching methods to data and robustly verifying their results requires knowledge both of the algorithm and the application. A computer scientist may believe she has arrived at a significant and interesting answer, but this may not be useful to an earth scientist if, for example, the uncertainty related to the spatial position of an identified region is too great. Along with a method's power must also come an appropriate level of skepticism and emphasis on validation, statistical or otherwise. NEMI is an answer to the issue of not having satisfying metrics to use to determine statistical significance of clustering results. Clustering is the task of dividing data into sub-groups so that data points within each group are similar and dissimilar to the data points in other groups. Clustering is largely regarded to be an 'unsupervised' machine learning methodology, meaning that the data is given to the method without explicit 'labels'. Clustering can be seen as the act of determining labels that can then be interpreted and offer insight to the practitioner.

There is a large and growing number of clustering algorithms available. It is beyond the scope of this article to give an overview of these. In general terms, common to all clustering methodologies is that the act of seeking to determine sub-groups within the data means that the differences between the overall data and the sub-sets that the clustering method has chosen should be determined. Put differently, every methodology should evaluate the differences between the overall data and clustered sub-sets. Note that such partitioning of the data's covariance space is also the backbone of, for example, neural networks. When applying a clustering algorithm, the resultant 'model' is the algorithm and the chosen values for any parameters or similar, referred to as clustering model hereafter. As such, it is critical to quantify how well the clustering model is able to represent the data. This is also seen in simple regression. Within clustering and regression, we search for an underlying and general model, or formula, to describe the data. To illustrate, the left column in 6 shows an underfitted model, attempting to partition the data with a straight line. This misclassifies large amounts of data. The rightmost column illustrates a model fit

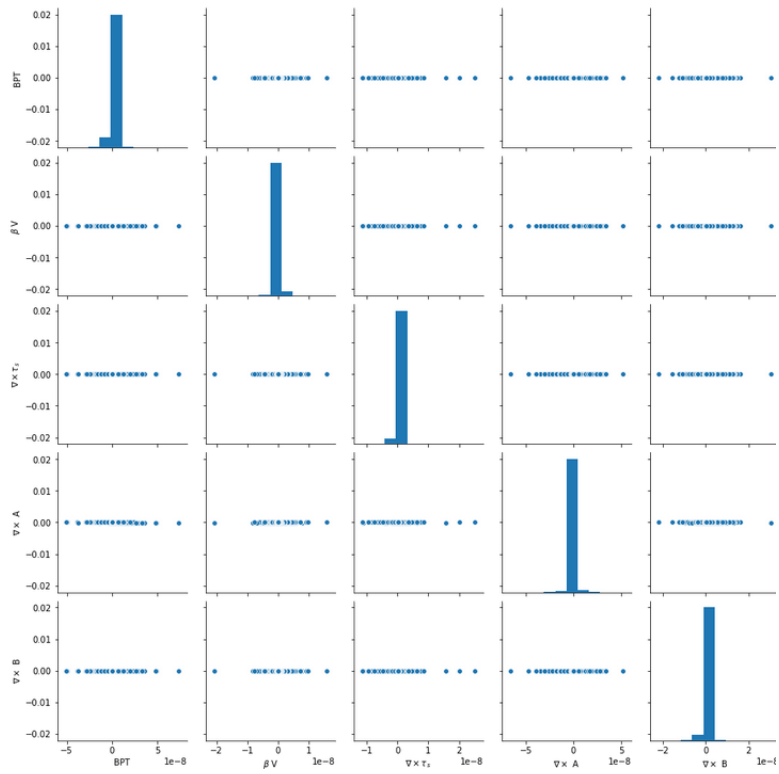

Figure 4. A basic look at the data. Each variable in the BV data is plotted against the other. Note that the data is not scaled, and the odd lack of structure indicates that the data is heavily skewed towards smaller numbers. See Fig. 5 for the scaled example.

where every point is accounted for, which also fails to reasonably approximate the underlying model. The middle column illustrates a general fit that represents a model that closely approximates the underlying model from which the data was drawn.

2.2.1 Validation

To validate a clustering application, showing that we have successfully discovered a reasonable representation of the underlying model, there are two main techniques: 1) external, and 2) internal validation. External validation requires a subset of the data to have known labels to compare to. Internal validation revolves around cohesion within a cluster and the degree of separation between different clusters. If the cohesion within a cluster is bigger than the degree of separation between clusters, then the clustering method is successful. However, recall the problem of overfitting. Many methods for verification of model skill exist, including the Silhouette coefficient, the Calisnki-Harabasz coefficient, the Dunn index, the Xie-Beni score, the Hartigan index, and the use of information criteria. It is beyond the scope of this article to go through all the above, but the example below will briefly introduce information criteria.

2.2.2 Practical example: k-means on idealised and BV data

A very popular method for clustering is called k-means (MacQueen, 1965). It is fast and conceptually simple, making it an excellent first choice for data exploration. 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. To initialize, the k-means algorithm makes a stochastic guess. This means that points are initially scattered across the data, and the algorithm iterates until a "maximum" in found. This maximum is determined by minimizing the objective function J:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||\mathbf{x}_i^j - \mathbf{c}_j||^2,$$

where k is the number of clusters, n is the number of data points, the vector $\mathbf{x}_i$ correspond to the five

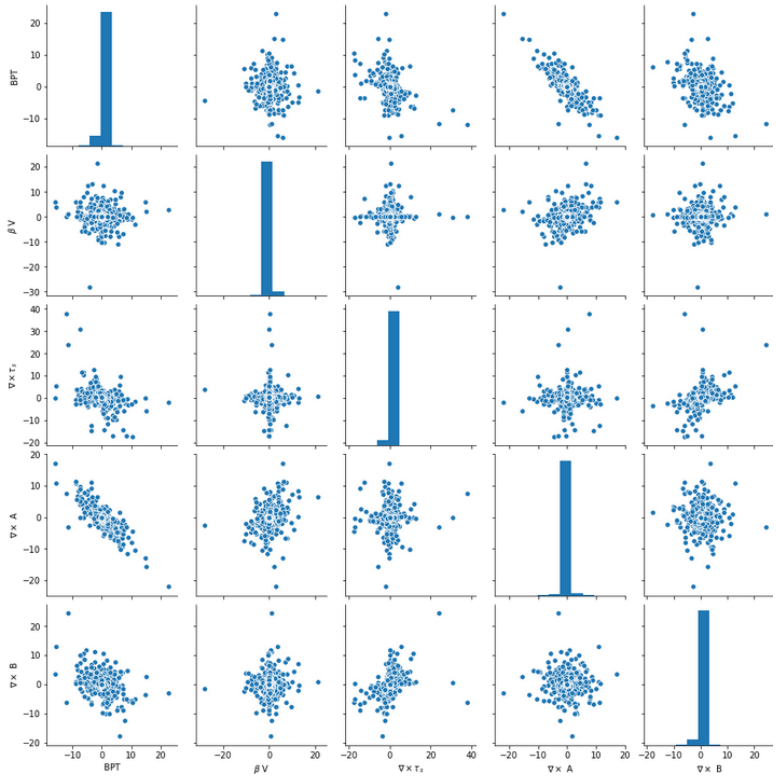

Figure 5. The scaled BV data. Each variable in the BV data is plotted against the other. Note that compared to Fig. 4 much more structure is visible.

terms in the BV budget, and $\mathbf{c}_j$ is the estimated location of cluster j. The number of k clusters is a free parameter that is chosen before the algorithm is applied. Initially, the cluster centers have random values scattered throughout the parameter space. Each cluster $j = 1, \ldots, k$ is represented by the five-dimensional characterizing vector $\mathbf{c}_j$, and the k-means classification attributes each vector $\mathbf{x}_i$ to a unique cluster c_j, so $\mathbf{x}_i = \mathbf{x}_i^j$. The distance between a data point is given by $\mathbf{x}_i^j$ and the cluster center $\mathbf{c}_j$ is determined as: $||\mathbf{x}_i^j - \mathbf{c}_j||^2$. In this way, each data point in $\mathbf{x}$ is associated with the closest k-cluster. Then, the position of $\mathbf{c}_j$ is calculated again, and the association is reassessed until the solution converges.

Note that k-means uses only one parameter (the number of clusters) and an initial stochastic guess for the cluster centers. Effectively, k-means clustering minimizes within-cluster variances (squared Euclidean distances), which also entails that k-means would work *perfectly* if the data were separated into tidy clumps with Gaussian distributions (round). Unfortunately, very few data have this type of covariance space and suffer from interconnected and decidedly non-Gaussian (and nonlinear) statistics. Put differently, the strength but also the weakness of this clustering method is that it works by partitioning the data into Voronoi cells. Effectively, the algorithm can only draw straight lines to partition the data and cannot isolate any more complex covariance structures.

In contrast to the BV example, Fig. 7 shows an idealized scenario that illustrates how k-means can be successfully used. The top left panel shows a dataset of tightly clustered points that are well-separated from neighbouring clusters and each has a Gaussian distribution (round). The top middle panel illustrates how k-means is successfully applied to discover this correct underlying structure in the data (here called 'true'). The top right panel shows the *same* data but with noise added. Using the labels discovered from the 'true' underlying model in the top middle panel, the bottom left panel shows where the data *should* be classified. The middle lower panel contain the classification results. The colors are arbitrary, but note that while there is some misclassification, the performance of the model determined using k-means is reasonable.

Each run of the k-means algorithm on the BV data results in one statistical 'model'. As such, we want to assess how well different models fit the data, where we can vary the number of clusters (k) and

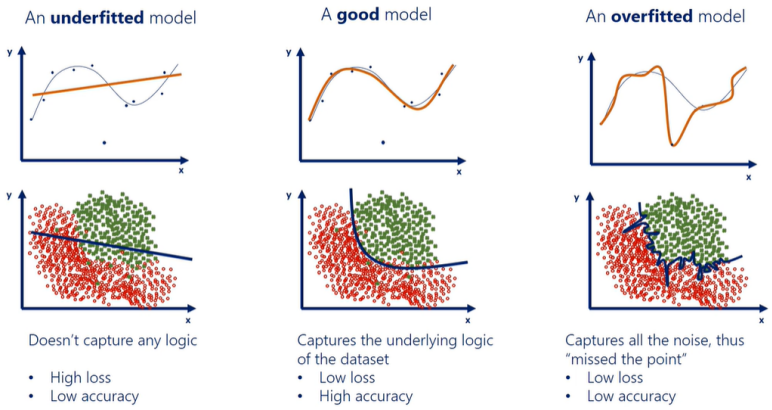

Figure 6. Figure illustrating a model fitting exercise. Adapted from 365datascience.com [1].

different model initializations, which give different results due to the initial stochastic guess (note that some implementations use a fixed seed which masks this potential source of model error). To assess the fit of the model, we apply validation methods. Here, the 'fit' of the model is assessed using information criteria (IC). The IC can tell us how 'complex' our k-means model should be, where increasing the number of k is equivalent to increasing the complexity. The IC illustrates if we are capturing **more information** if we add another k, or if the maximum is reached and the model is complex enough. Recalling the example in Fig. 6 where accounting for every data-point is not desirable, and this statistical model of the data is too complex, meaning it will not generalise well.

Different methods exist for estimating the IC, and here I will discuss the Akaike IC and Bayesian IC (AIC and BIC respectively). The AIC and BIC have been very well studied, and are therefore preferred. The maximum likelihood function is the basis of both and is the primary tool for estimating the parameters of an assumed probability distribution given data (here the data we cluster on). The likelihood ($\mathscr{L}$) is defined as:

$$\mathscr{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x),$$

Here, X is a discrete random variable with probability mass function p depending on a parameter θ. If thought of as a function of θ, it is the likelihood function, given the outcome x of the random variable X. Suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model (for example the number of cluster guesses from k-means). Then, $\hat{L}$ is the maximum value of the likelihood function for the model. Then the AIC value is estimated as follows:

$$\text{AIC} = 2k - 2\ln(\hat{L}).$$

With a set of different candidate models (for example, comparing models determined using different numbers of k clusters), the AIC with the lowest number will be the one that fits the data best. The goodness of fit is assessed by the likelihood function. To discourage overfitting, the penalty term ($2k$) increases as the complexity (number of k clusters) increases. As such, the AIC will in general asymptote, and a good model is determined when this happens.

The BIC also uses the likelihood function to determine the goodness of fit, but uses a different penalisation to determine if the model is overfitted:

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L}),$$

where n is the sample size. As discussed by Yang (2005); Harvey (1982), the AIC can overestimate the order, where the BIC penalisation term discourages this more strongly. See figures in section 2.2 for an example of how a model can fail to find an underlying model. In short, the AIC should asymptote, while the BIC should start increasing. A number of k somewhere between these two (if they both occur) could offer a good fit.

It follows that the AIC and BIC are inappropriate if the number of k is unmanageably large, or is close to the number of data points when we have no reason to suspect it should. The relative simplicity of

the AIC and BIC compared to many other model validation methods demonstrates the difficult nature of assessing if a 'good' approximation of the underlying model has been found, and stresses the importance of applying common sense, additional checks, and caution. Note the AIC and BIC are useful in many applications of model selection, for example auto-regressive model estimation (Sonnewald et al., 2018) as is commonly used without validation of the chosen order. Use of a statistical model without assessment of how well the model approximates the data can be highly unfortunate, including that a model that is needlessly complex is chosen or vise versa as discussed in Sonnewald et al. (2018).

Returning to the idealised example in Fig. 7, the bottom right panel illustrates the use of the AIC and BIC, where the 'correct' number of k is 5. In Fig. 7 the AIC appears to have asymptoted, and the BIC to reach its lowest point before going upwards again. As such, 5 clusters are correctly identified as the optimal number.

In Sonnewald et al. (2019), using BV data on a 1° horizontal resolution ocean model (See Sonnewald et al. (2019) for details on the model), k-means was used to good effect. This was astonishing, as the success of k-means suggested that large proportions of the ocean had an underlying linear distribution. Sonnewald et al. (2019) both illustrated that there were dominant partitioning within the BV data at this relatively low resolution, but also that the data was relatively normally distributed. This was unexpected in an oceanographic context using data from a realistic model. The partitioning of the data has led to scientific insight as the ML effectively performs an empirical leading order analysis that can subsequently be explored.

Running an AIC/BIC check on BV data from MOM6 at 1/4° (Fig. 9) illustrates that k-means is an inappropriate method for exploring this data. This is evident in that the AIC has not asymptoted after adding even 350 k, and while the BIC has started turning upwards the standard deviation (shown in the blue shading) is fairly large. To illustrate spatially on the ocean, Fig. 8 shows the spatial patterns associated with k set to 50 and 200. Neither are helpful, and section 2.2 illustrates that k-means is actually doing with its inability to work with nonlinear data.

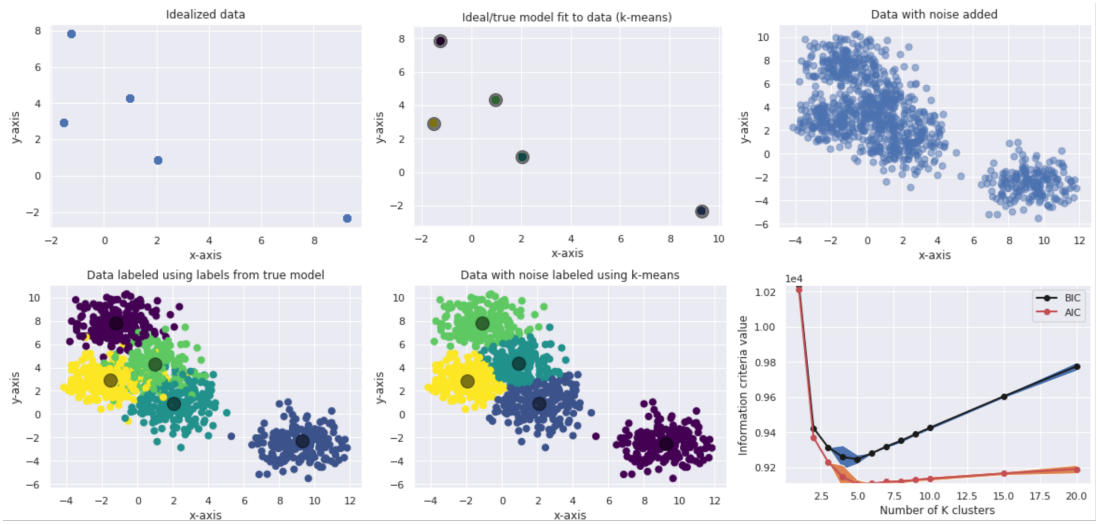

Figure 7. An illustration of concepts on idealised data.

3 MANIFOLD APPROXIMATION OF THE UNDERLYING COVARIANCE STRUCTURE

The NEMI methodology takes the approach that validation is of the utmost importance. As discussed above, validation can take multiple forms. The data-mining challenge that NEMI addresses uses a symbolic methodology to characterize the 'latent space'. The latent space is the covariance structures in the data hidden to our human perception. This can be imagined as how variables relate and change according to one another. The 'symbolic' methodology refers to reducing the size of the latent space from the original dimensions (over 270 for the closed momentum budget of the ocean model) down to a few

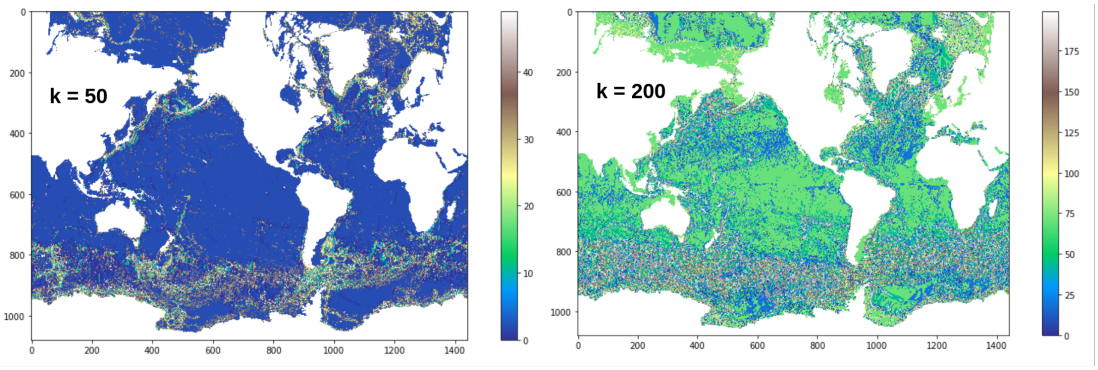

Figure 8. An illustration of running k-means on the BV data. To the left a k of 50 is chosen. To the right a k of 200 is used.

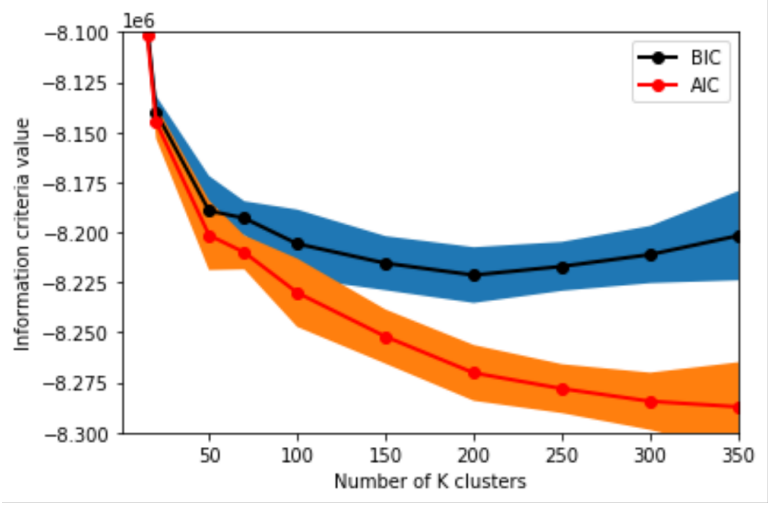

Figure 9. An illustration of the AIC and BIC criteria run on the BV data. Note that the AIC fails to converge and the BIC stays fairly flat. The AIC and BIC indicate that the k-means algorithm is not converging.

(i.e., five) using oceanographic theory. However, five is still too many for human perception. As such, we further characterize the latent space using a manifold methodology.

A mathematical manifold is a construct from topology: any local point resembles the Euclidean space near each point. Effectively, the 'distances' between different datapoints are used to determine relations. This has the convenient property, which makes it useful in NEMI, that the space is homeomorphic. This homeomorphism means that one shape can be transformed to another, without violating the relationships between the datapoints. One common example is that a doughnut (torus) can be transformed into a coffee mug, as both have one hole. It is beyond the scope of this article to give a thorough introduction to topology, but the key utility is that a confusing ball of covariances can efficiently be untangled **without losing the nonlinear structures**. For a visual example, imagine a scarf tangled on a table. The scarf has various patterns that may look oddly disjointed and tangled together. If you spread out the scarf, the complicated 3D structure becomes a smooth and fully visible (approximately) two-dimensional object. Any patterns on the scarf can be seen. In NEMI, the threads making up the scarf and its pattern are the barotropic vorticity equation terms.

Manifold methodologies have two further convenient properties: they can be used to reduce the dimensionality for visualization and 'strengthen' associations between different areas of the data, allowing patterns to emerge more clearly. NEMI employs the manifold methodology UMAP (Uniform Manifold Approximation and Projection McInnes et al. (2018)) as a processing step with considerable advantages. First, the UMAP methodology projects the data into three dimensions, meaning that the data can be

visualized. This allows an additional external validation step that will be discussed later. Second, the UMAP methodology works by assessing the connectedness of the data as described above. Third, through the UMAP application, the noise that posed an issue in the k-means application is lessened.

So what is a manifold in relation to actual data? We start with simple combinatorial building blocks (called simplices) of the distances between the data points. One data point is a 0 simplex, two connected points is a 1 simplex, three connected points is a 2 simplex (a triangle), a 3 simplex has four connected points (a pyramid), and we can continue upwards adding dimensions. We can construct different simplexes and combine these together, and in practice the simplexes do not need to have very high order to cover their local space. If this sounds similar to a k nearest neighbour graph (distinct from k-means) note that there the choice of the radii can have immediate detrimental impact on a graphs ability to approximate the underlying space, which is amplified if a dimensionality reduction is attempted (if the space were uniformly sampled it would work). The problem of having non-ideal data remains, and we can only assume that we are not uniformly distributed. Using Riemannian geometry the non-uniformness can be leveraged as fortuitous. In UMAP we **assume** that we have uniformly distributed data, and then use the actual distances between the data-points to create a map of the underlying manifold. Effectively, to map out the manifold we choose a unit 'ball' about a point stretches to the k-th nearest neighbor of the point, where k is the sample size we are using to approximate the local sense of distance (I use 'k' to conform with the overall machine learning literature, but note that this is distinct from the k in k-means). In UMAP, each point is given its own unique distance function. This lets us choose a number of 'neighbouring' data-points to use, rather than needing to determine the distance as k nearest neighbours would have required.

We now add to the concept of the manifold that it is locally connected, meaning that it describes one space, rather than a set of disconnected spaces. However, in a simplified sense, because we looked at the neighbouring points to assess the distances, two neighbouring points may individually have different values describing the same distance. As such, a useful mental construct with which to envision this set of UMAP is to think of it as a weighted graph, where the weights describe the distances. If there are conflicting weights associated with the simplices we will interpret the weights as the probability of the simplex existing.

Embedding the manifold into a lower-dimensional space can now happen based on the notion that we have the information about the manifold approximated by the data points, and we wish to conserve the associated probabilities between the data points in the lower dimensional space. In comparing the original topological structure of the manifold with a lower dimensional candidate. Both would share the same 0 simplices, and we can imagine that we are comparing the two vectors of probabilities indexed by the 1-simplices. For this we use the cross-entropy. In Information theory, the cross-entropy is a concept describing if two probability distributions are drawn from the same set.

To estimate the cross-entropy, say the set of all 1-simplices is E, and we have arrived at weight functions so the weight of the 1-simplex e is $w_h(e)$ in the *high* dimensional case. Now $w_l(e)$ is the weight of e in the *low* dimensional case, and the cross entropy will be:

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right).$$

Now, we minimize the cross-entropy to arrive at our low dimensional embedding of the migh dimensional manifold. Here, the first term $w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right)$ can be thought of as a force that attracts the points whenever there is a large weight associated in the high dimensional case. If $w_l(e)$ is as large as possible, the term will be minimized. This occurrs when the distance between the points is as small as possible, and effectively when the UMAP algorithm is focusing on the very local structure. In contrast, a repulsive force is found in the $(1 - w_h(e)) \log\left(\frac{1-w_h(e)}{1-w_l(e)}\right)$ term between the ends of e when $w_h(e)$ is small.

Key concepts to understand the limitation of such 'manifold' based methods is that we are assuming that our data points populate the manifold of the underlying model well enough. For example, if we think about a landscape with mountains, we may have less data points among the mountains than in the surrounding areas that are also 'smoother'. A manifold representation of the landscape based on this data would only roughly describe the true landscape in the mountaineous regions.

Now, what does a UMAP rendition of the highly complex and **complicated** BV data look like? In Fig. 10 a three dimensional rendition is demonstrated from different angles. The shape can vary depending

on the parameters chosen, as stressed above. In Fig. 10 we can see that there are clear areas that, from all angles, are more dense and some that are more sparsely populated. We will use the visualization for choosing a clustering algorithm below. The sensitivity to parameters (or how 'brittle' the method is) is highly dependent on how the data's complexity. In this example, a large ensemble sweeping through the parameters (described above) was needed to arrive at a reproducible manifold representation. The concept of a reproducible manifold means that one should be able to run the algorithm on the data and recover the same (or sufficiently similar) structure. Here, small differences can have large impact, and they can be difficult to pick up by eye. The importance of small differences is part of the reason why NEMI employs the additional checks and leverages the associated uncertainty. In Fig. 11 three renditions of running the UMAP algorithm on the processed BV data. The plots in panels a-c in Fig. 11 may look very similar to the human eye, but note the differences in the arrays. For example, the first number in the array goes from 7.895877 in the manifold in Fig. 11a, to 7.892489 and 7.875971 in b and c respectively. These differences may appear small, but they are present and can skew results. Determining the acceptable and appropriate level of difference is critical to the success of NEMI.

UMAP is similar to other methods such as t-SNE (t-Distributed Stochastic Neighbor Embedding, van der Maaten and Hinton (2008)). NEMI as presented here uses UMAP, but note that the t-SNE method was used in Sonnewald et al. (2020). Both UMAP and t-SNE have drawbacks, and one should weight carefully if these are appropriate for the data. These include that t-SNE, like UMAP, does not completely preserve density. UMAP, like t-SNE, can also create tears in clusters that should not be there, resulting in a finer clustering than is necessarily present in the data. Overall, such issues are exactly why NEMI was developed with additional validation steps. As such, NEMI uses both external and internal validation.

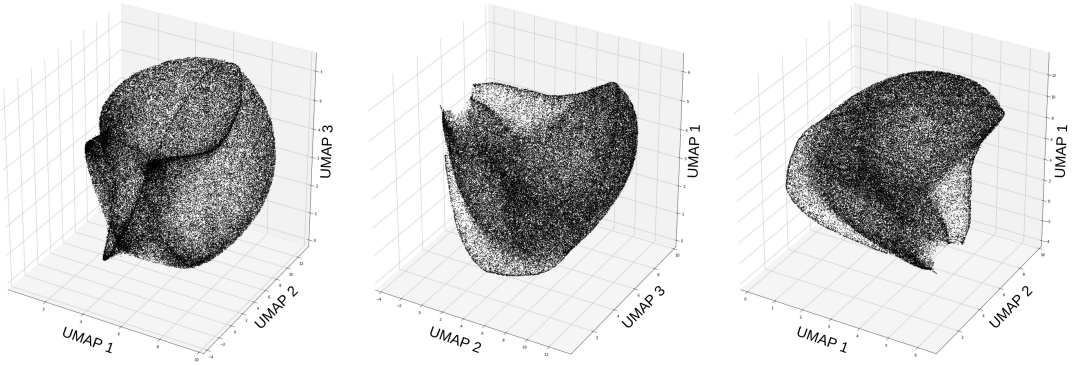

Figure 10. One UMAP manifold from different angles.

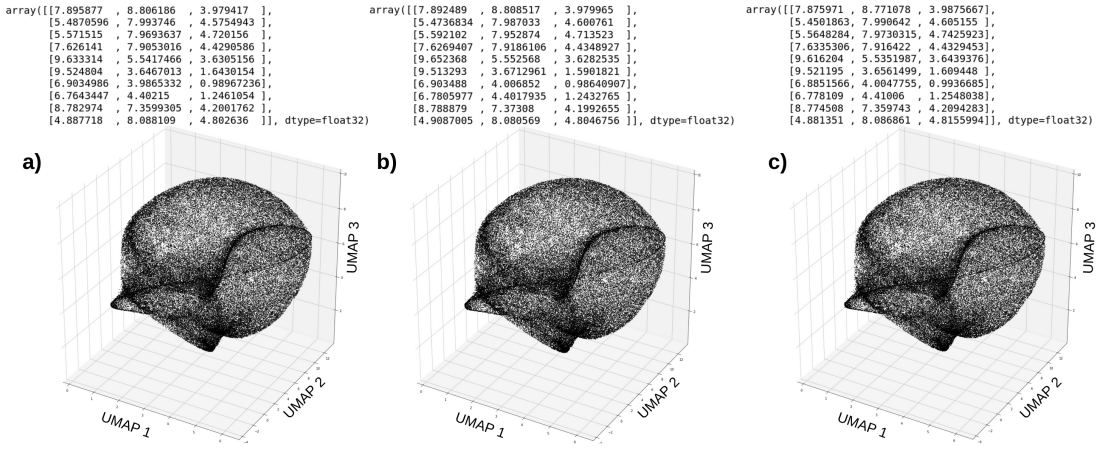

Figure 11. Three different ensemble members, with a part of the associated data. Note that while the manifold renditions look very similar, the data associated highlights the slight differences.

4 CLUSTERING: LEVERAGING MANIFOLD APPROXIMATION

The use of manifold and dimensionality reduction methodologies in NEMI leads to a very convenient three-dimensionsl visualization of the data (Figs. 10 and 11). In terms of clustering (or data-mining) this makes it visually apparent that an algorithm needs to be able to deal with data that is: 1) not well-separated (e.g., one continuous-seeming structure), 2) highly nonlinear, and 3) of varying densities meaning that the points are more likely to be found in certain areas.

There is growing number of different clustering algorithms available to the practitioner. For validation NEMI uses a hierarchical cluster analysis (HCA, the clusters found by the ML method will hereafter be refrered to as 'HCA clusters'), specifically an agglomerative methodology. Here, an agglomerative algorithm initially assumes that each data-point is its own cluster, and pairs of clusters are merged as one moves up the hierarchy. This is a "bottom-up" approach, where a "divisive" approach would be the opposite ("top-down") and assume that the initial step is to have one cluster represent the whole dataset and proceed to divide the data. Note that the agglomerative clustering methodology is not stochastic. The hierarchical element is useful as it means that running the algorithm on the same data will not introduce uncertainty in what clusters are found. In NEMI this refers to the same manifold rendition of the BV data. Using a hierarchical method is intuitively useful both for global (for example the whole Earth in the present example) or more local applications (for example a basin or more regional assessment).

The agglomerative hierarchical clustering methodology is presented as a cartoon in Fig. 12. Here, Fig. 12a shows the data points 1 to 6 in a two dimensional space (here 'UMAP 1' and 'UMAP 2' for simplicity in relating to the present section, although this is strictly a cartoon and UMAP was not applied). The data points have a certain distance to each other within this space. In Fig. 12b, initially each data point is progressively clumped together in relation to the distance between the points in panel a. As such points 4 and 5 and initially grouped, as are 3&2, while 6 and 1 remain isolated. In the next aggregation level, 6 is brought into the 5&4 cluster, becoming 6&5&4. The other points remain disaggregated as the distance between them is still too large (see Fig. 12a)). At the next level, the 3&2 and 6&5&4 clusters are merged into 6&5&4&3&2. Finally, on the next level the data point 1 is brought into the cluster, comprising now of the entire data set.

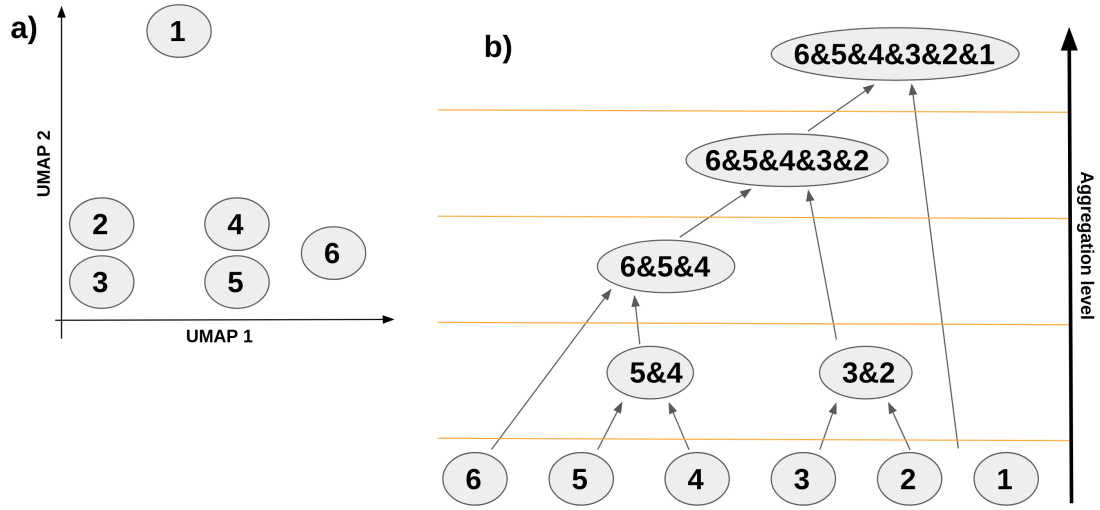

Figure 12. Sketch of the agglomerative clustering functions.

Having chosen an agglomerative methodology, I will highlight two hyperparameters that are of greatest relevance to the practitioner. The choices are that of the distance metric and linkage method. The distance metric is an expression of how the separation of the points is quantified. To illustrate, imagine a room of people, such as an auditorium with a lecturer on a podium and students sitting at a distance. If grouping the people using physical distance, the students would be clustered together because the gap between the students would be smaller than the distance to the lecturer. That is, the gap between the teacher and the closest students would be a defining feature of the data. However, if one used a metric such as how well the students know each other (e.g., how many friends they have in common),

there would likely be clear groupings within the students. As such, the distance metric chosen should be considered carefully. In NEMI, the use of the manifold methodology and a *closed budget* means that we can directly link the distance in UMAP space (seen in Figs. 11 and 10 as the distance between points) to the clustering. The use of budgets implies that a Euclidean methodology is appropriate. A Euclidean metric effectively uses Pythagoras's theorem in Cartesian coordinates. Next, the choice of the linkage method is often dictated by computational capacity. Note that methods scale differently with the size of the dataset. Here, the simplest method (single linkage) scales as $\mathcal{O}(n^3)$ where n is the number of points and should be avoided unless the dataset is very small. The second hyperparameter of interest, the linkage method, groups points. Recall the validation methodologies that look at internal versus external validation. Here, the linkage method can be seen in relation to the internal method. In NEMI we have used the Ward linkage method Ward (1963). Ward's method uses a minimum variance criterion that minimizes the total within-cluster variance. Let d be the distance between points i and j in data vector $\mathbf{x}$. The initial distances in Ward's method are Euclidean distances between points:

$$d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||^2.$$

Note that in NEMI the use of the 'cut' is equivalent to the direct number of clusters that are returned (the HCA clusters). So why not just use these? The reason for this was illustrated in Fig. 12, where in our BV data example the more 'extreme' outliers would be immediately focused on, and the wide swaths of the open ocean that are dynamically highly interesting would not be identified. For example constitute term balances that are opposite. The HCA clusters, approached naively, therefore have limited utility.

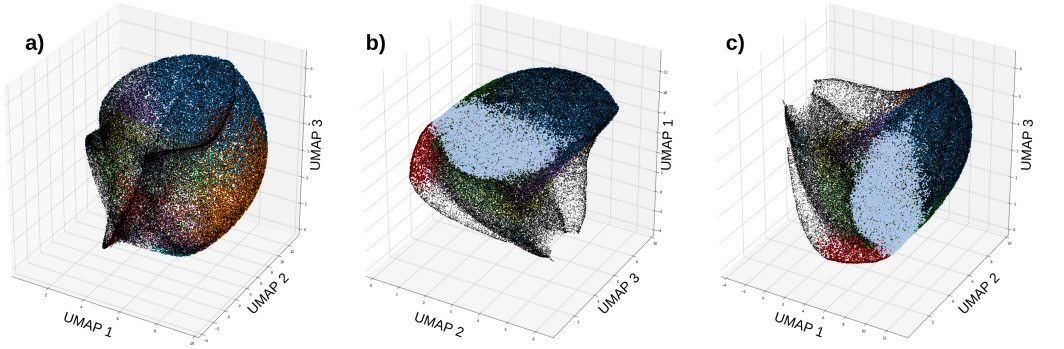

Figure 13. The agglomerative clustering on UMAP with 17 clusters. Panels a, b, and c show the same manifold from different angles. See sub-sampled version of a) in Fig. 14 to highlight shapes that are picked up by NEMI.

In Fig. 4 the application of the hierarchical clustering to a UMAP rendition is illustrated from a few angles (panels a-b are from the same manifold with the same clusters). The colors indicate the 17 different clusters (more detail on this below) and show how the clusters successfully isolate the ridges running along the sides of the data (see Fig. 14 for a sub-sampled version of panel a from Fig. 14 where details are highlighted). Note also that Fig. a displays one arbitrary iteration (i.e., ensemble member (i.e., ensemble member) of UMAP, with clusters determined on another UMAP ensemble member. In Fig. 15, a k-means rendition with 200 k (as looked visually reasonable in section 2.2.2) is displayed on the manifold used in Fig. 4 and 14 (the pale and translucent colours were chosen to enhance the readability due to the large number of colors). Note that the clustering was performed on the BV data before the UMAP algorithm and only subsequently projected onto the UMAP manifold. Each data-point is projected onto three dimensions from a five-dimensional space the locations are retained, the number of data-points remain the same, but the number of dimensions change. In Fig. 15 colors do not delineate the areas that are observed to be grouped together; this is a visual demonstration of how k-means fails to identify key regions. The figures illustrates what k-means does: the algorithm is applied to the manifold rendition in Fig. 16 and is forced to artificially separate the data coarsely using 'straight lines' across the entire data volume. Remembering that the UMAP rendition of the BV data is used to 'simplify' and 'clean' the data, it becomes apparent how difficult it would be to apply k-means to the non-transformed data. In supplement to the information criteria, this additional visual appraisal of the performance of the algorithm

underscores that the k-means algorithm is a poor choice. This method of validation can be applied widely beyond the examples used here.

As with most clustering and machine learning applications, there is no guarantee of finding the optimum solution. There might not even be one. However, if an optimum does exist for the agglomerated clusters, it is guaranteed to be found via single-linkage. Due to computational costs the application of single-linkage application is largely impractical. Other methods, such as the Density-based spatial clustering of applications with noise (DBSCAN Ester et al. (1996)) used in Sonnewald et al. (2020) can be useful, especially if the data is more separated. However, in this example arriving at a robust set of clusters was difficult using DBSCAN. Note that DBSCAN performs **considerably** better, in terms of scaling to larger datasets, so if possible this method is recommended.

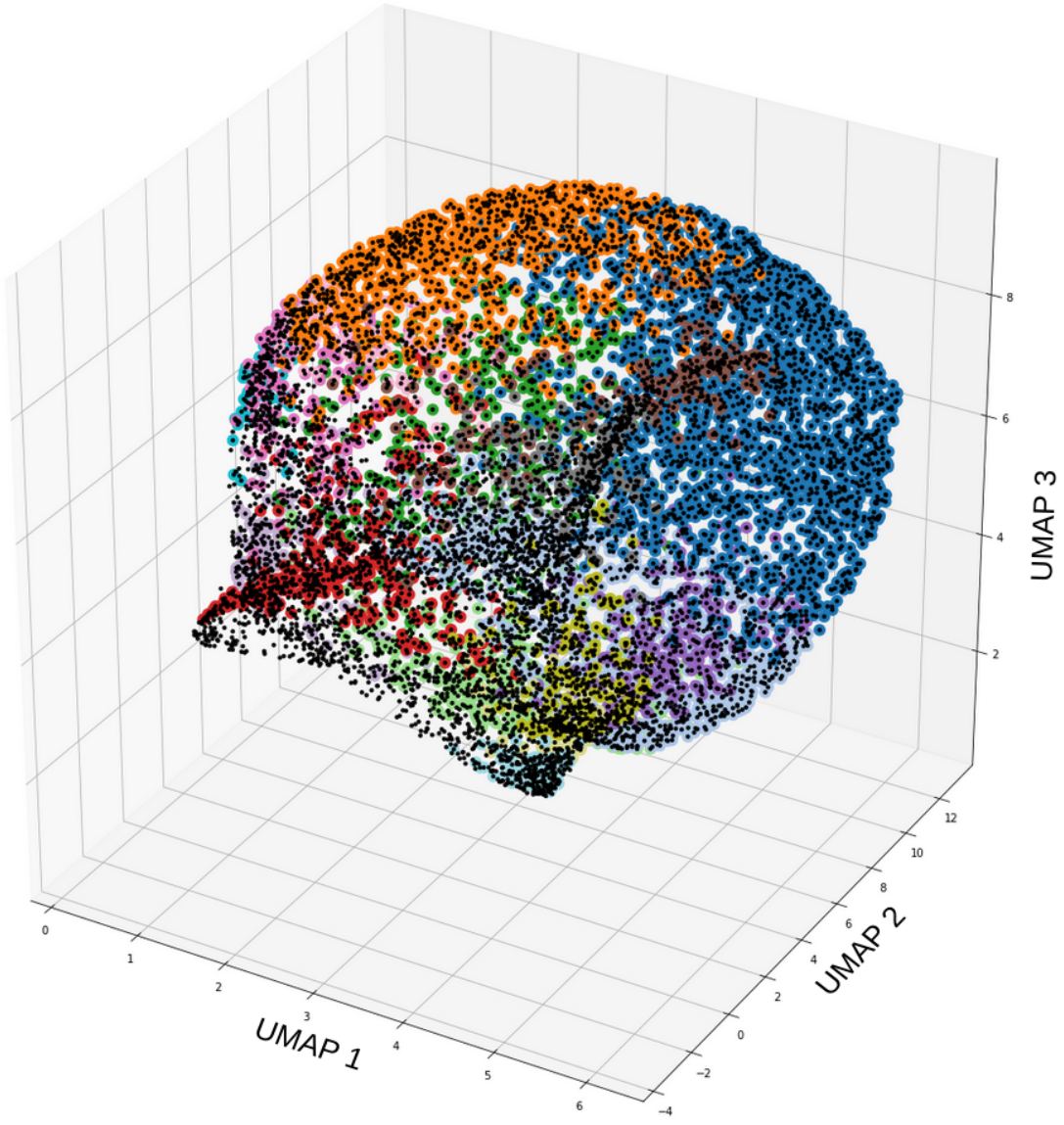

Figure 14. The agglomerative clustering on UMAP with 17 clusters, heavily sub-sampled. Illustration to supplement Fig. 4. Note I use an arbitrary ensemble member for the manifold and a different ensemble member for the clusters.

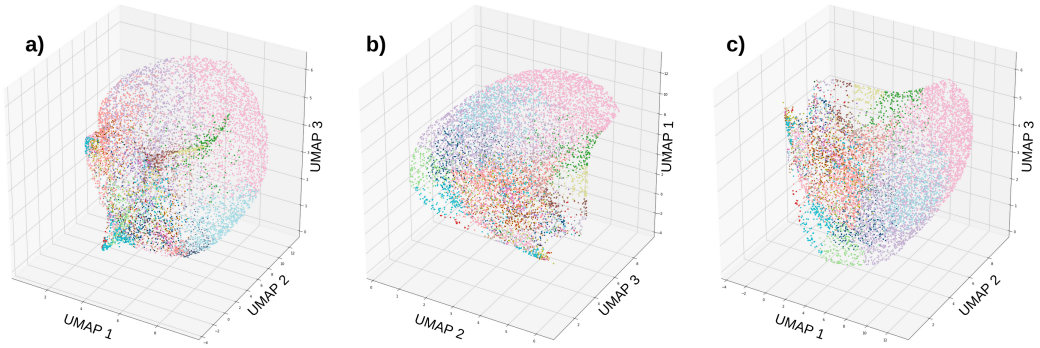

Figure 15. The k-means algorithm with k = 200 result projected onto a UMAP manifold. Panels a, b, and c show the same manifold from different angles. Note that the clusters should be coherent on the manifold if the method is successful. Note there is poor coherence and the clusters are somewhat arbitrarily separating chunks of the space. This confirms earlier suspicions that the k-means algorithm was not succeeding in arriving at a good model representation.

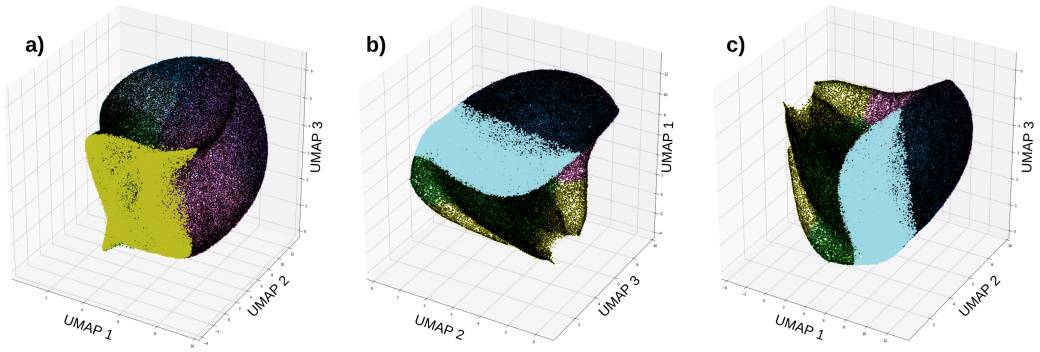

Figure 16. The k-means algorithm applied to a UMAP manifold. Panels a, b, and c show the same manifold from different angles. Here the impact of k-means is illustrated. Note how the manifold is artificially 'chopped' up in ways that clearly do not respect the data.

SORTING FOR DESIRED TRAITS: UTILITY AND VALIDATION

The visual check of the clustering algorithm chosen in NEMI as discussed in section 4 is the first step towards validation. However, the second validation step also builds on the hierarchical aspect of the clustering algorithm. For additional external validation in NEMI, we turn to established oceanographic theory and leverage that the data we are using is the BV budget. While this may appear specific to the BV data, it is generalizable such as in Sonnewald et al. (2020), who used the idea of 'provinces' in ecology and how they compared to established notions.

For validation and utility, let us return to a concept introduced in **?** in relation to cluster validation and assessment. The concept of using what is **useful** in an oceanographic context. Put differently, having a model that is a good fit to the data can be completely useless and misleading, for example, if a key parameter was missing from the data (think of the hydrodynamic paradox where missing boundary layer friction stood in the way of progress for over 100 years). A focus on the scientific problem at hand can be very powerful (in the hydrodynamic paradox this would be working on the equation terms **?**). Here, as in **?**, it is critical that the algorithm can robustly recover and reproduce geographical sub-regions. Namely, if the algorithm does not repeatedly recover the same geographical areas, the identified clusters, however reasonable it may look given statistical checks or other validation, have no utility. Ultimately, a criteria, defined here by the practitioner as finding the same spatial area, is the final objective. From Figs. 4, 14 and 11, it may seem surprising that the same area is not recovered precisely after each iteration of this component of NEMI. However, despite the precision apparent in the Figs. 4, 14 and 11, there is geographical variability. This variability, as discussed in the next section, is intrinsically useful.

The next step in NEMI is to sort the clusters for each UMAP rendition by spatial similarity. For this sorting, we weight by geographical extent is used as a weighting because large areal extents are seen as a relevant feature to favor. As such, the next component of NEMI sorts the clusters from each UMAP and agglomerative clustering iteration and then assesses which clusters are most similar in the geographical region covered (scaled by area covered which varies widely across the model grid) across the ensemble members.

In addition to sorting via coherent spatial cover across the ensemble of UMAP and agglomerative clustering repeats, the agglomerative methodology allows the selection of different aggregation levels, with NEMI having these be the number of HCA clusters. As such, NEMI is designed to be appropriate both for global and regional applications. Specifically, a practitioner in need of a globally representative set of clusters would select a small level of aggregation, while a regional application should choose a higher one.

In combination, the choice of aggregation level, as well as sorting by area size, allows one to select the **number of clusters**, together with the **spatial level** one is wishing to focus on. Note that it is up to the practitioner to determine a reasonable level and effectively number of clusters, as well as acceptable uncertainty/entropy (discussed below). Overall, note that this feature is of specific concern if working in the equatorial region compared to high latitude regions. Mid-latitudes see much less impact, as is expected.

The level of aggregation as well as the number of clusters is illustrated in Fig. 17. Three different ensemble members are shown separately (rows), with an aggregation level of 350 with 6 clusters in the first two columns, and an aggregation level of 350 with 20 clusters in the third and fourth columns. Columns one and three show the global ocean, and columns two and four show the North Atlantic. Note that the three members look very similar, particularly in their global distributions. I omit plotting the 350 clusters as this offers limited insight due the colour scale.

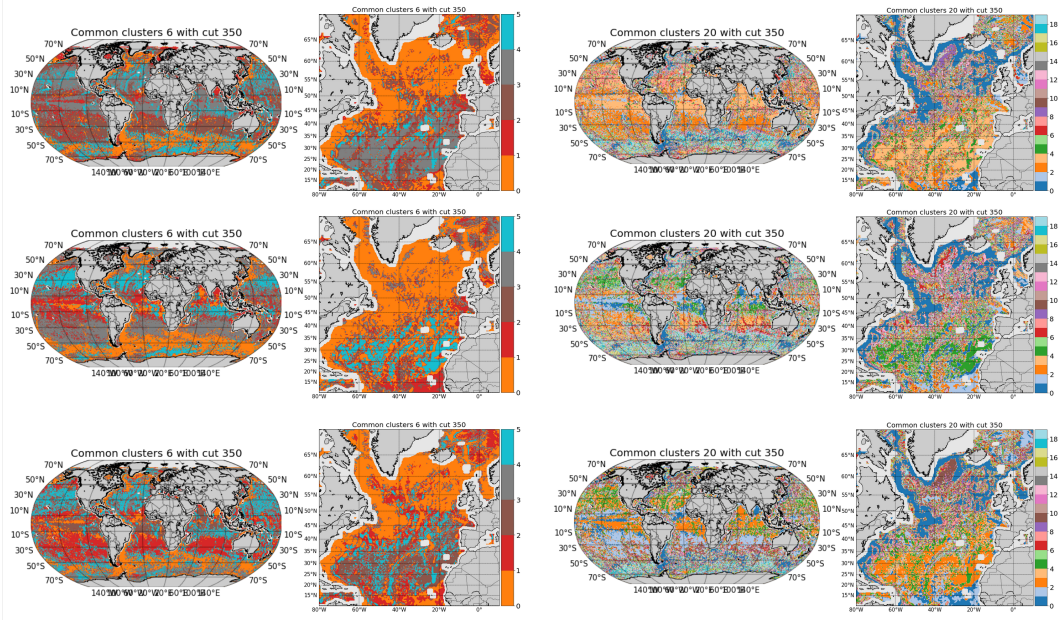

Figure 17. Demonstration of the changes in cluster locations within the ensemble. Three arbitrarily chosen different ensemble members are shown separately (rows), with an aggregation level of 350 with 6 clusters in the first two columns, and an aggregation level of 350 with 20 clusters in the third and fourth columns. Columns one and three show the global ocean, and columns two and four show the North Atlantic. Note plotting the 350 clusters offers limited insight due the colour scale.

Having determined the desired level of aggregation as well as number of clusters, validation via theory, or field-specific intuition should also occur. For example, within the BV budget certain balances are known and expected in certain regions. Specifically, a canonical balance between the windstress curl and advective component (see Sonnewald et al. (2019); Sonnewald and Lguensat (2021) for extensive

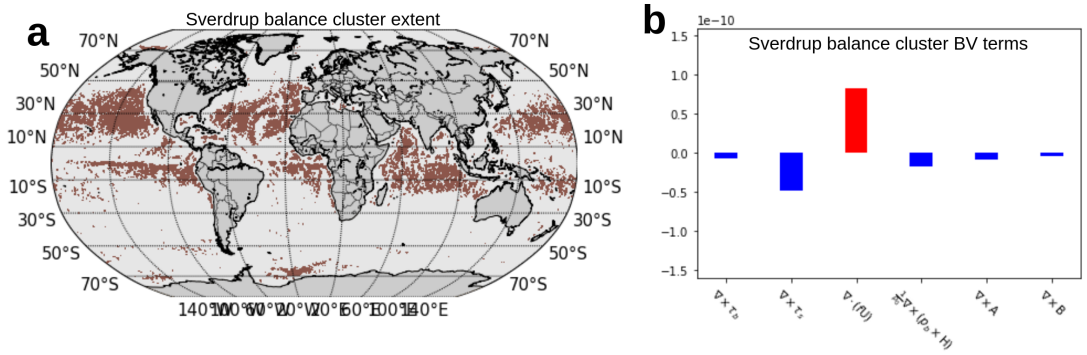

Figure 18. Figure showing canonical and expected balance for validation via expert judgement. The expected balance (meaning that the terms add up to zero) between the windstress curl and advective component can be seen. Fig. 18a shows its geographical extent, and Fig. 18b shows the area averaged terms balances in the locations NEMI highlighted.

description). If this balance between dominant terms is **not seen**, this does not necessarily invalidate the results, but it should mean that the results are treated with increased caution. For example, NEMI here is applied to a realistic coupled model **?**, where intuition and experience strongly suggests the balance between the windstress curl and advective component should emerge in the subtropics in the Northern Hemisphere (Munk, 1950; Sverdrup, 1947). In Fig. 18, just this balance (meaning that the terms add up to zero) between the windstress curl and advective component can be seen, where Fig. 18a shows its geographical extent, and Fig. 18b shows the area averaged terms balances in the locations NEMI highlighted. The exact locations where the balance does not hold (where there are other clusters mixed in, for example, over ridges) can lead to new studies and new scientific insight (For example **?**). As such, NEMI is an avenue towards generating new knowledge with machine learning. However, if this were a BV balance in an idealized channel-model set-up one would not necessarily flag the absence of this balance as suspicious. As a general tool, this step of NEMI requires field specific intuition, where the machine learning and scientist should interact to forge and identify new avenues of discovery.

5 LEVERAGING AND MANAGING NOISE

The issue of noise and stochasticity within data and methods may at first appear to be a challenge that only increases the difficulty of building applications interpeting them. In this section I will describe the final notion and step of NEMI and make a case that a stochastic-friendly methods are needed for crafting methodologies applied to 'real' data.

No data is perfect, and methods, like most from machine learning, must find optimal ways of approximating the 'underlying' model. However, as demonstrated in Fig. 6, being able to account for the slight variations, for example in the sine curve in the top middle panel, can improve a model's utility. Having a methodology that is able to reflect the uncertainty of the model fit can be highly beneficial. The two-dimensional examples in Fig. 6 are simple cases, but the highly nonlinear BV data poses a more difficult problem. In NEMI, as with any neural network application or optimization algorithm, the method application will determine the best fit given its initial conditions (i.e., parameters), including a stochastic or random seed. In many cases, a slight perturbation in initial conditions can lead to a different result, meaning a different model representation. In NEMI, this would be a different manifold, as was demonstrated in Fig. 11. What this sensitivity to initial conditions means in practice is that there are multiple landscape of possible solutions that the model can converge to and that these different states can be reached given just a small difference in parameters.

The sensitivity to parameters may appear to be a weakness in a methodology, and will be if a model of sufficient utility is not arrived. However, in the application of NEMI to the BV data the slight sensitivity to parameters allows the exploration of the complex covariance space of the BV data. Consequently, NEMI allows an estimation of the uncertainty. Thought of in the framework of bias versus variance, having a good approximation of the variance within the covariance space of the data a methodology describes is highly beneficial. The application of a manifold methodology facilitates this.

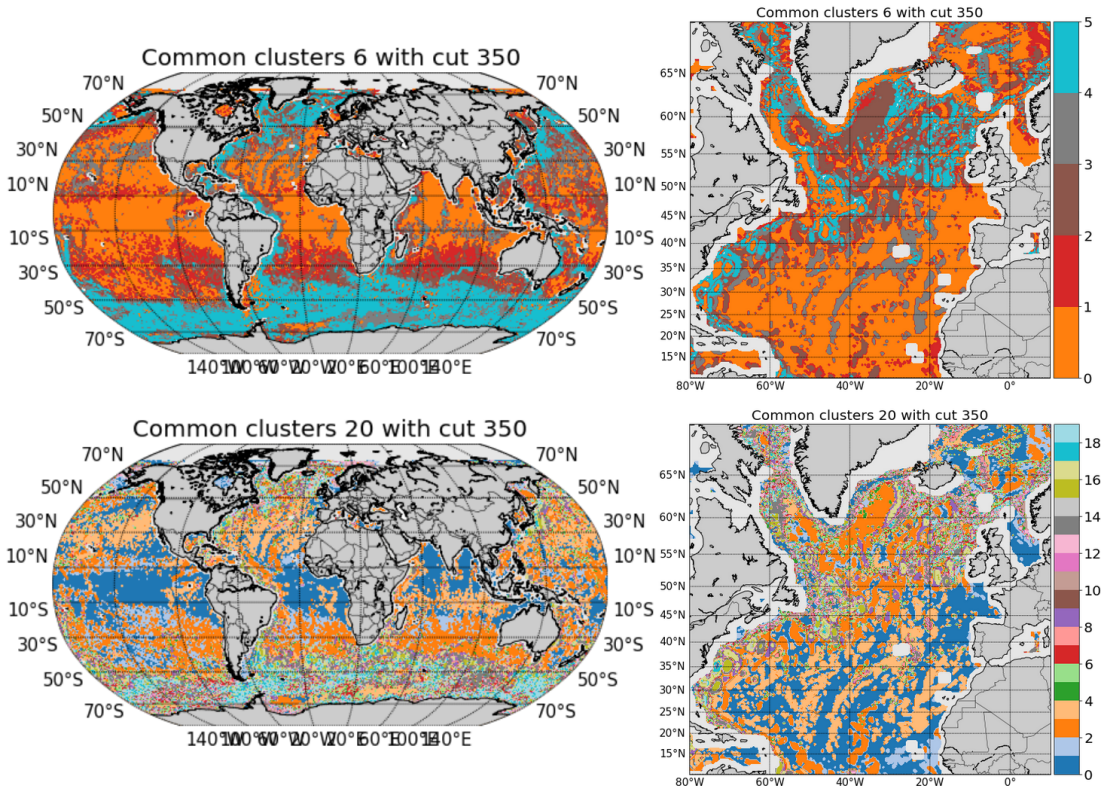

Figure 19. The clusters (BV dynamical regimes). For aggregation level 350, cluster numbers of 6 and 20 are shown. Note that in comparison to Fig. 17 the clusters here are much smoother.

A number of methods and approached can be used to estimate the uncertainty. In the BV example a geographic majority vote was used. This means that for each geographical location, the cluster that was most often flagged throughout the ensemble was the one chosen. Here, this was done largely for simplicity. Note that other methods, such as entropy are highly suitable as described in appendix A. Using entropy would allow the assessment of how *many* different clusters were chosen. If the majority was between two very different ones, this could be important information or If an area were highly contested then our confidence in that area would be lowered. Naturally, having this information is highly valuable in an of itself, and for the interested practitioner I recommend exploring this avenue and feature of NEMI.

5.1 Oceanographic interpretation of regimes

In Fig. 19 the product of applying NEMI to the BV data is shown. An HCA cluster aggregation level of 350 is chosen, and six (top row) and twenty (bottom row) clusters are demonstrated. Comparing to Fig. 19, we can see that while the figures look somewhat similar, the 'static' (the result of the clusters changing ever so slightly) has been greatly suppressed. Overall, the clusters are much smoother and crucially *reproducible*. To illustrate the utility of these choices of cluster numbers, I will briefly give two examples of the utility. Note however, that the number of clusters (here dynamical regimes) is entirely up to the practitioner and will likely depend on the research question at hand. To illustrate the now oceanographic context I will refer to the final cluster products as 'dynamical regimes' as these illustrate an objective empirical leading order analysis of the closed BV equation.

In Fig. 19, the top row shows the large overall dynamical regimes that are very interesting when assessing the global structures. Note that coherent areas in the areas where the wind stress curl ($\nabla \times \tau$) are largely coherent have been grouped together, despite having opposite signs. Note how we know from Fig. 18 and from oceanographic intuition that these areas should be similar but have opposite main drivers. For example, in the Northern Hemisphere in the large wind gyre areas (see Sonnewald et al., 2023 for a detailed theoretical description) the wind stress curl is negative and balanced largely by positive planetary advection. In similar areas in the Southern Hemisphere this effect is opposite, which is also

intuitive due to the symmetry of the Earth around the equator. From a clustering perspective, the fact that the terms are similar allows them to be grouped together, and different areas to stand out more. See for example the grey streaks through the North Atlantic running approximately latitudinally from 43 to 17°N. These are colocated with where significant areas of variability in the sea floor are found (for example the mid-Atlantic ridge). The clustering here illustrates what an important feature this is, and that this should be paid close attention to.

In the bottom row of Fig. 17 an example where there are 20 dynamical regimes is shown. Here, the area where the 'classical' wind gyres are found are seen and the Northern Hemisphere and Southern Hemisphere dynamical regimes are distinguished. Note the increased detail around, for example, the coast. As a thought example, imagine that a current is flowing along the coast. The coasts have large features such as canyons. Moving south to north, a current moving into a canyon would suddenly have more room, and the vorticity contributed by the bottom pressure torque would decrease significantly. As the current moves further north the other side of the canyon would be reached and the current would become more constricted again, where the bottom pressure torque term would increase. These would emerge as separate dynamical regimes in a study where a larger number of dynamical regimes is chosen, but most likely not appear in a study choosing a lower number.

Note that the two examples above use examples where one as 'equal but opposite' scenarios being grouped together. This was chosen as an accessible example but should by no means be seen as the only possible cancelling effect. Recall the complicated covariance space being queried and the highly nonlinear data. Further investigation of the dynamical regimes in the BV equation in MOM6 is the topic of another study.

6 CONCLUSION

Here, I presented the method Native Emergent Manifold Interrogation (NEMI), which is a generalisation of the methodology presented in Sonnewald et al. (2020). NEMI is designed for 'data mining', or put differently, to find underlying patterns within data. Nemi is a generalisation of the methodology in Sonnewald et al. (2020) that targeted plankton ecosystems, in that is is designed to scale to larger datasets. Scaling is a formidable bottlenecks in data mining for scientific applications. In NEMI I have generalised a workflow that can accommodate a wide array of data, where the particular example application used here is geospatial data. An explicitly hierarchical approach is used, making NEMI less parametric (fewer parameters to tune and less danger of noise interference) and intuitively useful both for global (for example the whole Earth in the present example) or more local applications (for example a basin or more regional assessment). NEMI does not use a fixed field-specific benchmark criteria (used in Sonnewald et al. (2020)), but is generalised so a field agnostic option is available. Lastly, NEMI invites the use of a range of uncertainty quantification options in the final cluster evaluation, from a majority vote to entropy. I demonstrate NEMI's application to a numerical ocean model, namely MOM6 (Griffies et al., 2023), and represent the barotropic vorticity balance of a time-mean of a model run. Here, the data serves as an example of a highly nonlinear and complicated covariance structure, within which reside highly valuable oceanographic patterns. NEMI is used to extract these patterns and facilitate further scientific discovery. However, NEMI is entirely general, and can be used on a range of data from the Earth sciences and beyond.

OPEN RESEARCH

The code for the Native Emergent Manifold Interrogation (NEMI) method is available here:
https://github.com/maikejulie/NEMI
DOI: 10.5281/zenodo.7764719

ACKNOWLEDGMENTS

I gratefully acknowledge students and colleagues who have requested details regarding this methodology to the extent that writing it down seemed appropriate.

Funding: Cooperative Institute for Modeling the Earth System, Princeton University, under Award NA18OAR4320123 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and

do not necessarily reflect the views of Princeton University, the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce.

APPENDIX

A: Entropy for uncertainty estimation

Entropy (H) can be used as a measure of uncertainty. As discussed in Clare et al. (2022): In information theory, entropy is the expected information of a random variable, and for each sample i is given by

$$H_i = -\sum_{j=1}^{N_l} p_{ij} \log(p_{ij}), \tag{2}$$

here N_l is the number of possible outcomes for each location and p_{ij} is the probability of each outcome j for sample i (Goodfellow et al., 2016). The larger the entropy, the less skewed the distribution will be and the more uncertain the outcome. The concept of entropy can be directly applied to manage the potentially different results from NEMI for each geographic location within the ensemble. If this is better than a simpler method, such as a majority vote, depends entirely on the application.

REFERENCES

Beucler, T., Ebert-Uphoff, I., Rasp, S., Pritchard, M. S., and Gentine, P. (2021). Machine learning for clouds and climate (invited chapter for the agu geophysical monograph series "clouds and climate").

Clare, M. C., Sonnewald, M., Lguensat, R., and Deshayes, J. (2022). Explainable artificial intelligence for bayesian neural networks: toward trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, 14(11):e2022MS003162.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise.

Fleming, S., Watson, J., Ellenson, A., Cannon, A., and Vesselinov, V. (2021). Machine learning in earth and environmental science requires education and research policy reforms. *Nature Geoscience*, 14.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Götz, M., Richerzhagen, M., Bodenstein, C., Cavallaro, G., Glock, P., Riedel, M., and Benediktsson, J. A. (2015). On scalable data mining techniques for earth science. *Procedia Computer Science*, 51:2188–2197. International Conference On Computational Science, ICCS 2015.

Harvey, A. C. (1982). Spectral analysis and time series, m. b. priestly. two volumes, 890 pages plus preface, indexes, references and appendices, london: Academic press, 1981. price in the uk: Vol. i, £49-60: Vol. ii, £20-60. *Journal of Forecasting*, 1(4):422–423.

Hughes, C. W. and de Cuevas, B. A. (2001). Why Western Boundary Currents in Realistic Oceans are Inviscid: A Link between Form Stress and Bottom Pressure Torques. *Journal of Physical Oceanography*, 31(10):2871–2885.

MacQueen, J. (1965). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Math., Stat., and Prob*, page 281.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861.

Munk, W. H. (1950). ON THE WIND-DRIVEN OCEAN CIRCULATION. *Journal of Meteorology*, 7(2):80–93.

Sonnewald, M., Dutkiewicz, S., Hill, C., and Forget, G. (2020). Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science advances*, 6(22):eaay4740.

Sonnewald, M. and Lguensat, R. (2021). Revealing the impact of global heating on north atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, 13(8):e2021MS002496. e2021MS002496 2021MS002496.

Sonnewald, M., Lguensat, R., Jones, D., Düben, P., Brajard, J., and Balaji, V. (2021). Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, 16.

Sonnewald, M., Wunsch, C., and Heimbach, P. (2018). Linear predictability: A sea surface height case study. *Journal of Climate*, 31:2599–2611.

Sonnewald, M., Wunsch, C., and Heimbach, P. (2019). Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, 6(5):784–794.

Stommel, H. (1948). The westward intensification of wind-driven ocean currents. *Transactions, American Geophysical Union*, 29(2):202–206.

Sverdrup, H. U. (1947). Wind-driven currents in a baroclinic ocean; with application to the equatorial currents of the eastern pacific. *Proceedings of the National Academy of Sciences*, 33(11):318–326.

van der Maaten, L. and Hinton, G. (2008). Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.