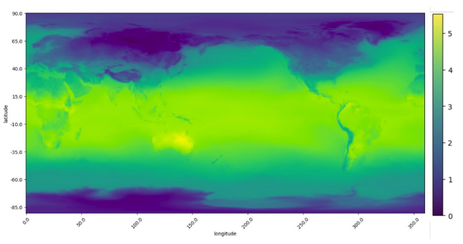
Generating new realizations of large-scale climate ensembles with conditional variational autoencoders

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.

Problem setting / Abstract

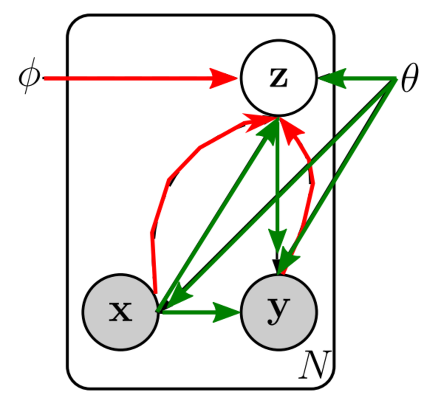
- Climate Model simulations are expensive – how can we get the most from the realizations available?
- Consider large ensembles of smaller independent simulations and utilize shared information across realizations
- Standard off-the-shelf machine learning methods cannot represent multiple independent realizations well
- Plan: Develop customized, flexible deep generative model approach to - capture internal variability in low-dimensional latent spaces with low reconstruction error - represent complex spatiotemporal data and generate samples from their distributions - help reduce the cost of obtaining new realizations from large-scale Earth system models
- ERA5 model [1] output: 10 independent realizations of monthly reanalysis for mean surface temperature from 1940–present



Deep Conditonal Generative Models

Conditional Variational Autoencoders [2]:

- **Encode time series**: embed original full-length time series into low-dimensional, *disentangled* latent space
- Geography should influence time series embedding: similar geographic coordinates⇒similar latent coordinates - aids visualization/interpretability - uses available info to enrich encoder
- 3 types of variables: input vars x (geo location), output vars y (observed time series), and latent vars z (latent coordinates)
- The *conditional* generative process of the model: for given observation y, z is drawn from the prior distribution $p_\theta(z|x)$, and the output y is generated from the distribution $p_\theta(y|x, z)$



Condition latent embedding of a time series y on geographic and latent coordinates, x and z
- generative params θ and variational params ϕ
- green arrows = generative process of y
- red arrows = approximate inference of z

Optimize parameters θ, ϕ jointly: *variational approximation* to the posterior, $q_\phi(z|y\ x)$ for $p_\theta(z|y)$, by minimizing the *ELBO*:

$$\log p_\theta(y|x) \geq \mathcal{L}_{CVAE}(x, y; \theta, \phi)$$
$$= -KL(q_\phi(z|x, y)||p(z|x)) + \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(y|x, z)]$$

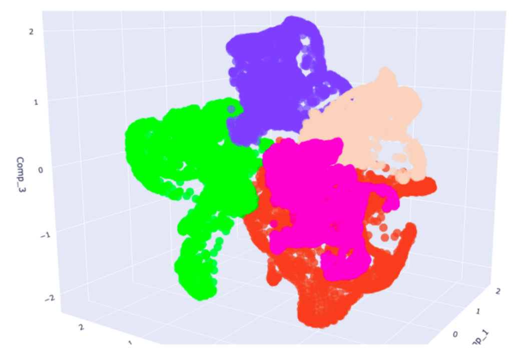
with variational approximate posterior of z

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

→ KL-divergence acts as a regularizer, expectation as reconstruction error; mean μ and s.d. σ learned by eg CNN (are nonlinear functions of datapoint y^i and variational params ϕ

Deep Conditional Generative Modelling Workflow

CVAE trained on ensemble of *all* 10 realizations *simultaneously* into 3D



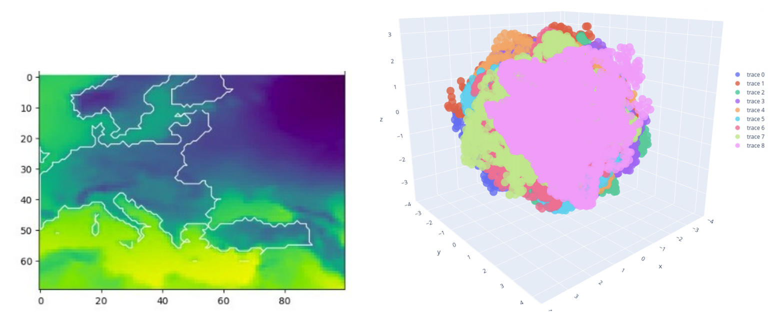
Latent space fragmented – no discernible shape or correspondence between realizations, each occupying own subspace within CVAE latent space, regardless of known geographic space correspondence

Vanilla CVAE cannot represent time series from an unseen realization properly
⟹ fragmented embedding cannot reconstruct or generate new sample

Idea: predict new realizations from a *small* sample of *new* data, transferring relationships learned from (training on) *other* realizations

Promote homogeneous structure of latent space across realizations

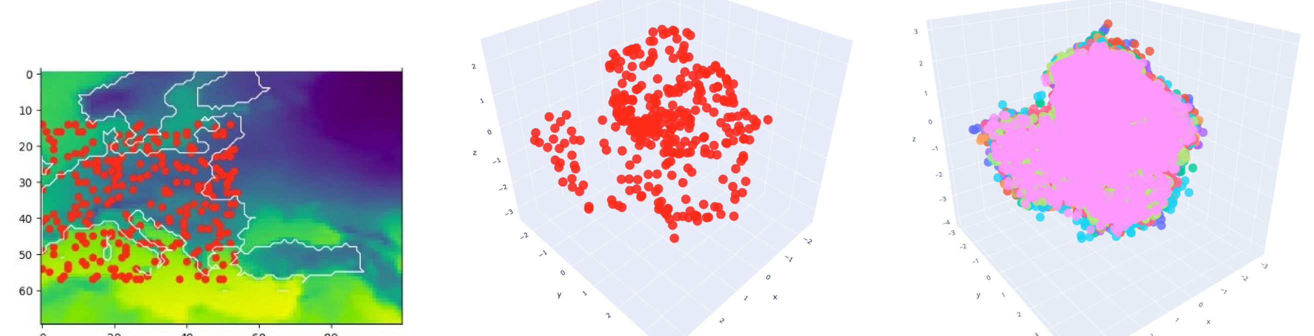
Train a CVAE for each realizations ***separately***



- Intuition: points near each other in latent space have similar temporal behavior
- **Latent-Constrained Conditional VAE**: add cross-realization latent homogeneity constraint, optimize new objective:

$$\mathcal{L}_{LC-CVAE}(x, y; \theta, \phi) = -KL(q_\phi(z|x, y)||p(z|x)) + \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(y|x, z)]$$
$$-\lambda^T \mathbb{E}\, \rho_{Fp}(x, y)\{||q_\phi(z|x, y) - z^f_{x,y}||^2 - D^2_{z,max}\}$$

for constraints on max. distance $D_{z,max}$ of latent encodings $q_\phi(z|x, y)$ at small sample of geographic locations x to fixed points $z^f_{x,y}$ in latent space
⟹ establish common structure across realizations



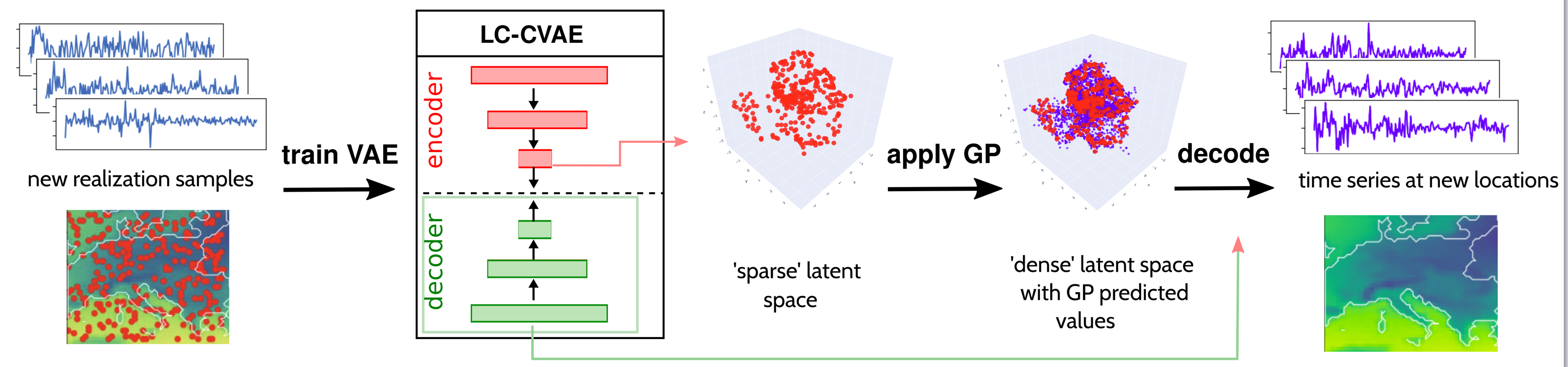
sampled x coordinates | fixed points $z^f_{x,y}$ | constrained embeddings

Predict geographic location's coordinates in latent space

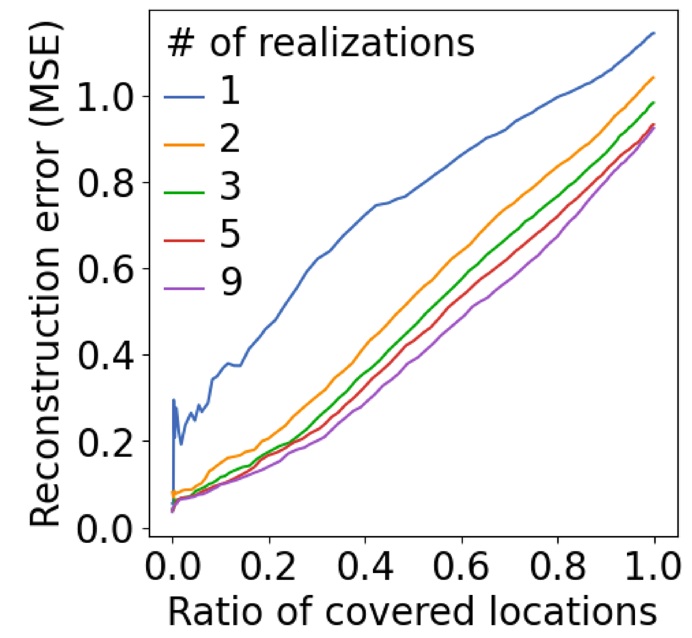
Multi-output Gaussian Process Regression [3]

- Flexible nonparametric model learning a function that maps from the observed data (fixed points' latent coords) to a property of the latent variables (new point's latent coords), e.g. $f: \mathbb{R}^D \to \mathbb{R}^{N_L}$ for $(F^{r_1}(x_1), q^{r_1}_\phi(z|x_1, y_1)), \ldots, (F^{r_P}(x_P), q^{r_P}_\phi(z|x_P, y_P))$
- Training data: features $F^{r_i}(x_i)$ (concatenated latent coords of point x_i's k-nearest-neighbors, in realization r_i), and regression target (true latent coords of x_i), $q^{r_i}_\phi(z|x_i, y_i)$)
- Model: *each* latent coord l is approximated by a Gaussian process $g_l \sim \mathcal{GP}(0, k_l)$, where $k_l(\cdot, \cdot)$ is the covariance kernel, parameterized to represent (nonlinear) relationships between variables – trained via sparse variational inference

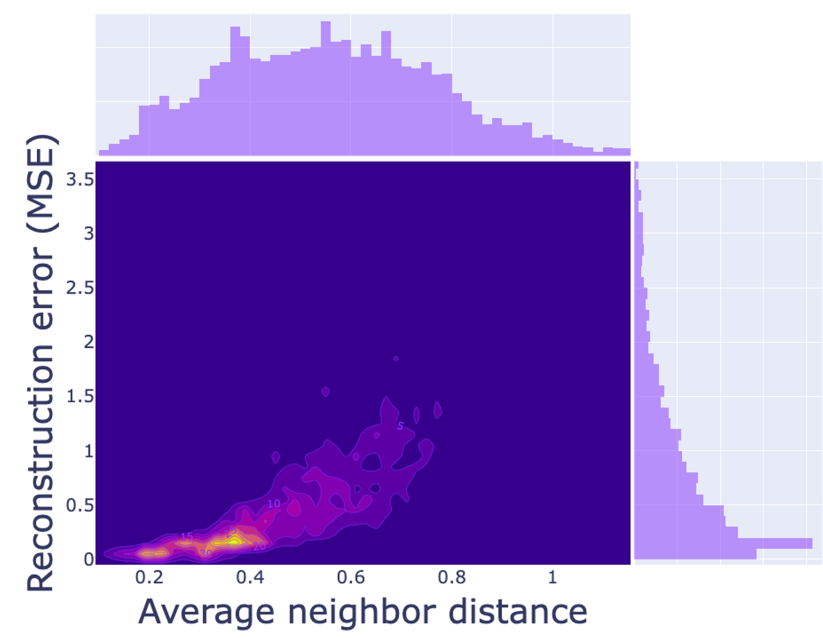
Completing a new realization with CVAE



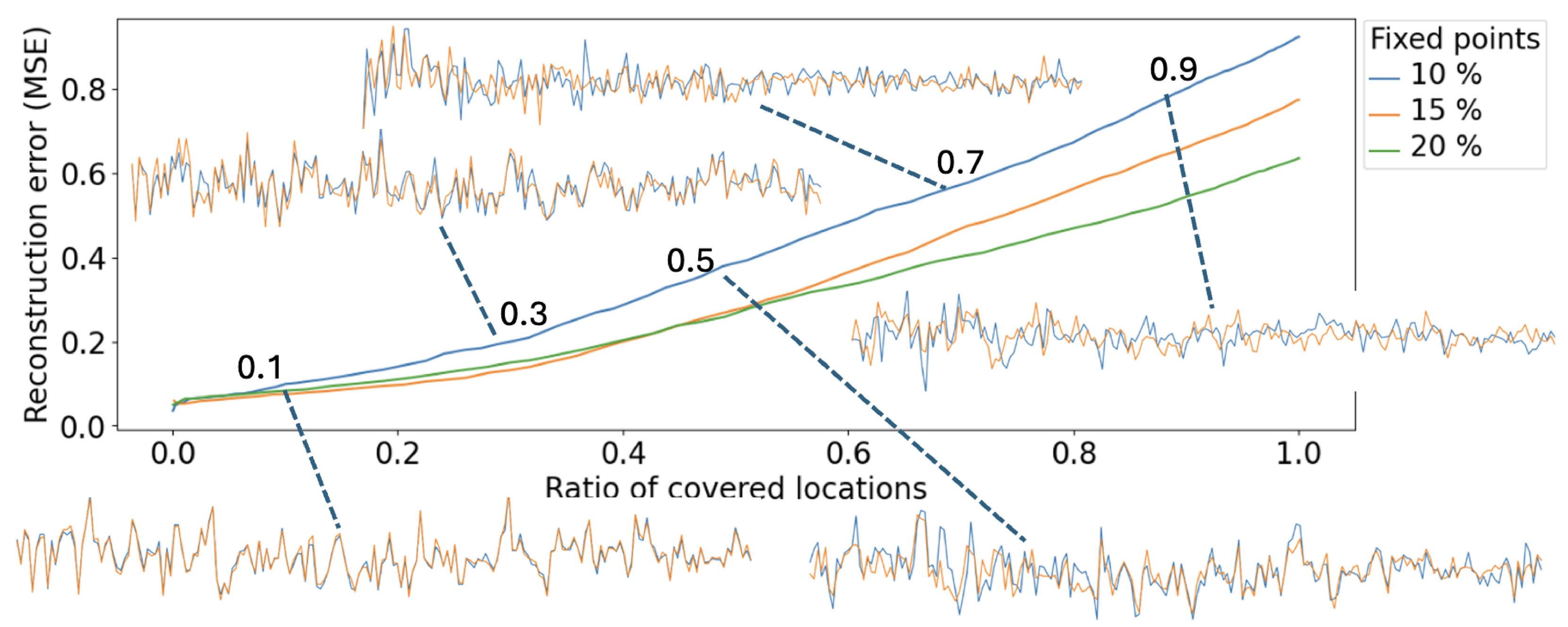
Experimental evaluation and ablation study



→ single realization is unstable, diminishing returns after 5 realizations

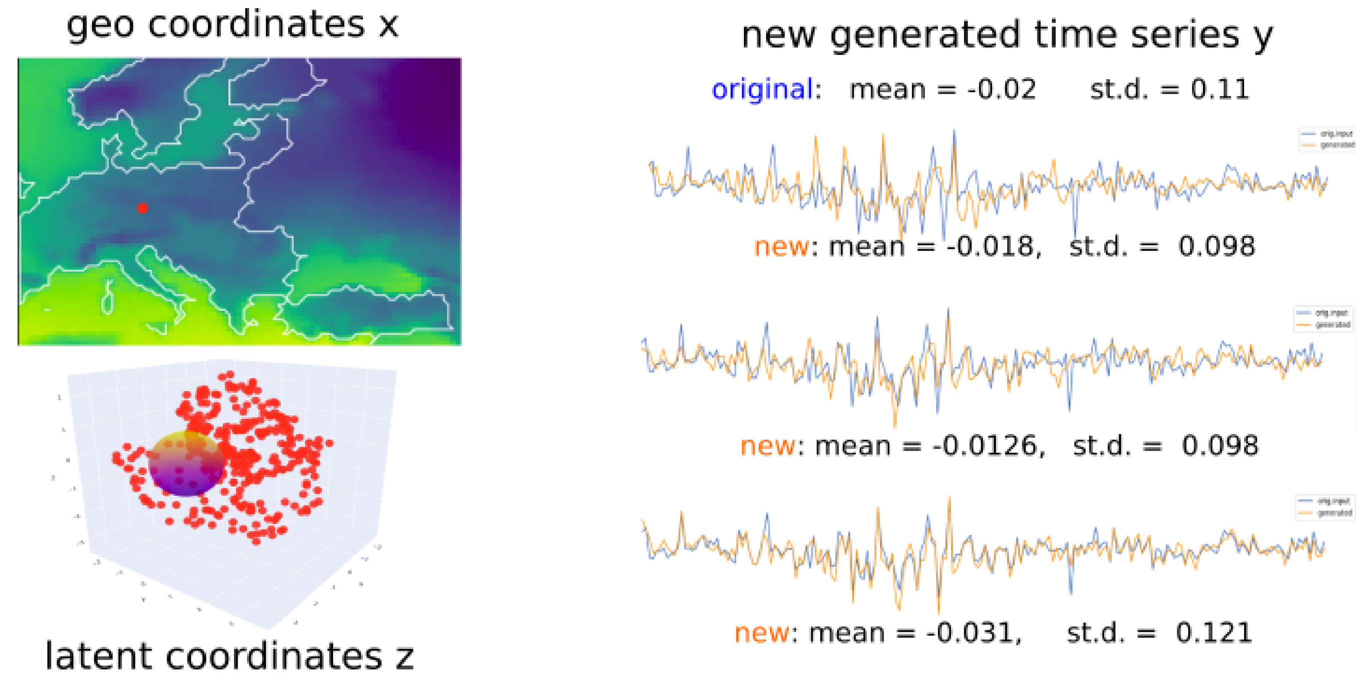


→ after a certain threshold, reconstruction error is correlated with average neighbor distance



→ Trade-off between ratio of covered locations and quality of reconstruction (order by neighbor distance).
→ Time series at selected locations show increasing deviation between true and reconstructed curves.

Generate new samples with full schnurfle run



→ New time series retains statistical and temporal properties

References
[1] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Accessed 2.2024.
[2] Sohn, K., Yan, X., and Lee, H. Learning Structured Output Representation using Deep Conditional Generative Models. NeurIPS, 2015.
[3] Rezende, D., J., and Viola, F. Taming VAEs. aarXiv:1810.00597, 2018.
[4]van der Wilk, Mark and Dutordoir, Vincent and John, ST and Artemev, Artem and Adam, Vincent and Hensman, James. A Framework for Interdomain and Multioutput Gaussian Processes. arXiv:2003.01115, 2020.