ABSTRACT. We use importance sampling in a redefined way to highlight and investigate rare events in the form of trajectories trapped inside a target almost invariant set. We take a transfer operator approach to finding almost invariant sets of a reconstructed 2-dimensional flow of the atmosphere from wind velocity fields provided by the Portable University Model of the Atmosphere. Motivated by extreme value theory, we consider an observable \( \phi(x) = -\log(d(x, \gamma)) \) maximized at the center \( \gamma \) of a chosen target almost invariant set. We illustrate that importance sampling using this observable provides an enriched data set of trajectories that experience a rare event, in this case defined as ending near \( \gamma \). Backwards reconstruction of these trajectories provides valuable information on initial conditions and most likely paths a trajectory will take toward a rare event.

In this specific setting, one can think of an almost invariant set as a region in the atmosphere where only a small number of particles move in and out of the region. For short time intervals, these regions are represented physically by individual eddies in the atmosphere; while longer time intervals suggest formation by eddy interaction such as particles trapped between the spin of two eddies. Particle movement in and interaction with these sets can provide useful information on the most probable path of a storm (in the former case), or origin of pollution (in the latter), for example.

1. INTRODUCTION

Atmospheric eddies play a major role in extreme weather phenomenon such as heatwaves, hurricane movement, and pollution distribution \[15, 16\]. Topologically, these eddies can be seen as almost invariant sets of a flow where there is minimal movement of particles from within and outside of a region \[8\]. Understanding where these eddies occur and the likelihood of trajectories ending up inside them can provide a new and useful perspective on atmospheric movement.

We reconstruct a 2-dimensional model of atmospheric flow defined on a space \( Y \subseteq \mathbb{R}^2 \) from wind velocity fields provided by the Portable University Model of the Atmosphere (PUMA) \[13\]. Following recent literature, we estimate the transition probability matrix...
(TPM) of the flow on a subspace $X \subset Y$ by taking a fine grid of boxes and measuring transitions of particles from one box to another over a finite time interval. Intuitively, one can think of an almost invariant set as a region where only a small number of particles move in and out of the region. We take a transfer operator approach to finding these almost invariant sets by approximating the Perron-Frobenius operator with the associated TPM. Spectral properties of this operator provide information on the invariant structure of the space. This approach is discussed in detail in [5, 6, 4, 7] and applied numerically to an ocean flow model in [8]. For adaptation purposes, our methods differ slightly from those of the listed literature by including variations on the TPM [14] and employing a spectral clustering approach equipped with $K$-means [17, 1, 2]; however the foundational arguments remain the same. We discuss this foundation and set of differences in detail in the following sections.

Motivated by extreme value literature [10, 12], we consider an observable $\phi(x) = -\log(d(x, \gamma))$ where $\phi : Z \rightarrow \mathbb{R}$ is defined for every $x \in Z$, $X \subset Z \subseteq Y$ and $\gamma$ is the euclidean center of the almost invariant set. In this way trajectories of the observable under the flow are maximized as they approach the center of the almost invariant set. Under some flow $f_t : Y \rightarrow Y$, it is often of interest to consider a set of random variables defined by $X_N = \phi \circ f_t(x_N)$ for a set of initial values $x_N \in Z$ at some fixed time $t$. For our choice of $f_t$ and $\phi$, we show that $(X_N)$ behaves as though it comes from some shifted exponential distribution where $f_t(x_N) \rightarrow \gamma$ gives $X_N \rightarrow \infty$.

We apply an importance sampling method, called genealogical particle analysis (GPA), that exponentially tilts the distribution of $(X_N)$ so that the probability of observing larger values (and hence, values of $f_t(x_N)$ closer to $\gamma$) is increased [3, 18, 15]. GPA works by killing and cloning trajectories under the flow at specified sampling times based on a weight function that determines the performance of a trajectory. Large values of the weight function indicate that a trajectory behaves as though it comes from the target (tilted) distribution. In the end, the surviving trajectories represent the set that has a higher probability of ending near $\gamma$. Backwards reconstruction of these trajectories allows us to find the set of most likely paths that end in the almost invariant set within a specified time interval.

2. Transfer Operators and Almost Invariant Measures and Sets

We begin with a set of definitions from [6] that will be used to establish reasoning for the methods used in this investigation. A function $p : X \times \mathcal{B} \rightarrow [0, 1]$ in space $X$ with associated $\sigma$-algebra $\mathcal{B}$ is said to be a stochastic transition function if

1. $p(x, \cdot)$ is a probability measure $\forall x \in X$
2. $p(\cdot, A)$ is Lebesgue measurable $\forall A \in \mathcal{B}$

If $\mu \in \mathcal{M}$, with $\mathcal{M}$ the space of all probability measures on $\mathcal{B}$, satisfies

$$\mu(A) = \int p(x, A) \, d\mu(x) \ \forall A \in \mathcal{B}$$
then \( \mu \) is said to be an *invariant measure* of \( p \). Under the assumption that the probability measure \( p(x, \cdot) \) is absolutely continuous w.r.t Lebesgue \( m \) \( \forall x \in X \) we have,

\[
p(x, A) = \int_A k(x, y) \, dm(y) \quad \forall A \in \mathcal{B}
\]

with **transition density function** \( k : X \times X \rightarrow \mathbb{R} \), \( k(x, \cdot) \in L^1(X, m) \) and \( k(x, y) \geq 0 \).

The Perron-Frobenius operator \( P : \mathcal{M} \rightarrow \mathcal{M} \) is defined by,

\[
P\mu(A) = \int p(x, A) \, d\mu(x).
\]

If \( p \) is absolutely continuous with density function \( k \), then \( P \) is defined on \( L^1 \) by,

\[
Pg(y) = \int k(x, y)g(x) \, dm(x) \quad \forall g \in L^1.
\]

A measure \( \mu \in \mathcal{M} \) is invariant if and only if it is a fixed point of \( P \) e.g. the invariant measures correspond to eigenfunctions of \( P \) for the eigenvalue \( \lambda = 1 \) (\( P\mu = \lambda \mu = \mu \)).

If we let \( p(x, \cdot) = \delta_{f(x)} \) with \( f : X \rightarrow X \) a diffeomorphism on \( X \) then one can show by definition,

\[
\mu(A) = \int p(x, A) \, d\mu(x) = \mu(f^{-1}(A))
\]

Hence, in knowing the spectral information of the operator \( P \) for deterministic systems we can obtain invariant measures for the diffeomorphism \( f \). Of course, it is important in our case to consider a stochastic system rather than a deterministic one. In this sense, we may start by introducing a small perturbation into the original system. The density function \( k \) associated with this \( \varepsilon > 0 \) perturbed system is introduced in [4] as,

\[
k_\varepsilon(x, y) = \frac{1}{\varepsilon^nm(B)}1_B\left(\frac{1}{\varepsilon}(y - x)\right), \quad x, y \in X
\]

where \( B = B(0) \) is an open ball of radius 1 in \( \mathbb{R}^n \). Then,

\[
p_\varepsilon(x, A) = \int_A k_\varepsilon(f(x), y) \, dm(y)
\]

and the Markov process defined by any initial probability measure \( \mu \) and transition function \( p_\varepsilon \) is a small random perturbation of the deterministic system \( f \). However, many invariant maps form more complicated structures in the phase space with regions of periodic and chaotic movement occupying closely neighboring spaces. In this case, perturbations of the map may cause the space to no longer be decomposed into invariant sets [11]. Fortunately for certain small enough perturbations, these can be recovered as *almost invariant* sets. Informally, an almost invariant set is one where the probability of mapping from set \( A \) back to set \( A \) is [4],

\[
\rho(A) = \frac{\int_A p_\varepsilon(x, A)d\mu(x)}{\mu(A)} \approx 1.
\]
Furthermore, $p_\varepsilon(x, \cdot) \to \delta(f(x))$ as $\varepsilon \to 0$ and hence in the limit we have the deterministic setting \[4\],
\[
\int_A p_0(x, A) d\mu(x) = \int_A \delta(f(x))(A) d\mu(x) = \mu(f^{-1}(A) \cap A).
\]

If $f_\varepsilon$ is a differentiable, invertible time $t$ map of a smooth flow, then $P_\varepsilon$ simplifies to
\[
P_\varepsilon g(y) = \frac{g(f_\varepsilon^{-1}y)}{|\det Df_\varepsilon(f_\varepsilon^{-1}y)|}
\]
where $\|P_\varepsilon\|_1 = 1$ and $P_\varepsilon g \geq 0$ if $g \geq 0$ \[7\]. Numerically, $P_\varepsilon$ is often approximated by a finite dimensional Galerkin approximation based on a fine partition $\{B_1, \ldots, B_m\}$ of $X$ \[4, 6, 8\].

Following the setup of \[8\] we let $(x, t, \tau) \to f_\varepsilon(x, t; \tau)$ represent the non-autonomous flow where $[t, t + \tau]$ is the full time interval of interest and $\tau$ is taken on the order of a day. In this way, $f_\varepsilon(x, t; \tau)$ is the end position in $X$ of a trajectory beginning at $x \in X$, time $t$, and flowing for $\tau$ time. A trajectory $x(t) := f_\varepsilon(x_0, t_0; t)$ is a solution to the non-autonomous ODE $\frac{dx}{dt} = V(x(t), t)$ with initial condition $x(t_0) = f_\varepsilon(x_0, t_0; 0)$ and vector field $V$. For our purposes, we will only consider the case when $X \subseteq \mathbb{R}^2$, however many of these results extend in an obvious way to flows on $\mathbb{R}^n$. The transition probability matrix that is formed under the flow from $t$ to $t + \tau$ given by,
\[
P_{t,\tau,i,j} = \frac{m(B_i \cap f_\varepsilon(B_j, t + \tau; -\tau))}{m(B_i)}
\]
represents the approximation to $P_\varepsilon$.

Remark. It is sometimes of interest to consider the flow over a restricted region $X$ in the larger space. In this setting, a point $x$ may flow outside of the region $X$ in $t + \tau$ time causing the rows of the TPM to no longer sum to 1. Normalization of the rows has been successfully applied in the literature \[8\]; however, a more natural setting for this study is allowing $x$ to flow outside of $X$. To accomplish this, we introduce an "external" tile \[14\] to the system represented by an additional row/column pair in $P_{t,\tau,i,j}$ with the probability of transitioning into this tile equal to $1 - \sum_{j=1}^m P_{t,\tau,i,j}$ and transitioning out to any tile with equal probability. Intuitively, the addition of this tile can be thought of as a point $x$ flowing under the map until it reaches the boundary of $X$ and then being sufficiently mixed before being injected back into $X$.

The following nontrivial result from \[3, \text{Prop. 5.7}\] relates the measure $\rho(A)$ to any eigenvalue $\lambda$ of the corresponding Perron-Frobenius operator $P_\varepsilon$ (and hence, $P_{t,\tau,i,j}$),
\[
\lambda \mu(A) = (\rho(A) + \rho(X \setminus A) - 1) \mu(A).
\]

When $\lambda \approx 1$ we have that the probability measure $\mu$ is close to the invariant measure of the system. In a similar way, if we consider the right hand side where the sets $A$ and $X \setminus A$ form a partition of the space then finding an almost invariant measure can be viewed as a maximization problem of both $\rho(A)$ and $\rho(X \setminus A)$. This approach is discussed in detail in \[3\].
3. Graph Partitioning and Invariant Sets

Following the relationship described in (1), we may consider $q$ partitions of the space where we seek to maximize the value,

$$
\rho(A_1, \ldots, A_q) = \frac{1}{q} \sum_{k=1}^{q} \rho(A_k)
$$

by varying the partitions $A_1, \ldots, A_q$ such that $A_k \cap A_\ell = \emptyset$ for $k \neq \ell$ and $\bigcup_{k=1}^{q} A_k = X$. For our $\varepsilon$-perturbed flow $f_\varepsilon$, we may equivalently write the measure $\rho$ of any set $A$ as,

$$
\rho(A) = \frac{\mu(A \cap f_\varepsilon(A, t + \tau; -\tau))}{\mu(A)}.
$$

Given a fine partition of the space $X = \{B_1, \ldots, B_m\}$ and a corresponding transition matrix $P_{t,\tau,i,j}$ let,

$$
p_i = \frac{m(B_i)}{m(X)},
$$

where in practice $m(B_i) =$ area of $B_i$ if $f_\varepsilon$ is the defined flow on $\mathbb{R}^2$, for example. In this setting it is natural to define the probability measure \cite{6} by,

$$
\mu_m(A) = \sum_{i=1}^{m} \frac{m(A \cap B_i)}{m(B_i)} p_i.
$$

In fact, $\mu_m \rightarrow \mu$ strongly for small $\varepsilon$ perturbed deterministic dynamical systems \cite{4}. From \cite[Prop. 6.4]{6} we have that for a subset $A = \bigcup_{i \in I} B_i$ made up of discrete boxes $B_i \subset X$ which form the fine partition of the space of the flow $f_\varepsilon$ and $I \subset \{1, \ldots, m\}$ of indices,

$$
\rho(A) \approx \frac{\sum_{i,j \in I} p_i P_{t,\tau,i,j} \sum_{i \in I} p_i}{p_i}
$$

Generally, $P_{t,\tau,i,j}$ is not reversible because,

$$
p_i P_{t,\tau,i,j} \neq p_j P_{t,\tau,j,i}
$$
or equivalently,

$$
\mu(B_i \cap f_\varepsilon(B_j, t + \tau; -\tau)) \neq \mu(B_j \cap f_\varepsilon(B_i, t + \tau, -\tau)).
$$

In other words, the probability of being in box $B_i$ at time $t$ and then $B_j$ at time $t + \tau$ is not equal to the probability of being in box $B_j$ at time $t$ and then $B_i$ at time $t + \tau$. However, if we define the time-reversed quantity,

$$
\hat{\rho}(A) = \frac{\sum_{i,j \in I} p_i \hat{P}_{t,\tau,i,j} \sum_{i \in I} p_i}{p_i}
$$

where $\hat{P}_{t,\tau,i,j} = p_j P_{t,\tau,j,i}/p_i$ then in \cite{5} it is shown that $\rho(A) = \hat{\rho}(A)$. A major consequence of this relation is that the total cost function $\rho(A_1, \ldots, A_q)$ remains unchanged under time
reversal and without loss of generality we may replace $P_{t,\tau,i,j}$ with the time reversible matrix with entries,

$$R_{t,\tau,i,j} = \frac{1}{2} \left( P_{t,\tau,i,j} + \frac{p{j}P_{t,\tau,j,i}}{p{i}} \right)$$

(5)

The authors in [6] have shown that solving a relaxation of the min-cut problem produces near optimal values of the total cost function. Here, the greedy algorithm looks for a predefined number of $q$ partitions on the graph defined by the Laplacian corresponding to the symmetric matrix $(P_{t,\tau,i,j} + P_{t,\tau,j,i})/2$. On the other hand, it is shown in [5] that near optimal values can be obtained by performing a weighted fuzzy clustering algorithm on the space generated by the eigenvectors of the time reversible matrix $R_{t,\tau}$ directly.

We take a different approach by performing spectral clustering equipped with $K$-means to the reversible matrix $R_{t,\tau}$. This approach combines the benefits of both methods introduced in [5] and [6]. The Laplacian of the matrix $R_{t,\tau}$ ensures the eigenvalues are positive and real. The eigenvalues near zero and corresponding eigenvectors form the lower-dimensional subspace upon which the vectors of $R_{t,\tau}$ are projected. This projection can be understood as a lower-dimensional representation of $R_{t,\tau}$ without much loss of information. On the other hand, choosing the matrix $R_{t,\tau}$ over the symmetrized $P_{t,\tau}$ matrix provides spectral information on the number of expected almost-invariant sets and hence the appropriate $K$ value prior to clustering.

We use the unnormalized graph Laplacian of the time reversible matrix $R_{t,\tau}$ defined by,

$$L_{t,\tau} = D_{t,\tau} - R_{t,\tau}$$

where $D_{t,\tau}$ is the diagonal matrix with entries $D_{t,\tau,i,i} = \sum_i R_{t,\tau,i,j}$. The matrix $L_{t,\tau}$ is symmetric and positive semi-definite with $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$. Moreover, solving for the optimal $q$ partition of $L_{t,\tau}$ is equivalent to solving a relaxation of the min-cut problem [17].

The spectral approach to partitioning $L_{t,\tau}$ involves projecting the eigenvectors of $L_{t,\tau}$ onto a smaller $\ell$-dimensional subspace made of the ”most important” eigenvectors of $L_{t,\tau}$. Provided the perturbation $\varepsilon$ is sufficiently small to produce a stochastic, irreducible matrix $R_{t,\tau}$, $L_{t,\tau}$ has one eigenvalue equal to 0 and $q - 1$ eigenvalues close to 0 representing the $q$ (almost) invariant sets under the flow $f_\varepsilon$ [6 Thrm. 7.1]. Hence, the subspace $\text{sp}\{v_2, \ldots, v_q\}$ can be used to identify the $q$ almost invariant sets of $P_{t,\tau}$ where $\ell = q - 1$.

In practice, it is common to choose fewer than $q - 1$ eigenvectors of $L$ with nearly optimal results from clustering algorithms [9]. In our application, we choose to project onto $\text{sp}\{v_2\}$, that is an $\ell = 1$-dimensional subspace made of $v_2$ corresponding to the second smallest eigenvalue of $L_{t,\tau}$.

Given $m$ blocks $B_i$, each $B_i$ is represented by a point in $\mathbb{R}^{m+1}$ where the coordinate value is given by the corresponding probability transition values to each block $B_j$ provided in $R_{t,\tau}$. The goal of $K$-means is then to find hyperplane separations between the clusters in $\mathbb{R}^{m+1}$ by minimizing the objective function defined by the sum of the inner cluster distances. We refer the reader to Appendix A for a more rigorous explanation. Introducing $L_{t,\tau}$ allows us to run $K$-means using the projected, lower-dimensional representation of the vectors in $R_{t,\tau}$ which often results in a set of more easily separable clusters. The result from $K$-means
clustering with \( K = q \) initial centroids will provide our optimal \( q \) spatial partition result for \( R_{t,\tau} \). For a flow \( f_\epsilon \) representing movement in the atmosphere the partitioning result for \( R_{t,\tau} \) provides the geographic location of almost invariant sets (or eddies) where movement is trapped over the time interval \([t, t + \tau]\) for \( \tau \) on the order of a day.

4. Importance Sampling: Finding Trajectories Likely to End in an Almost Invariant Set

We will refer to the subset \( A = \bigcup_{i \in I} B_i \) made up of \( B_i \subset X = \mathbb{R}^2 \) as the almost invariant set over the time interval \([t, t + \tau]\) for \( \tau \) on the order of a day. We will require that \( A \) be connected and define the center of \( A \) as \( \gamma = ((x(\min) - x(\max))/2, (y(\min) - y(\max))/2) \). We consider the observable

\[
\phi(x) = -\log(d(x, \gamma)) \quad x \in X
\]

where \( d \) is the Euclidean metric so that \( \phi(x) \) is maximized as it approaches the center of the almost invariant set. Let \( X_N = \phi \circ f_\epsilon(x_{N,0}, t; \tau) \), where \( x_{N,0} \) is the \( N \)th starting position, be a sequence of random variables representing the value of our observable \( \phi \) as a function of the end position of the \( N \) trajectories on \( X \) run under the flow \( f_\epsilon \) from \( t \) up to time \( t + \tau \).

Remark. The importance sampling methods outlined in this paper are suggested as a starting point for almost invariant sets of any topology; however, they work best for symmetric, convex sets by the definition of the center \( \gamma \). It is possible to obtain varying results for different choices of \( \gamma \) even for nonsymmetric convex sets; for example, consider the set \( A \) as an ellipse and \( \gamma \) as either one of the focus points. Our justification for using this method relies on the fact that GPA results in a spatial distribution of points only with the highest density near \( \gamma \).

Remark. Allowing \( \phi(x) = -\log d_H(x, S) \) where \( S = \{x : x \in \delta A\} \) and \( d_H = \min d(x, S) \) is the Hausdorff distance to \( S \) would be an interesting and more physically relevant choice for many climate questions. We have tested this definition and have found that it introduces several new complexities which future work would have to overcome including: (1) the distribution of \( X(x) = \phi \circ f_\epsilon(x) \) is numerically unclear with little theoretical support and (2) GPA would inevitably force surviving trajectories to group near the boundary.

Genealogical particle analysis (GPA) is an importance sampling method that uses weights to perform a change of measure in a reversible way so that rare events are sampled more often. These weights can be thought of as measuring the performance of the trajectory at specified sampling times. Large values of the weight function imply that the trajectory is behaving as if it comes from the target distribution. These trajectories will be cloned while low weight values indicate a trajectory that will be killed. Importance sampling algorithms are often used to lower relative error of tail probability estimation because the change of measure provides a set of trajectories that are more likely to end in a rare event. In our context, running GPA will provide a pool of trajectories that are most likely to end in our almost invariant set over the time interval \([t, t + \tau]\).
One difficulty with GPA is determining a weight function that will change the measure in an appropriate way so that rare events are sampled more often. This choice depends on the distribution of $X_N$. We describe the explicit method used in this paper, however, we refer to [3, 15, 18] for more details on the general case. We assume that $X_N$ is distributed according to a unimodal distribution with tails decaying fast enough ($o(1/N^{1+\delta})$) to zero.

GPA has been successfully implemented numerically for random variables with symmetric, heavy-tailed distributions [3, 15, 18]. This assumption is supported by the numerical behavior of the distribution of $X_N$ (see Figure 1).

Remark. Preliminary work using GPA on the Lorenz '63 model equipped with the observable $\phi(x) = -\log(d(x, \gamma))$ where $\gamma \in X$ is some fixed point in the space also suggests $X(x) = \phi \circ f_\varepsilon(x)$ follows a unimodal distribution with fast-decaying tails.

1. Initiate $N = 1, \ldots, M$ starting trajectories uniformly distributed over the space $X$.
2. For $n = 1 \ldots \lfloor \tau/T \rfloor$ where $\tau$ is the final integration time. $T$ is referred to as the resampling time.

Remark. It is important to balance $T$ between the correlation time of $X_N$ and the Lyapunov time. Values of $T$ taken too small can result in highly correlated trajectories (many of clones of a single trajectory) while too large can result in a relaxation back to the original distribution.

2a. Iterate each trajectory from time $t_{n-1} = T(n - 1)$ to $t_n = Tn$.
2b. At time $t_n$, stop the simulation and assign a weight to each trajectory $x_{n,N}$ given by,

$$W_{N,n} = \frac{\exp(C(\phi \circ f_\varepsilon(x_{n-1,N}, t_{n-1}, t_n) - \phi \circ f_\varepsilon(x_{n-2,N}, t_{n-2}, t_{n-1})))}{Z_n}$$

where

$$Z_n = \frac{1}{N} \sum_{N=1}^{M} W_{N,n}$$

2c. Determine the number of clones produced by each trajectory,

$$c_{N,n} = \lfloor W_{N,n} + u_N \rfloor$$

where $\lfloor \cdot \rfloor$ is the integer portion and $u_N$ are random variables generated from a uniform distribution on $[0, 1]$.

2d. The number of trajectories present after each iteration is given by,

$$M_n = \sum_{N=1}^{M} c_{N,n}$$

Clones are used as inputs into the next iteration of the algorithm. For large $N$, the normalizing factor ensures the number of particles $N_n$ remains constant; however, in practice the number of particles fluctuates slightly on each iteration $n$. To ensure $N_n$ remains constant it is common to compute the difference $\Delta N_n = N_n - M$. If $\Delta N_i > 0$, then $\Delta N_i$ trajectories are randomly selected.
SOURCES AND SINKS OF RARE TRAJECTORIES IDENTIFIED BY IMPORTANCE SAMPLING  

(without replacement) and killed. If \( \Delta N_i < 0 \), then \( \Delta N_i \) trajectories are randomly selected (with replacement) and cloned.

3. The final set of trajectories \( \tilde{X}_N = \phi \circ f_\epsilon (x_{\tau/T-1,N}, t_{\tau/T-1}, \tau) \) tends to a new distribution as \( N \to \infty \) exponentially tilted by the constant \( C \).

![Figure 1. Exponentially tilted distributions for different \( C \) values under GPA. Densities are estimated with a normal kernel. \( C = 0 \) corresponds to the original distribution.](image)

The probability of sampling events for increasing values of \( \phi \circ f_\epsilon (x, t; \tau) \) under the exponentially tilted distribution increases for increasing values of \( C \). Since \( \phi \) is maximized at \( \gamma \), the set of end trajectories is the set that has the highest probability of entering or remaining in the almost invariant set over the time interval \([t, t + \tau]\). For a flow \( f_\epsilon \) representing movement in the atmosphere this set of trajectories form the paths of particle movements which are most likely to enter or remain near the center of a target eddy.

5. AN APPLICATION TO THE PORTABLE UNIVERSITY MODEL OF THE ATMOSPHERE FOR POLLUTION MOVEMENT

PORTABLE UNIVERSITY MODEL OF THE ATMOSPHERE (PUMA) AND RECONSTRUCTING THE ATMOSPHERIC FLOW. The PUMA module is a subset of PlaSim, a planet simulation model of intermediate complexity provided by the Universitat Hamburg Meteorological Institute. Like most atmospheric models, PUMA is a simplified model derived from the Navier Stokes equation in a rotating frame of reference. In contrast to a full atmospheric general circulation model (GCM), moist processes such as humidity, cloud water, evaporation and the like are omitted. For more information on this model we refer to [13]. Among the many available output options in PUMA, one is the time-dependent northern
and eastern wind velocities for a user-defined set of latitude and longitude pairs taken at a number of different atmospheric levels.

Our goal for this section is to reconstruct a 2-dimensional time dependent atmospheric flow model. We begin with a set of time series of northern and eastern wind velocities each taken at a single (lat, lon) pair (64 latitude and 128 longitude in total) making up our whole space $Y$ and 1-day intervals with geopotential height equal to 1,000 hPA (approximately 100 meters above sea level). We will occasionally refer to a set of (lat, lon) pairs as the grid and a single (lat, lon) pair as a grid point. Values of the wind velocity are given initially in m/s. We convert this value to km/day. We create a new set of time series with values taken at 0.1-day intervals by linearly interpolating the km/day time series at each grid point over time. Let

$$V(\vec{x}(t), t) = \begin{bmatrix} V_x(\vec{x}(t), t) \\ V_y(\vec{x}(t), t) \end{bmatrix}$$

where $V_x(\vec{x}(t), t)$ is the eastern velocity magnitude and $V_y(\vec{x}(t), t)$ is the northern velocity magnitude for $\vec{x}(t) = (x(t), y(t))$ points on the grid at fixed time $t$. Then a trajectory $\vec{x}(t) := F_t(\vec{x}_0, t_0; t)$ is a solution to the non-autonomous ODE $\frac{d\vec{x}}{dt} = V(\vec{x}(t), t)$ and $F_t(\vec{x}, t, \tau)$ is the end position of a trajectory in the 2-dimensional space of latitude longitude pairs beginning at $\vec{x}$, time $t$ and flowing for $\tau$ time. $F_t$ can be understood as some $\varepsilon$ perturbation $f_\varepsilon$ of the deterministic flow $f$ in the previous discussion.

We numerically approximate $F_t$ by the following Runge-Kutta method. We first note that, by design, there is a fixed vector field at every $10/\text{day}$ of a day. Hence, we may integrate over the fixed vector field at each discrete time step $t_n$ where $t_n = 0.1n$ day(s) from $t$ (starting time of the flow), $T = 0.1$, and $n = 1, \ldots, \lceil \tau/T \rceil$. For this application, we choose $t = 360$ and $\tau = 4$ days.

PUMA and many standard GCMs require time for the system to reach an state where the initial conditions of the model (temperature gradients, $CO_2$ concentrations, radiation, etc.) provide meaningful and accurate time-series outputs (velocity fields, temperature, etc.). In a dynamical sense, this is time it takes for a trajectory to move from its initial position to the (chaotic) attractor. We choose a starting time $t \geq 360$ days to ensure the model has reached this state. Furthermore, we choose our specific $t$ based on a predetermined guarantee that (at least) one eddy will pass through our region of interest by running a simulation of the velocity field movement over several different time-intervals.

Let $F_t(\vec{x}_{n-1}, t_{n-1}, t_n) = \vec{x}_n = \begin{bmatrix} x_n \\ y_n \end{bmatrix}$ represent the discretized flow and calculate,

$$\ell_1 = V(\vec{x}_n, t_n) = \begin{bmatrix} V_x(\vec{x}_n, t_n) \\ V_y(\vec{x}_n, t_n) \end{bmatrix} = \begin{bmatrix} V_x(x_n, y_n, t_n) \\ V_y(x_n, y_n, t_n) \end{bmatrix}$$

$$\ell_2 = V(x_n + 0.05\ell_1(1), y_n + 0.05\ell_1(2), t_n)$$

$$\ell_3 = V(x_n + 0.05\ell_2(1), y_n + 0.05\ell_2(2), t_n)$$

$$\ell_4 = V(x_n + 0.1\ell_3(1), y_n + 0.1\ell_3(2), t_n)$$
And,

\[ \vec{x}_{n+1} = \vec{x}_n + 0.1(\ell_1 + 2\ell_2 + 2\ell_3 + \ell_4)/6 \]

where \( \vec{x}_{n+1} \) is the end position of the trajectory starting at \( \vec{x}_n \) and flowing from time \( t_n \) to \( t_{n+1} \). For those end trajectories \( \vec{x}_n \) that lie between the original grid points \((x,y)\), we linearly interpolate the vector field \( V \) over the \((x,y)\) grid at the fixed time \( t_n \) to obtain \( V(\vec{x}_n,t_n) \).

**Computing the Transition Probability Matrix over a Fixed Time Interval.**

We begin by converting the latitude longitude coordinates to kilometers from the origin (equator, center) so that integration over the km/day velocity is appropriate. We create a finer grid over the box \([36^\circ N - 70^\circ N, 11^\circ W - 25^\circ E]\) corresponding to Europe at \(1^\circ\) intervals. We will refer to this box as \( X \subset Y \) and each box made of four corners of \((\text{lat, lon})\) pairs inside \( X \) as \( B_i \) \( i = 1 \ldots m \). At this resolution, \( m = 4,896 \) making up a \( 68 \times 72 \) size grid of boxes. In each \( B_i \), we take \( k = 25 \) uniformly distributed points \( \vec{b}(k,i) \) with coordinates (in km) at \( \vec{b}(k,i) = (x(k,i), y(k,i)) \) and integrate to find the path over a fixed time interval \([t, t + \tau]\) using the Runge-Kutta method described above.

**Remark.** We are interested in restricting our search for almost-invariant sets to the small region of the atmosphere over Europe and investigating those sets that may be the result of multiple eddy interactions. To locate individual eddies in an expanded region, it is often of interest to first use the Okubo-Weiss criterion as a Eulerian way of quickly approximating the location of an eddy based on a combination of the vorticity, stretching, and shearing rate of the velocity field. After the location of an eddy is estimated, one restricts numerics to the identified region and uses the transfer operator approach to find the almost invariant set based on the Lagrangian dynamics inside the restricted region [14].

We compute the transition probability \( P_{t,\tau,i,j} \) as the number of points \( \vec{b}(k,i) \) that start (at time \( t \)) in \( B_i \) and end (at time \( t + \tau \)) in \( B_j \) [6]. We refer to Figure 2 for an illustration of this flow. Note that the \( \sum_{j=1}^m P_{t,\tau,i,j} \) is not necessarily equal to 1 because it is possible for \( \vec{b}(k,i) \) to land outside of the large box \( X \) under the flow. Following the work of [14] we introduce a mixing “tile” into this system defined as a box \( B_{m+1} \) in which all \( \vec{b}(i,k) \) landing outside of \( X \) are injected with equal probability into any \( B_i \) in the next step. Addition of this tile is done at the end of computation of the transition probability matrix \( P_{t,\tau} = P_{t,\tau,i,j} \) by adding a column at the \( i = m + 1 \) position with value(s) equal to \( 1 - \sum_{j=1}^m P_{t,\tau,i,j} \) for each \( j \) and adding a row at the \( j = m + 1 \) position with value(s) equal to \( 1/(m+1) \) for each \( i \).

Next, the probability measure on \( X = \cup_{i=1}^m B_i \) is approximated by (3) where,

\[ p_i = \frac{m(B_i)}{m(X)} = \frac{\text{Area of } B_i}{\text{Area of } X} \]

so that the cost function is exactly as stated in (4), that is for any \( A = \cup_{i \in I} B_i, I = \{1, \ldots, m\} \)

\[ \rho(A) \approx \frac{\sum_{i,j \in I} p_i P_{t,\tau,i,j}}{\sum_{i \in I} p_i} \]
Hence, we may solve for the almost invariant sets of the system by looking at the corresponding time-reversible matrix $R_{t,\tau}$ given by (5) with entries,

$$R_{t,\tau,i,j} = \frac{1}{2} \left( P_{t,\tau,i,j} + \frac{p_j P_{t,\tau,j,i}}{p_i} \right)$$

**Figure 2.** Starting boxes $B_i$ (black) and particles under the flow $F_t$ at time $t + \tau$. Velocity field at time $t + \tau$ is indicated by arrows.

**Spectral Clustering of the Time-Reversible Matrix.** We perform spectral clustering of $R_{t,\tau}$ to find the almost invariant sets in the box $X$ defined as the European spatial region in the following steps.

1. Form the unnormalized Laplacian of the matrix $R_{t,\tau}$,

$$L_{t,\tau} = D_{t,\tau} - R_{t,\tau}$$

where $D_{t,\tau}$ is the diagonal matrix with entries $D_{t,\tau,i,i} = \sum_{i=1}^{m} R_{t,\tau,i,j}$.

2. Choose the first $\ell$ eigenvalues $\lambda_1, \ldots, \lambda_\ell$ and corresponding eigenvectors $v_1, \ldots, v_\ell$ of $L_{t,\tau}$.

3. Form a subspace made of $S_\ell := \text{sp}\{v_1, \ldots, v_\ell\}$ and project the $m$-dimensional row vectors of $L_{t,\tau}$ onto $S_\ell$.

4. Run $K$-means on the projected $\ell$-dimensional row vectors with a predetermined $K$ value. $K$ essentially tells the algorithm how many almost invariant sets are expected.
Remark. Although the first $q$ eigenvalues $\lambda_1, \ldots, \lambda_q \approx 0$ of the matrix $L_{t, \tau}$ can give a rough estimate on the number of almost invariant sets in the space (and hence $K$), this may result in finer level sets of the same “parent” almost invariant set.

The boxes $B_i$ which form the partition of the space $X$ are now represented by $m$-dimensional points in $\mathbb{R}^m$. The $i$th row is an $m$-dimensional vector with entries corresponding to the transition probabilities between $B_i$ and every other box in the space. Ideally, the spectral method finds groups of boxes $B_i$ which are most likely to have transitions within groups and least likely to have transitions between groups. Physically, this is equivalent to finding regions in the atmosphere where trajectories move within the region but are unlikely to move outside of the region. This result may fail if $K > q$ the number of required regions is forced to be larger than the number of true almost invariant sets which is why it is suggested that the eigenvalues of $P_{t, \tau}$, $R_{t, \tau}$, and $L_{t, \tau}$ are considered before choosing this value.

Over the subspace $X = 36^\circ N - 70^\circ N$ and $11^\circ W - 25^\circ E$ representing a portion of the atmosphere over Europe (1,000 meters above sea level), we find two connected almost invariant sets indicated by a single cluster in the spectral clustering outcome. We choose the set corresponding to Europe for importance sampling. See Figure 3 for an illustration of these sets.

**Figure 3.** Two atmospheric eddies as almost invariant sets under the flow $F$ over $[t, t + \tau]$. Regions are indicated by different colors. Center of the chosen target eddy is marked with a red $X$. 
Importance Sampling and Finding Trajectories Likely to End in an Atmospheric Eddy. We initialize a set $\vec{x}_{0,N}$ of $M = 12,000$ uniformly distributed points over the larger domain $30^\circ N - 80^\circ N$ and $20^\circ W - 40^\circ E$. We will refer to this domain as $Z$ and emphasize that $X \subset Z \subset Y$. We refer to Figure 4 for an illustration of these regions. The objective of importance sampling is to enrich the data set of trajectories which are likely to end near the center $\gamma(A)$ of the almost invariant set $A$. By definition, this set has some mixing with the external space. This is illustrated in Figure 5.

For ease of computation, we choose the resampling time $T = 0.1$ so that $t_n = Tn$ for $n = 1, \ldots, \tau/T$ is equivalent to the step-size of the Runge-Kutta approximation of $F_t$; however we may theoretically choose any value $T \geq 0.1$ provided this value is smaller than the Lyapunov time. Moreover, the autocorrelation time of $\phi \circ F_t(\vec{x}(t), t)$ is near zero at $t = 0.1$ which further supports the choice of resampling time $T$. For pragmatic reasons, we add a constant value $D$ to the observable so that $\phi(x) = -\log(d(x, \gamma)) + D \geq 0$. This shift by $D$ ensures that negative values in the exponent of the weight function are the result of a true decrease of the observable value from the previous step.

At each step $n$ of the importance sampling algorithm outlined in Section 4 the weight function

$$W_{N,n} = \frac{\exp(C(\phi \circ \vec{x}_{N,n} - \phi \circ \vec{x}_{N,n-1}))}{Z_n}$$
Figure 5. End position of trajectories under the flow different colors illustrate regions of mixing: (green) starting from the almost invariant set $A$ and ending in the external space $Z/A$, (purple) starting in $A$ and ending in $A$, (red) starting in $Z/A$ and ending in $A$ and (blue) starting in $Z/A$ and ending in $Z/A$. The center of the almost invariant set is marked with an X.

is applied to exponentially tilt the original distribution by $C$ where $F_t(\vec{x}_{N,n-1}, t_{n-1}, t_n) = \vec{x}_n$ is the end position of the $N$th trajectory under the (numerically approximated) flow $F_t$ beginning at $\vec{x}_{N,n-1}$, time $t_{n-1}$ and running until time $t_n$. We choose the tilting value $C = 2.5$ which results in a higher probability of observing trajectories near $\gamma$ (see Figure 1 for an illustration) while higher values of $C$ force almost all trajectories toward the tail and result in a break-down of the distribution. This phenomenon in importance sampling methods is studied in numerical detail in [3]. The algorithm is run over the interval $[t, t+\tau] = [360, 364]$ where the killing and cloning procedure is performed frequently enough to provide a set of sample trajectories that (approximately) follow the exponentially tilted distribution.

Figure 6 shows the end location distribution of surviving trajectories after genealogical particle analysis (GPA). Of course, most surviving trajectories are located near the center of the almost invariant set. We note that surviving trajectories with clearly low probabilities of entering the almost invariant set are due to the distributional aspect of GPA. That is, there is still a probability, though much lower, of obtaining trajectories far from the center.
Remark. Following previous remarks from Section 4, we observe that trajectories that begin near the center $\gamma(A)$ of the almost invariant set can be sent further away from $\gamma(A)$ under the flow but still remain inside the almost invariant set. These trajectories are often killed in importance sampling because the value of the observable $\phi = -\log(d(x, \gamma(A)))$ is too low toward the edges of the almost-invariant set. This is illustrated in Figure 7 (a). Decreasing the tilting value $C$ decreases the number of killed trajectories that start in the almost invariant set; however, this adjustment also decreases (at possibly a different rate) the number of killed trajectories in the general space. In this investigation, we focus on trajectories that end in a reasonably large ball about $\gamma(A)$; however this phenomenon is our motivation for considering some modification of the Hausdorff distance in future work.

Since $\hat{X}_N = \phi \circ \tilde{x}_{N,n}$ is maximized at the center of our invariant set, the set of backward reconstructed trajectories provides a collection of initial points that are most likely to enter and/or remain in the almost invariant set $A$ over the time interval $[t, t + \tau] = [360, 364]$. Figure 7 shows that these initial points form clusters indicating which geographic regions are most likely to have trajectories end in the almost invariant set.

Figure 6. End position of trajectories (blue) under the flow and (red) after genealogical particle analysis. Note that most surviving trajectories end near the center of the almost invariant set (marked with an X).
We may use this set of trajectories to measure the probability of the region \( E \) sending a particle near the center \( \gamma \) of \( A \) under the flow by defining,

\[
p_E(E \rightarrow B(\gamma(A), r)) = \frac{\# \{ x_{0,n} \in E : x_{N,n} \in B(\gamma(A), r) \}}{\# \{ x_{N,n} \in B(\gamma(A), r) \}}
\]

where we take \( E \subseteq \mathbb{Z} \) and \( B(\gamma(A), r) \subseteq \mathbb{Z} \) is the ball centered at \( \gamma(A) \) with radius \( r \). In this way, we can determine the region that has the highest probability of trajectories ending up in \( A \). We can also determine the most probable paths and set of initial conditions a region will take to get to \( A \). Figure 8 shows the set of 10° × 10° boxed regions \( E \) color coded in terms of this probability. Application of this method can be used to determine the regions with the highest probability of pollution exchange with a target eddy.

6. An Application to the Portable University Model of the Atmosphere for Storm Tracking

We have shown in the previous example that the methods outlined in this paper can be used to track the collection of particles (or pollution) in the atmosphere to a fixed almost invariant set that is formed by the background movement of multiple eddies through the space. Now we consider a time-dependent almost-invariant set formed from a single eddy. Storm systems, such as hurricanes, have Lagrangian dynamics similar to that of an almost-invariant set so we can use these naturally occurring atmospheric eddies as a foundational model for storm movement.

We investigate a time-dependent almost invariant set as a function of a shorter time step and use importance sampling to find its most likely path. Each step in the path is determined by a transition probability matrix, built over a small time interval on which the almost invariant set is defined, with states given by a spacial grid. Transitions are taken as the probability of trajectories starting in a region and ending near the center of the almost-invariant set. Estimates of these probabilities are found by using genealogical particle analysis to enrich the set of trajectories likely to end near the almost-invariant set. As a rule, all notation in this section is carried over from the previous example.

We numerically approximate the flow \( F_t \), built from the same northern and eastern velocity field outputs of PUMA, by the Runge-Kutta method described in Section 5. Next, we use the method described previously to find the almost invariant set in the region \( X \subset Y \) over \( J = 3 \) non-overlapping, consecutive time windows of length equal to 1 day, \( [t + J, t + J + \tau] = [360 + J, 360 + J + 1] \). The result is a time-dependent almost invariant set found over three discrete time intervals; one set is found for each time interval. The length of the chosen time intervals is relative to the movement speed of the almost invariant set. Time windows of a shorter length do not show a significant amount of movement of the almost invariant set while time windows of a longer length produce overlapping eddies resulting in almost invariant sets of a different form. Our result is an almost invariant set that moves spatially as a function of discrete time without much change in geometry.

For each \( J \), we run genealogical particle analysis on the set of uniformly distributed particles over \( Z \subset Y \) with starting time \( t+J \) and termination time \( t+J+1 \). Recall that GPA returns a set of trajectories that behave as though they come from the exponentially tilted
distribution where there is a higher likelihood of obtaining larger values of the observable 
\( \phi(x) = -\log d(x, \gamma(A(J))) \) where \( \gamma(A(J)) \) is the midpoint (center) of the \( J \)th corresponding almost invariant set. Hence, the outcome is the set of trajectories most likely to end near \( \gamma(A(J)) \). The resampling time is again taken at \( T = 0.1 \) with sampling times \( t_n = Tn, n = 1, \ldots, [\tau/T] = 10 \). Backwards reconstruction of surviving trajectories is then used to determine the set of initial points which are most likely to end near \( \gamma(A(J)) \).

In this example, the set of possible starting regions is defined after GPA as the set of \( 5^\circ \times 5^\circ \) boxes covering all of \( Z \). For each \( J \), we have an associated region \( E_J \subset Z \) corresponding to the starting region that has the highest proportion of initial points from surviving trajectories (over all starting regions). Since \( E_J \) has the highest probability of sending trajectories near \( \gamma(A(J)) \) at time \( t + J + 1 \), this region should provide us with the movement direction of the almost invariant set \( A(J) \rightarrow A(J + 1) \) (and its corresponding center \( \gamma(A(J)) \rightarrow \gamma(A(J + 1)) \)) defined over \([t + J + 1, t + J + 2]\). Using each of the \( J = 3 \) invariant sets found previously from the PUMA flow approximation, we illustrate in Figure 10 that \( E_J \) can provide some reasonable indication of movement direction for the almost invariant set in the next time step.

Remark. The user-chosen resolution of the grid of regions \( E \) depends on the sparsity of the GPA outcome and the original resolution of the velocity field. This choice of resolution can produce slightly varying results for the predicted direction of movement.

7. Conclusion

Almost invariant sets in the atmosphere are physically interesting because they relate to single eddies and eddy interactions. In this investigation, we look at almost invariant regions of atmospheric flow represented by the numerically approximated flow taken from velocity fields of the Portable University Model of the Atmosphere. We use a modified set of tools revolving around a well-studied transfer operator approach to estimate these regions. In particular, we approximate the Perron-Frobenius operator by the transition probability matrix for the flow over the European subregion and use spectral \( K \)-means clustering to find almost invariant sets only located over Europe.

It can be seen for longer time intervals that the almost invariant sets are formed by multiple eddy interactions such as particles trapped between the spin of two eddies; whereas shorter time intervals have almost invariant sets corresponding to a single eddy. For the former, one can ask questions about the regional origin of trajectories ending inside the almost invariant set and their most probable paths. For the latter, one can ask questions about the path of such an eddy in the space. To study these trajectories, we employ a well-known importance sampling algorithm, called genealogical particle analysis, not used in this context to-date.

Current literature has focused on using importance sampling methods to decrease the relative error of an estimated rare event probability by forcing rare events to occur more frequently. We show that these methods can also provide useful information on the set of trajectories likely to end in an extreme event. For the interest of this study, we have introduced an observable that defines the extreme event as being near the center of an
almost invariant set. In this setting, we show that the surviving trajectories obtained from importance sampling can provide information on probable paths and initial regions of trajectories that end in an almost invariant set under an atmospheric flow. We complete our investigation by motivating and illustrating some important examples where information about trajectory movement toward the center of an almost invariant set in the atmosphere is useful and physically relevant: storm movement and origin of pollution.

In future work we plan to apply these techniques to real hurricane data where the fixed point $\gamma$ may be taken as some point outside of the time-dependent almost-invariant set. The outcome of importance sampling would then give us the probability of a hurricane moving over a given region. It would also be interesting to consider importance sampling methods for the sequence of maxima $M_n = \max\{X_1, \ldots, X_N\}$ where $X_N = \phi \circ f(x_N)$. This would limit the set of original distributions to the family of generalized extreme value functions and possibly provide a new way of using the Hausdorff distance in the definition of $\phi(x) = -\log(d(x, \gamma))$. Furthermore, a complete shift of the generalized extreme value distribution under exponential tilting would result in a higher density around $\gamma$ and less uniformly distributed points about the whole space.
Figure 7. (a) Parent initial positions of surviving trajectories after genealogical particle analysis. Note the cluster in the southwest region of the space. This distribution is used to determine the regions with the highest likelihood of ending near the center $\gamma(A)$ of the almost invariant set $A$ (marked with an X). (b) Initial positions of surviving trajectories ending in some ball $B(\gamma(A), r)$ of radius $10^\circ$ and not beginning in the almost invariant set. (c) Uniformly sampled initial positions (red) of (b) with full trajectories (blue).
Figure 8. Surviving trajectories after genealogical particle analysis (blue). The set of all surviving trajectories \( x_{N,n} \in B(\gamma(A), r) \) \textbf{and} \( x_{N,0} \notin A \) with \( r = 10^\circ \) are marked in red and their parent initial trajectories \( x_{0,n} \) (black). Grid boxes are \( 10^\circ \times 10^\circ \) with darker orange indicating a higher \( p(E \rightarrow B(\gamma(A), r)) \) value. The regions with the highest probability of ending near \( \gamma(A) \) are \( 10^\circ E - 10^\circ W \) and \( 30^\circ N - 40^\circ N \).
Figure 9. Almost-invariant sets found from the transfer operator method of $J = 3$ non-overlapping, consecutive time windows of length 1 day. Time intervals $[t + J, t + J + \tau] =$ (a) $[360, 361]$ (b) $[361, 362]$ (c) $[362, 363]$. Quivers indicate the velocity field.
Figure 10. Movement prediction of the almost invariant set. The center for step $J$ is marked with a red X and the corresponding initial positions of the end surviving trajectories within $B(\gamma(A(J)), r)$ are indicated by red points. The $J+1$ almost invariant set is highlighted in gray, its corresponding velocity field is represented by black quivers and center is marked with a black X. (a) $J = 1$ and (b) $J = 2$. The movement direction probability is taken as the proportion of (surviving) initial positions in the regions marked by the grid. These probabilities are highlighted in orange with darker values indicating a higher probability.
APPENDIX A. K-MEANS

The standard $K$-means algorithm for a set of $m$ nodes $n$ represented by $m$ vectors in $\mathbb{R}^H$ is given by,

Step 1 Start with $K$ random partitions $P_j$ of the space $\mathbb{R}^H$.

Step 2 Compute the centroids (means) of these partitions as
$$C_j = \sum_{n(\ell) \in P_j} n(\ell) / \text{card}(P_j)$$
where $C_j \in \mathbb{R}^H$.

Step 3 Assign $n(\ell)$ to the partition $P_j$ with the minimum (squared) euclidean distance between $n(\ell)$ and $C_j$.

Step 4 Update the algorithm by recalculating the centroids (means) of $P_j$.

The algorithm continues by repeating steps 3 and 4 until the assignments no longer change. This is equivalent to finding the steady state of the objective function given by,

$$\min_{j=1}^{K} \sum_{\ell=1}^{m} ||n(\ell) - C(j)||_{\mathbb{R}^H},$$

the minimum sum of the (squared) euclidean distances between each node and its assigned centroid.
References

[1] Carney, M., Azencott, R., Nicol, M.: Nonstationarity of summer temperature extremes in Texas, Int J Climatol. 40 620-640, 2020.
[2] Carney, M., Kantz, H.: Robust regional clustering and modeling of nonstationary summer temperature extremes across Germany, Advances in Statistical Climatology Meteorology and Oceanography. 6 61-77, 2020.
[3] Carney, M., Kantz, H., Nicol, M.: Analysis and Simulation of Extremes and Rare Events in Complex Systems arXiv: 2005.07573. preprint.
[4] Dellnitz, M., Junge, O.: On the Approximation of Complicated Dynamical Behavior, JSTOR. 36 (2) 491-515, 1999.
[5] Froyland, G.: Statistically Optimal Almost-Invariant Sets, Physica D. 200 (3-4) 205-219, 2005.
[6] Froyland, G., Dellnitz, M.: Detecting and Locating Near-Optimal Almost-Invariant Sets and Cycles, SIAM J. Sci. Comp. 24 (6) 1839-1863, 2003.
[7] Froyland, G., Padberg-Gehle, K.: Almost-Invariant and Finite-Time Coherent Sets: Directionality, Duration, and Diffusion, Bahsoun W., Bose C., Froyland G. (eds) Ergodic Theory, Open Dynamics, and Coherent Structures. Springer Proceedings in Mathematics and Statistics, Springer. 70, 2014.
[8] Froyland, G., Schwalb, M., Padberg, K., Dellnitz, M.: A transfer operator based numerical investigation of coherent structures in three-dimensional Southern ocean circulation, Proceedings of the International Symposium on Nonlinear Theory and its Applications, 2008.
[9] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction 2nd Edition, Springer, 2008.
[10] Holland, M., Nicol, M., Török, A.: Almost sure convergence of maxima for chaotic dynamical systems, Stochastic Processes and their Applications 10, 126, 3145-3170, 2016.
[11] Junge, O., Marsden, J., Mezic, I.: Uncertainty in the Dynamics of Conservative Maps, 2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601). 2 2225-2230, 2004.
[12] Lucarini, V., Faranda, D., Freitas, A.C.M, Freitas, J.M., Kuna, T., Holland, M., Nicol, M., Todd, M., Vaienti, S.: Extremes and Recurrence in Dynamical Systems, Wiley 2016.
[13] Lunkeit, F, Blessing, S., Fraedrich, K., Jansen, H., Kirk, E., Luksch, U., Sielmann, F.: Planetsimulator users guide version 15.0, Meteorological Institute of the University of Hamburg, Hamburg, 2007.
[14] Lünsmann, B., Kantz, H.: An extended transfer operator approach to identify separatrices in open flows, Chaos. 26 053101, 2018.
[15] Ragone, F., Wouters, J., Bouchet, F.: Computation of extreme heat waves in climate models using a large deviation algorithm, Proc. Natl. Acad. Sci. USA, 1 (115) 24-29, 2018.
[16] Vallero, D.: Fundamentals of Air Pollution 5th ed., Elsevier, 2014.
[17] von Luxburg, U.: A Tutorial on Spectral Clustering, Statistics and Computing. 17 (4), 2007.
[18] Wouters, J., Bouchet, F.: Rare event computation in deterministic chaotic systems using genealogical particle analysis, J. Phys. A., 49 (37), 374002, 2016.