AGU 2022 Fall Conference – C52C-0372

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Los Alamos NATIONAL LABORATORY

Generate Antarctic sub-shelf melt using recurrent neural network-based Generative Adversarial Models on pixel clusters

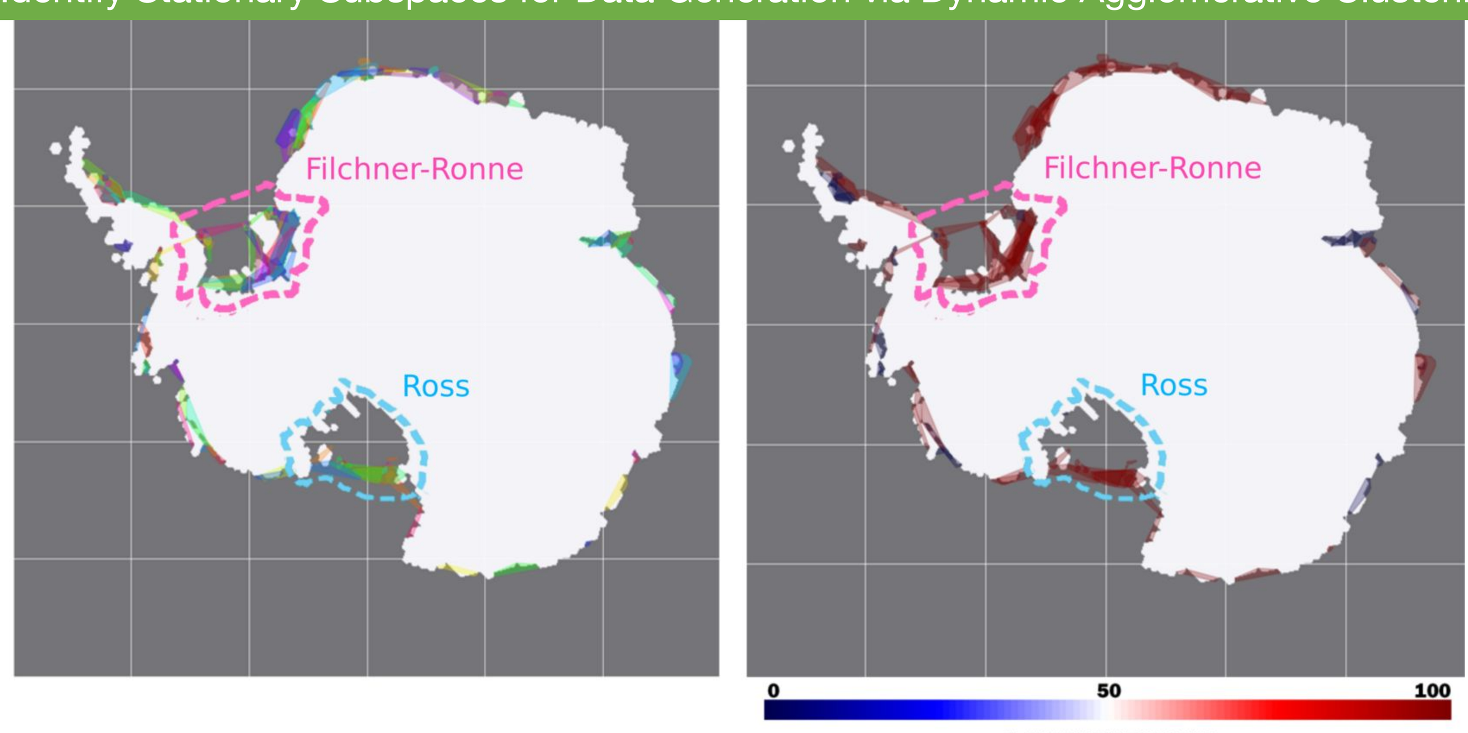
Jacquelyn A. Shelton[1], Alexander Robel[2], Matthew Hoffman[3], and Stephen Price[3]
[1]Hong Kong Polytechnic University [2]Georgia Institute of Technology [3]Los Alamos National Laboratory

Georgia Tech

E3SM Energy Exascale Earth System Model

- Antarctic Ice Sheet ice loss accelerated by surrounding ocean's extreme warming over last 30 years → dominant contributor to global sea level rise
- **Questions:**
  → How much ice loss due to anthropogenic changes and to **internal variability** [1]?
  → Does internal climate variability introduce significant uncertainty into projections of the Antarctic contribution to future sea level rise?
- **Goal:** using limited model output – *one realization* of 150-year simulation of pre-industrial variability of sub-shelf melt rates from *expensive* state-of-the-art Earth System model (E3SM) [2] – develop/apply machine learning methods to **generate additional realizations**
- **Previous work:** identified stationary subspaces of input data → realistic (physically consistent) and representative of complex spatiotemporal dynamics [3]
- **Results:** TimeGAN can generate realizations of basal melt rates preserving the **temporal dynamics** and **stationarity**
- **Evaluation metrics:** quality of synthetic vs emulated data → PCA [5], t-SNE [6], KPSS [7]

Approach: Identify Stationary Subspaces for Data Generation via Dynamic Agglomerative Clustering [3]



Prerequisites prior to data generation: identify individual subspaces in data that are:

1. *representative* of its spatiotemporal dynamics,
2. *realistic* in terms of consistency with physically observed dynamics, and
3. *stationary* over the entire time-series, regardless of the behavior of the data within each subspace relative to any other's (may vary independently over arbitrary time-scales)

Idea: construct dynamic hierarchical clustering pipeline to adaptively learn stationary subspaces, the number of which can grow or shrink in a data-driven fashion according to the data dynamics while simultaneously incorporating relevant prior domain knowledge (e.g. physical observations, problem setting). →breaks data into **sequence of temporal problems**

Notation: Let $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ be 2 pixel locations given by their 2D coordinates on the Antarctic Ice Sheet. Note that each pixel spans *10* km. Ice melt flux time-series at these locations: $F_1 = (f_1^1, \dots, f_1^M)$ and $F_2 = (f_2^1, \dots, f_2^M)$, where *M* denotes the number of time-steps in the simulation, which for this data is a monthly resolution over *150* years, for *M = 1800* time-steps. Normalized versions denoted: $\widehat{F_k} = \frac{F_k}{max\, F_k}$ for $k \in \{1,2\}$, and the spatial-distance threshold denoted: d_{thr}^s.

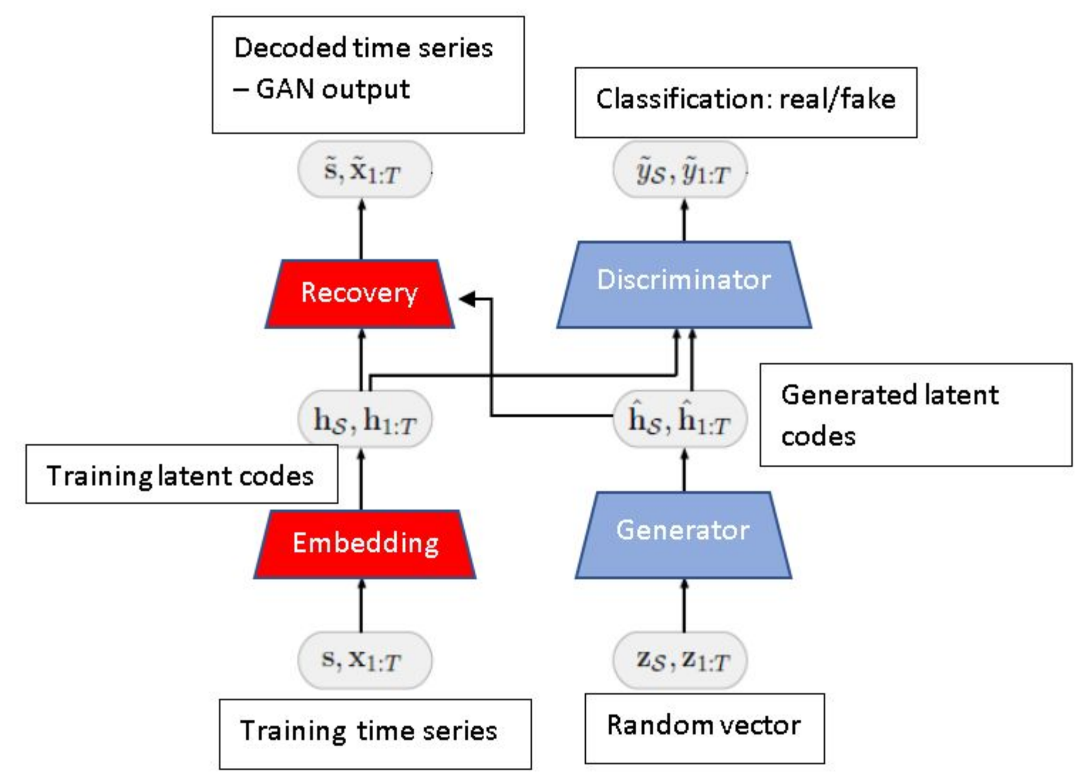
Aggregate Spatiotemporal Distance Criterion

$$d_{s,t}(p_1, p_2) = \begin{cases} \sum_{k=1}^{M} |\hat{F}_1(k) - \hat{F}_2(k)| & \text{if} \quad ||p_1 - p_2|| \leq d_{thr}^s \\ +\infty & \text{else} \end{cases}$$

Stationarity: evaluate stationarity of each identified cluster using Kwiatkowski-Philips-Schmidt-Shin (KPSS) hypothesis test with null hypothesis that the cluster's time-series is stationary around the mean → cluster deemed stationary if majority of pixels pass the KPSS test

Method: Time-series Generative Adversarial Network (TimeGAN) [4]

TimeGan Network Architecture

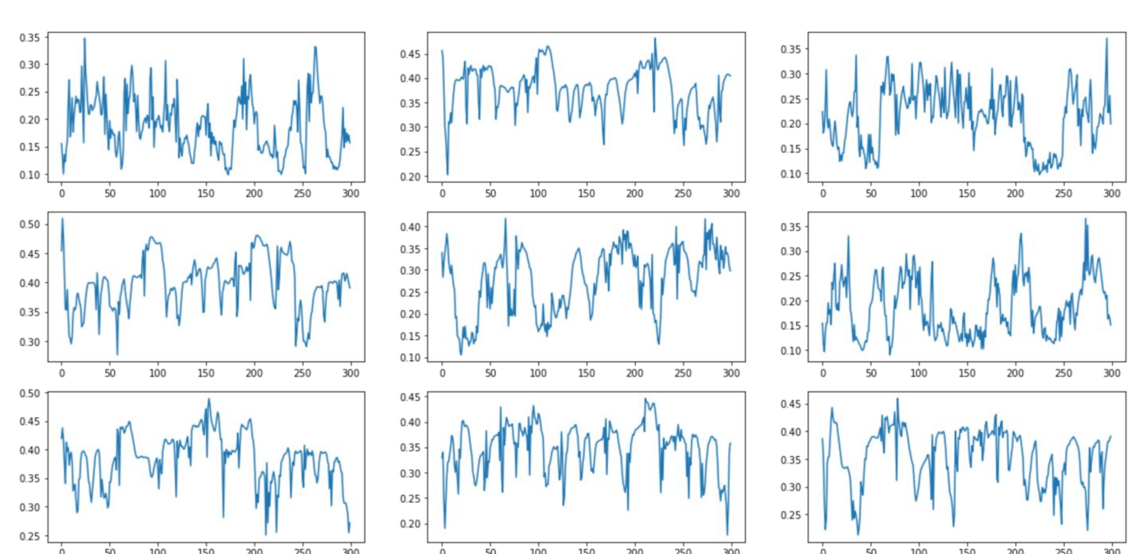


- **TimeGAN:** learns good generative model for time-series data that preserves temporal dynamics → new sequences respect original relationships between variables across time [4]
- Architecture: consists of four unique Recurrent Neural Networks (RNNs; e.g. LSTM, GRU): *embedder, generator, discriminator* and *recovery* RNNs

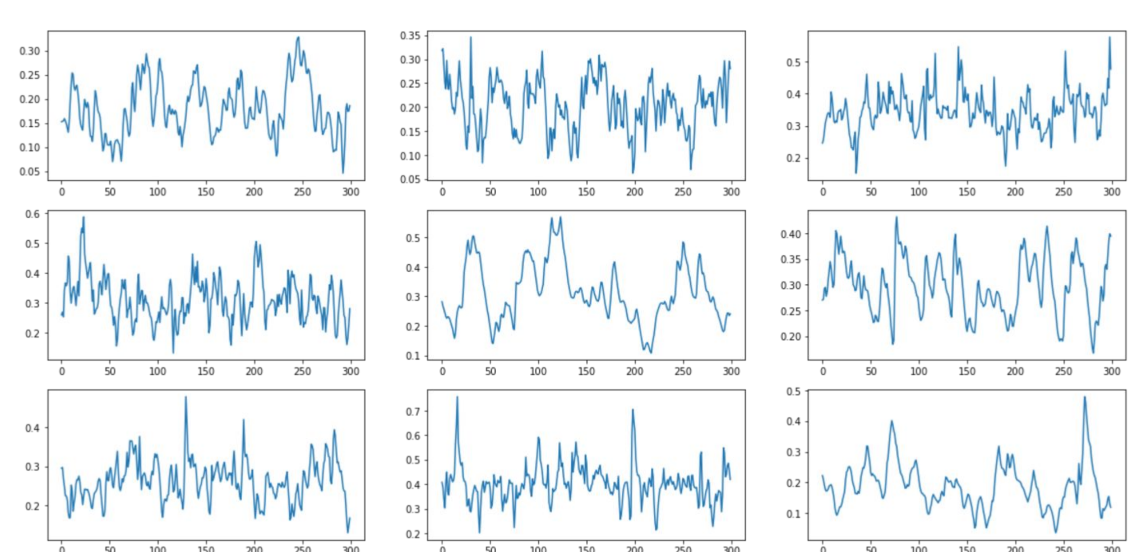
Data Generation: Examples of Original E3SM vs TimeGAN Generated Realizations

Cluster 1 – time-series data of T = 300 (~25 yrs)

Generated data examples

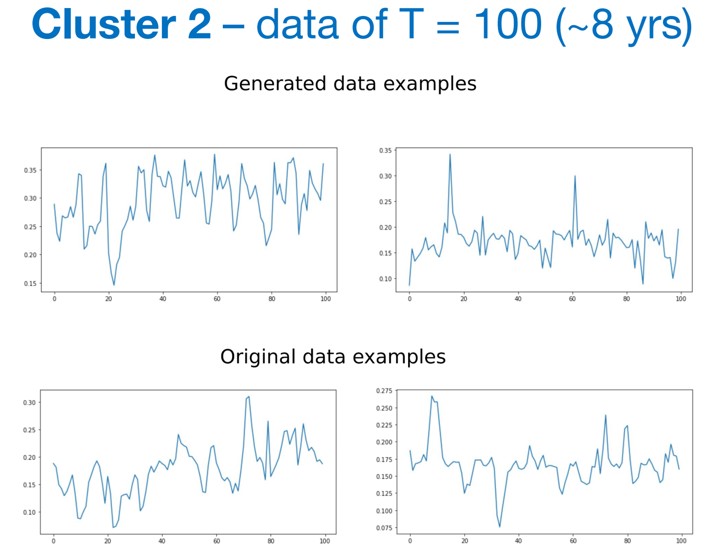


Original data examples



- **Train TimeGAN** on individual stationary cluster's respective 'real' time-series data
- **Each cluster** → treated as its own independent data distribution for GAN to train generative model
- **Generate data** from each cluster's generative model

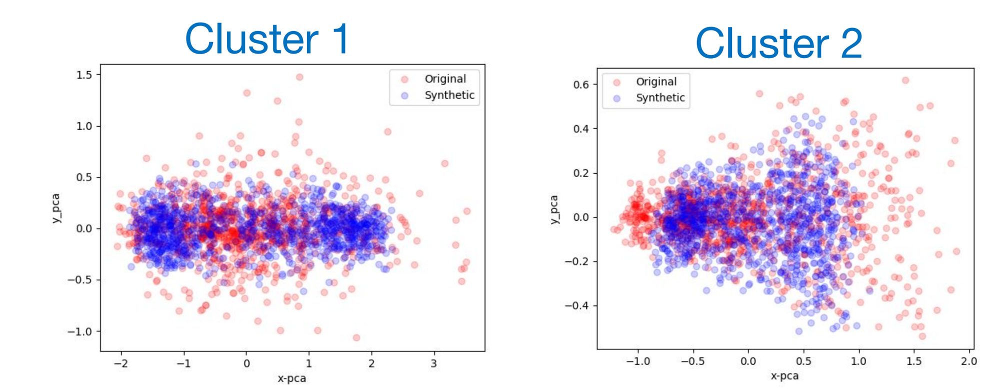
Cluster 2 – data of T = 100 (~8 yrs)



Evaluation: How *'good'* are the TimeGAN generated realizations?

Metrics to quantify distributions' consistency & similarity→ **generated data** vs. **input data**:

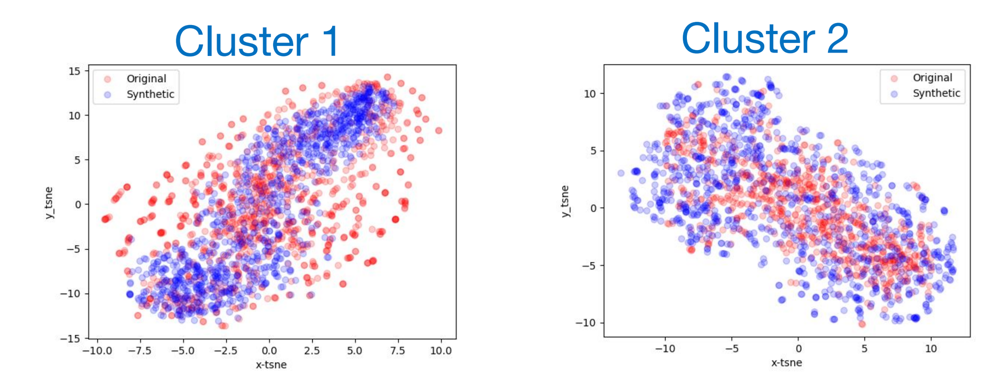
1. Principal Components Analysis (PCA) [5]



→ Flatten temporal dimension s.t. we can visualize data in 2D space: project all data onto the first 2 PCs of original data

→ **Shows:** significant overlap of and spread of input/generated data pts of both clusters 1 and 2

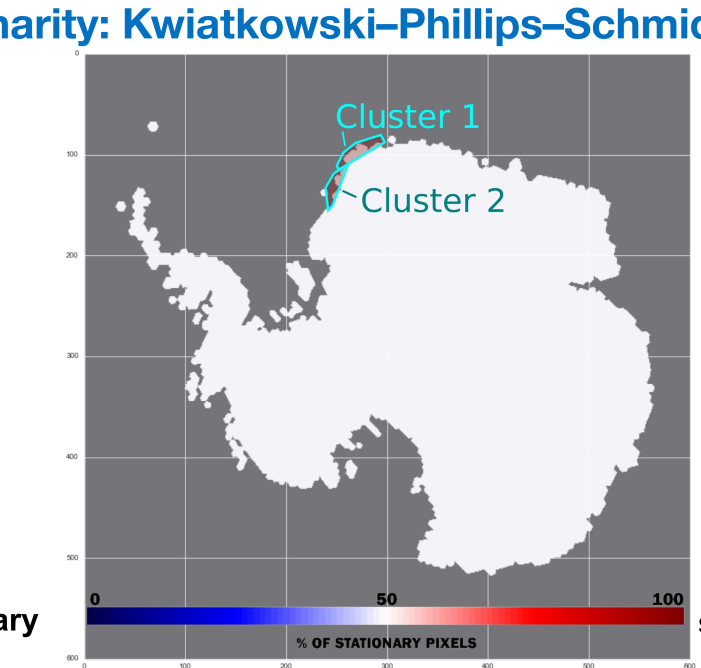
2. t-Distributed Stochastic Neighbor Embedding (t-SNE) [6]



→ **Method** to quantify/visualize similarity of data – capable of retaining local structure of (high dimensional) data and revealing important global structure

→ **Shows:** significant data overlap in both clusters generated & real data

3. Stationarity: Kwiatkowski–Phillips–Schmidt–Shin (KPSS) hypothesis test [7]



→ **Shows:** GAN preserves stationarity – like the real (training) data, both clusters' generated data pass the KPSS stationarity test

Conclusions, Impact, and Outlook

- **Results** show TimeGAN can generate realizations of variable Antarctic sub-shelf melt that preserves the **temporal dynamics** and **stationarity**
- **Evaluation summary:** all 3 metrics show that spatial clustering + TimeGAN can generate data similar to input data – preserving temporal dynamics and stationarity → **PCA** and **t-SNE**: data have similar temporal dynamics in a lower dimensional space, **KPSS** shows generated data retains the input data's stationarity
- This work addresses the general pervasive problem of **data scarcity** in the climate sciences → far more computationally efficient than running climate model
- **Spoiler:** Current work includes further, advanced quality metrics & incorporating advanced discriminator functions for built-in domain-agnostic non-parametric high-dim distribution comparisons [e.g. 8, 9]

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] Shelton, J. A., Robel, A. A., Hoffman, M., and Price, S.: Towards generating stationary realizations of simulated Antarctic ice shelf melt rates from limited model output. Climate Informatics, 2022. *(additionally, expanded preprint available soon)*
[4] Yoon, J., Jarrett, D., van der Schaar, M. Time-series Generative Adversarial Networks. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS). 2019.
[5] H. Hotelling. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24:417–441, 1933.
[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008.
[7] Kwiatkowski, D., Phillips, P. C., Schmidt, P., Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? In Journal of Econometrics, 54(1-3), 159-178. 1992.
[8] Cristóbal Esteban, Stephanie L. Hyland, G.: RätschReal-valued (Medical) Time Series Generation with Recurrent Conditional GANs. In arXiv:1706.02633v2. 2017.
[9] Binkowski, M., Sutherland, D., Arbel, M., Gretton, A. Demystifying MMD GANs. In Proceedings of the ICLR 2018 Conference Blind Submission, 2018.