# Beyond PCA: Machine Learning Approaches to Extract Dynamical Drivers

**Jonathan O. Melcher**[1], Jens H. Christensen[1], Chongyang Zhang[2], Peter L. Langen[2], Shuting Yang[3]

[1]Physics of Ice, Climate and Earth, Niels Bohr Institute, University of Copenhagen;
[2]Department of Environmental Science, iClimate, Aarhus University;
[3]National Centre for Climate Research, Danish Meteorological Institute

Contact: jonathan.melcher@nbi.ku.dk

KØBENHAVNS UNIVERSITET

## Abstract

Different areas of climate science uses dimensionality reduction. In the field of teleconnections, EOF/PCA and its rotated variant dominate[1]. Machine Learning offers and an alternative, that can capture nonlinear connections. This poster shows a proof of concept with a CVAE trained on 5-day low pass filtered mean sea level pressure (MSLP) for the entire year from ERA 5.

- **AutoEncoders (AE):** Is a neural network where the input and output is the same. It has a bottleneck with few neurons. This encodes (fig 1d) the data in a low dimensional latent space, that is decoded (fig 1f) to resemble the input.
- **Variational AE (VAE):** The bottleneck is replaced by a multi-dimensional unit Gaussian (fig 1e). This allows meaningful interpolation in the latent space[2].
- **Convolutional VAE (CVAE):** To reduce the number of parameters and encode spatial relations, convolutional layers are used (fig 1 d & f).

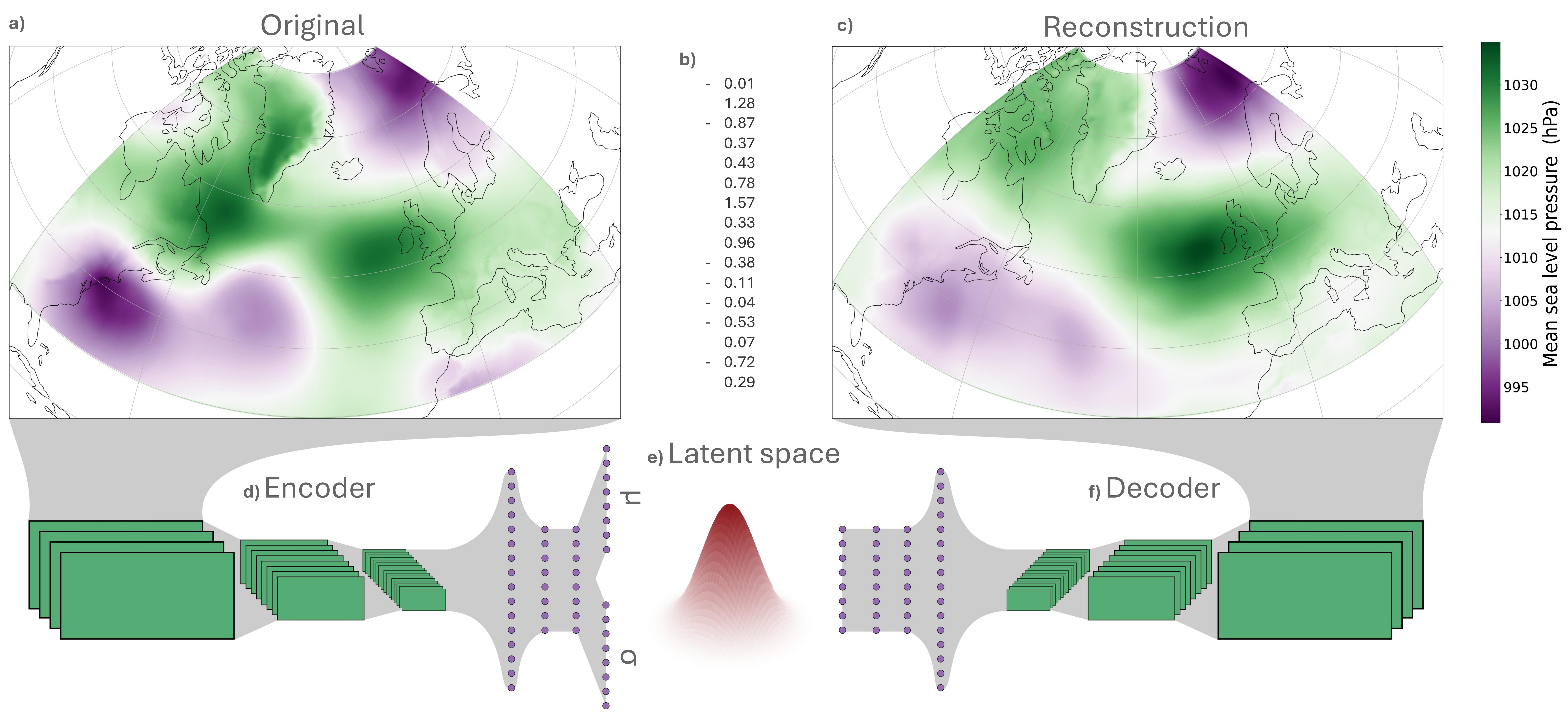


**Figure 1**: Schematic of dimensionality reduction using a CVAE. **a** The original MSLP field from ERA5 (1940-2023) spanning 90°W-30°E and 30°N-80°N has been preprocessed by a 5-day lowpass giving 30681 samples of size 201 x 481. To work with the convolutions, this is interpolated to 256x512 (not shown) and has the pointwise mean subtracted and standard deviation set to unity. **b** 16 numbers that the original MSLP field has been encoded to. **c** The reconstruction from decoding the 16 numbers from the latent space interpolated to the original resolution. **d** Graphical representation of the encoder structure, it is split in two components: a Convolutional (CNN) and a Feed Forward Neural Network (FFNN) . The CNN (green squares) convolutes the input into 128 filters of size 32x64 and average pooling takes this to a vector fed as input to the FFNN (purple circles). This goes through fully connected layers with ReLU activation of size: 256,128,64,64,64,32. This then splits into two vectors that describe the mean (μ) and standard deviation (σ) of the Gaussian in the latent space **e.** Using the reparameterization trick[2] a sample is drawn from this distribution to be decoded. **f** Graphical representation of the decoder, it runs in reverse of the encoder **d** from the sampled point.

## Variations on CVAE

The CVAE architecture can be further adapted to find teleconnections, by allowing for more Gaussians in the latent space (fig 1e), directly predicting variables of interest[3], and having more variables as in- and outputs, (fig 1a & c).

## Teleconnections

Dimensionality reduction is the first step in finding teleconnection.

Step 2: timeseries construction
- Self-organizing maps
- k-means
- PCA / EOF

Step 3: timeseries analysis
- Linear fits
- Correlation
- Causal networks

This work is part of the ArcticPush Project "Arctic push on Europe's extremes: The anatomy of Arctic impact on European climate change". We would like to acknowledge the support of the Independent Research Fund Denmark, Grant id 4258-00025B

References:
1. IPCC "Annex IV: Modes of Variability." In *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the AR6 of the IPCC* https://doi.org/10.1017/9781009157896.018.
2. Kingma, Diederik P., and Max Welling. "Auto-Encoding Variational Bayes." arXiv, December 10, 2022.
3. Spuler, Fiona R., Marlene Kretschmer, Magdalena Alonso Balmaseda, Yevgeniya Kovalchuk, and Theodore G. Shepherd. "Learning Predictable and Informative Dynamical Drivers of Extreme Precipitation Using Variational Autoencoders." Weather and Climate Dynamics 6, no. 3 (2025): 995–1014.