Decomposing Antarctic Sub-shelf Melt Variability using Generalized Clustering with Kernel Embeddings

Jacquelyn A. Shelton[1], Przemyslaw Polewski[4], Alexander Robel[2], Matthew Hoffman[3], Stephen Price[3]
[1]Hong Kong Polytechnic University [2]Georgia Institute of Technology [3]Los Alamos National Laboratory [4]TomTom North America Inc.

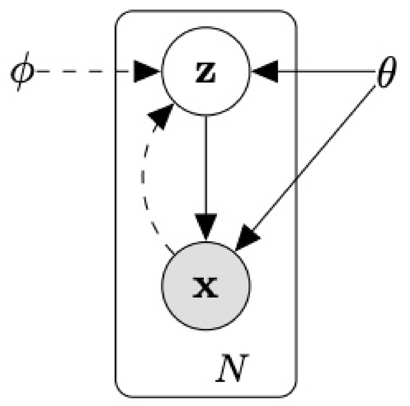
Introduction

- Antarctic Ice Sheet ice loss – ice-shelf basal melt flux main contributor to global sea level rise
- Predictions/uncertainties require large ensembles of realizations from earth system models → computationally costly
- To mitigate this bottleneck we can *learn* the variability of the stationary component of ice melt dynamics, and *generate* new time-series (realizations) using machine learning methods
- But, underlying ice melt dynamics are complex and multimodal → crucial to decompose ice melt variability into homogeneous sub-components that can be modeled independently

Step 1: Nonlinear Dimensionality Reduction

Variational Autoencoders [3]:

- **Encode** time series: embed original full-length time series into low-dimensional, disentangled latent space



Assumption: observed data x generated in 2 step process: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$, conditioned on *latent variables* z. As inference on the marginal $p_\theta(x)$ and/or true posterior $p_\theta(z|x)$ is often intractable, the VAE uses a *variational approximation* $q_\phi(z|x)$ for $p_\theta(z|x)$, and learns parameters θ, ϕ jointly by optimizing the lower bound on $p_\theta(x)$:

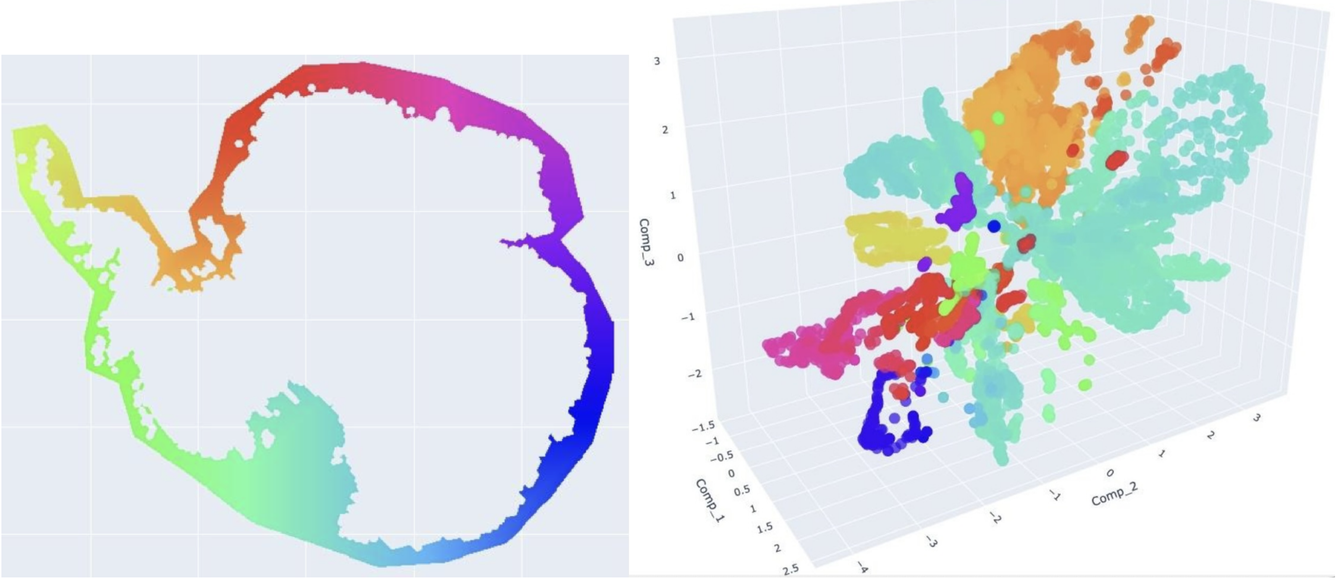
$$p_\theta(x) \geq \mathcal{L}(\theta, \phi; x) = -D_{KL}(q_\phi(z|x)||p(z)) + E_{q_\phi(z|x)}[\log p_\theta(x|z)]$$

Promote disentanglement betweenlatent dimensions, a Gaussian prior with a diagonal covariance structure is chosen:

$$\log q_\phi(z|x) = \log \mathcal{N}(z; \mu, \sigma^2 I)$$

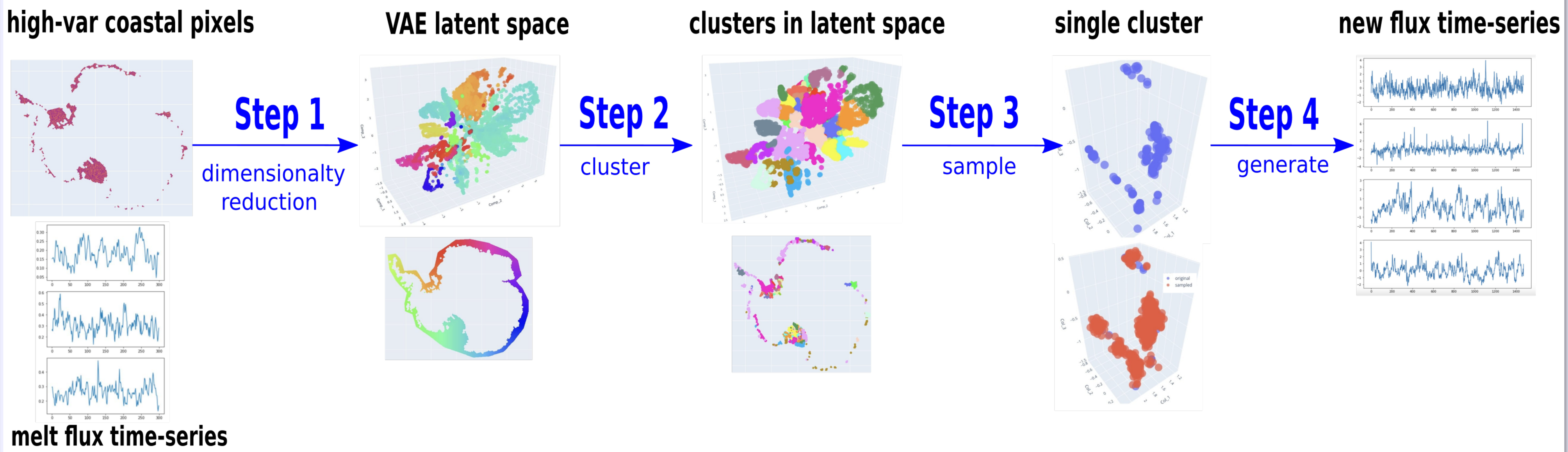
where μ, σ^2 are functions implemented via Convolutional Neural Networks (CNNs)

3-dimensional latent variable embedding of ca. 1660 timebin channels of ice flux timeseries:



Coastal distance colors   Latent space structure

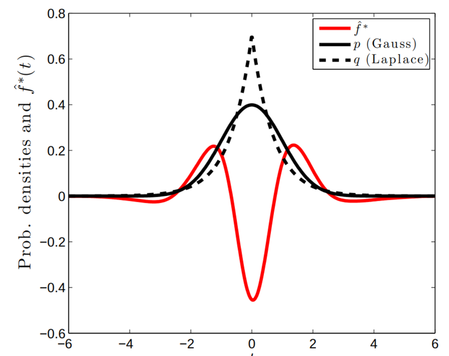
Method: Pipeline Components



Step 2: Clustering

Recursive clustering of latent space:

- Partition the low-dimensional latent space into regions with same dynamic behavior → generalized statistical clustering approach [4] based on Maximum Mean Discrepancy measure (MMD)
- Consider two distributions P_1, P_2 on latent space Z, and kernel function $k: Z \times Z \to R$ using associated reproducing kernel Hilbert space (RKHS) H:



$$MMD(P_1, P_2) = ||\mu(P_1) - \mu(P_2)||_H$$

Function maximizing the mean discrepancy between 2 distributions: Gaussian and Laplace w/ same mean and variance (zero mean & unit variance)

for two-cluster problem: $\alpha^i \in [0, 1]$ is assignment of data point i to cluster 1

$\hat{\pi}_1, \hat{\pi}_2$ proportion of points to clusters $1, 2$

→ Compute clusters by maximizing criterion:

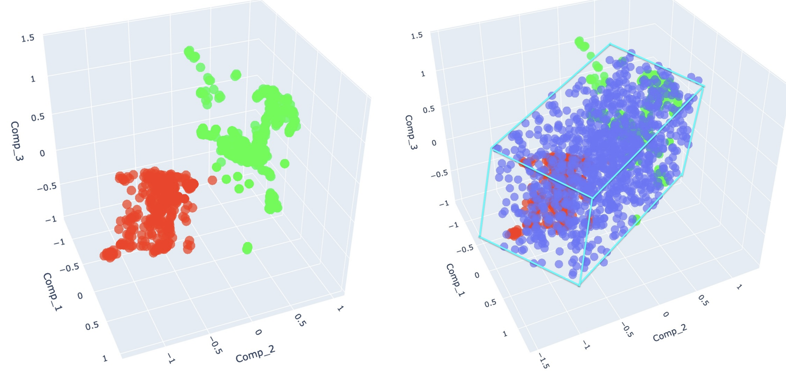
$$\max_{\alpha \hat{\pi}_1 \hat{\pi}_2} MMD(\hat{P}_1, \hat{P}_2) = \max_{\alpha} const - \sum_{k=1}^{2} \sum_{i=1}^{n} ||\phi_i - \mu[\hat{P}_k]||_H^2$$

- Cluster # unknown in advance – **Partition recursively** until stopping criterion met → Each iteration: choose best partition of subclusters $k \in \{2, 3, 4, 5\}$ given data subset $Y \subseteq Z$ using the **gap statistic** [5]:

$$g(k) = \log \frac{MSE_{Q*}(k)}{MSE_{Q*}(1)} - \log \frac{MSE_Q(k)}{MSE_Q(1)}$$

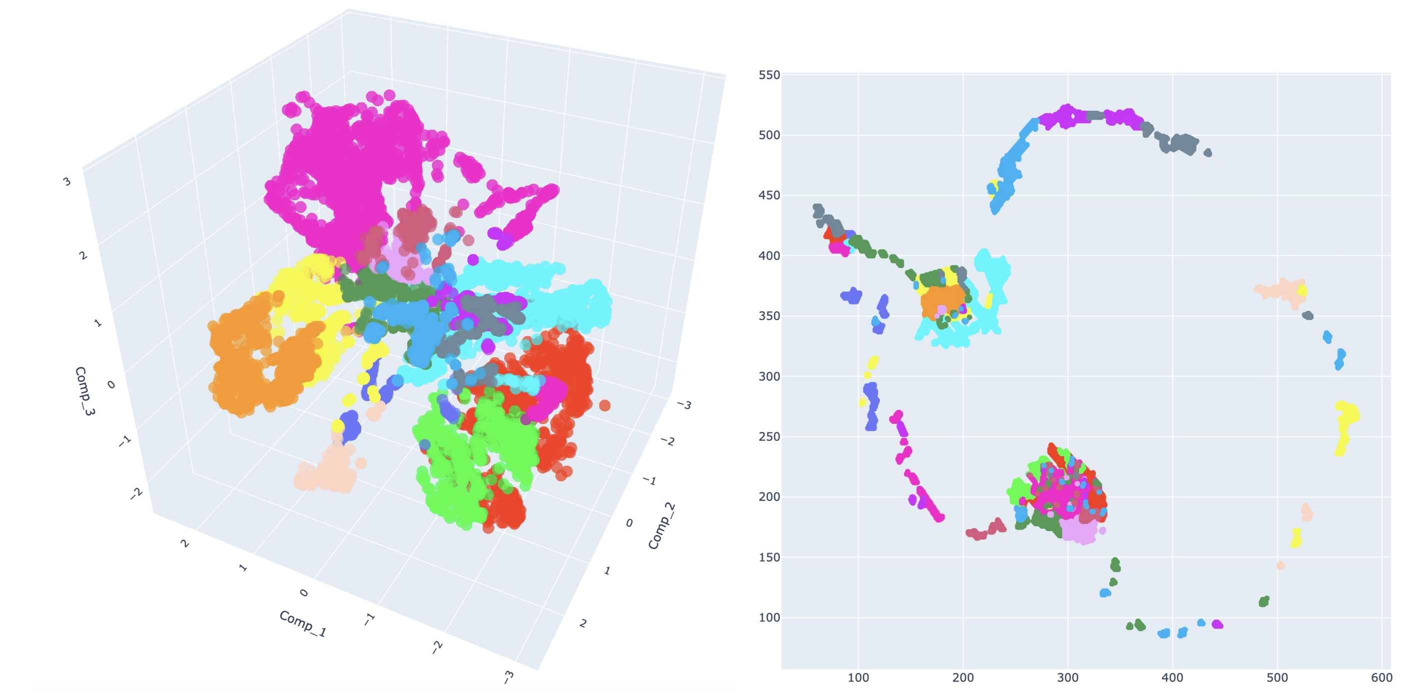
$MSE_Q(k) \equiv \min_\alpha \sum_i ||\phi_i - \mu[\hat{P}_\alpha]||_H^2$: distance from each point to its closest cluster 'center' in the kernel's RKHS

$Q*$: reference data distribution, uniformly sampled from oriented bounding box of Q:



Proposed clustering   Reference distribution

Step 2: Cluster results



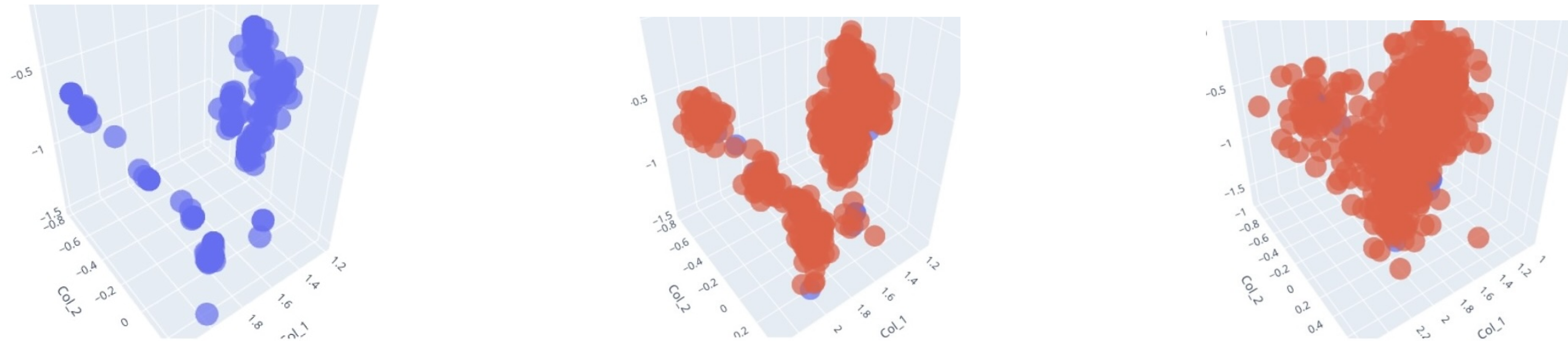
latent space   geographic space

Step 3: Sampling

KDE - Nonparametric sampling from clusters:

- For each discovered cluster C_i in the latent space, construct a kernel density estimator and sample from it to obtain desired number of new timeseries

$$\hat{f}_i(z; B) = \frac{1}{|C_i|} \sum_{q \in C_i} K_B(z - Z_{q_i})$$

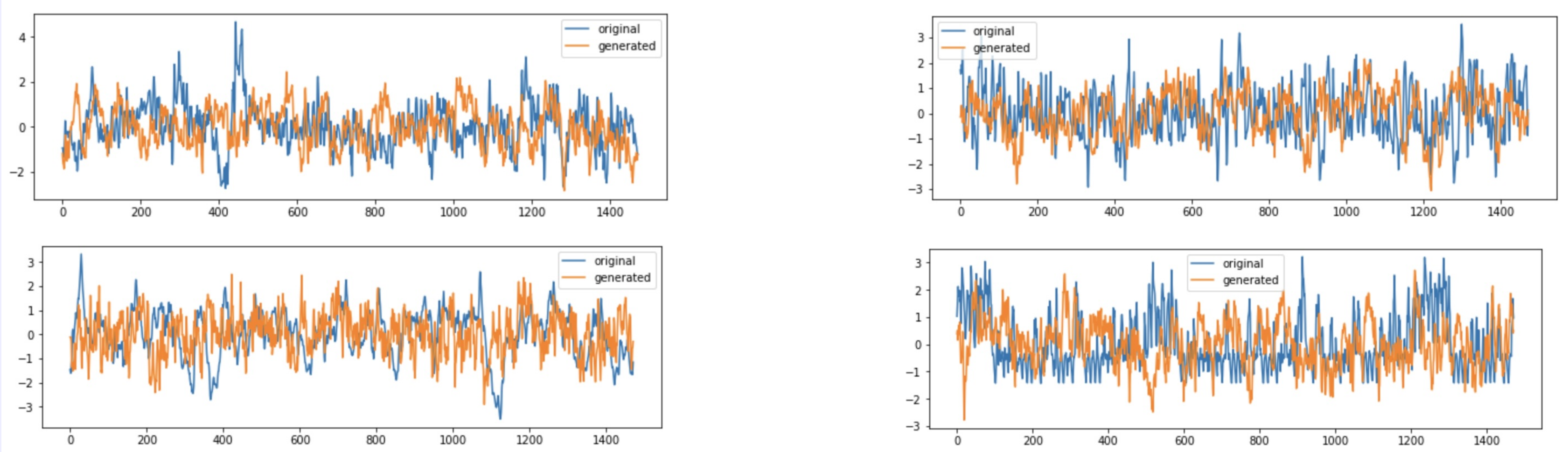


original points   sampled – low κ sampled – high κ

- Plug-in estimator [6] of bandwidth B with factor κ for greater variation
- Choose max. bandwidth factor κ allowed by 2-sample MMD test [7]

Step 4: Generation

- **Decode** latent space back to time-series (original) space and generated time-series to original spatial location belonging to cluster
- New time series retains statistical and temporal properties



References

[1] Robel, A., Seroussi, H., Roe, G. Marine ice sheet instability amplifies and skews uncertainty in projections of future sea-level rise. In PNAS, 116(30). 2019.
[2] Hoffman, M., Price, S. The DOE E3SM v1.2 Cryosphere Configuration: Description and Simulated Antarctic Ice-Shelf Basal Melting. In J of Advances in Modeling Earth Systems. 2022.
[3] Kingma, D. P., Welling, M. Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations. In ICLR 2014, Banff, AB, Canada, April 14-16, 2014.
[4] Jegelka, S., Gretton, A., Schölkopf, B., Sriperumbudur, B.K., von Luxburg, U. Generalized Clustering via Kernel Embeddings. In: Advances in Artificial Intelligence. KI, 2009.
[5] Tibshirani, R., Walther, G., Hastie, T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic, Journal of the Royal Statistical Society, Volume 63, Issue 2, 2001.
[6] Wand, M. P., Jones, C. Multivariate plug-in bandwidth selection. Computational Statistics, 9(2) pp. 97–116, 1994.
[7] Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., & Gretton, A. MMD Aggregated Two-Sample Test. ArXiv, abs/2110.15073. 2021.