IMPERIAL

Scale-Preserving Autoencoders for High-Dimensional Heavy-Tailed Data

A. Turner, N. Gnecco, A. Veraart
Department of Mathematics,
Imperial College London

High-dimensional, heavy-tailed data

High-dimensional data in fields such as climate, finance, insurance and astronomy possess low-dimensional structure and also often exhibit heavy tails. A single, perhaps extreme, observation may live in hundreds or thousands of dimensions, yet concentrate near a low-dimensional manifold.

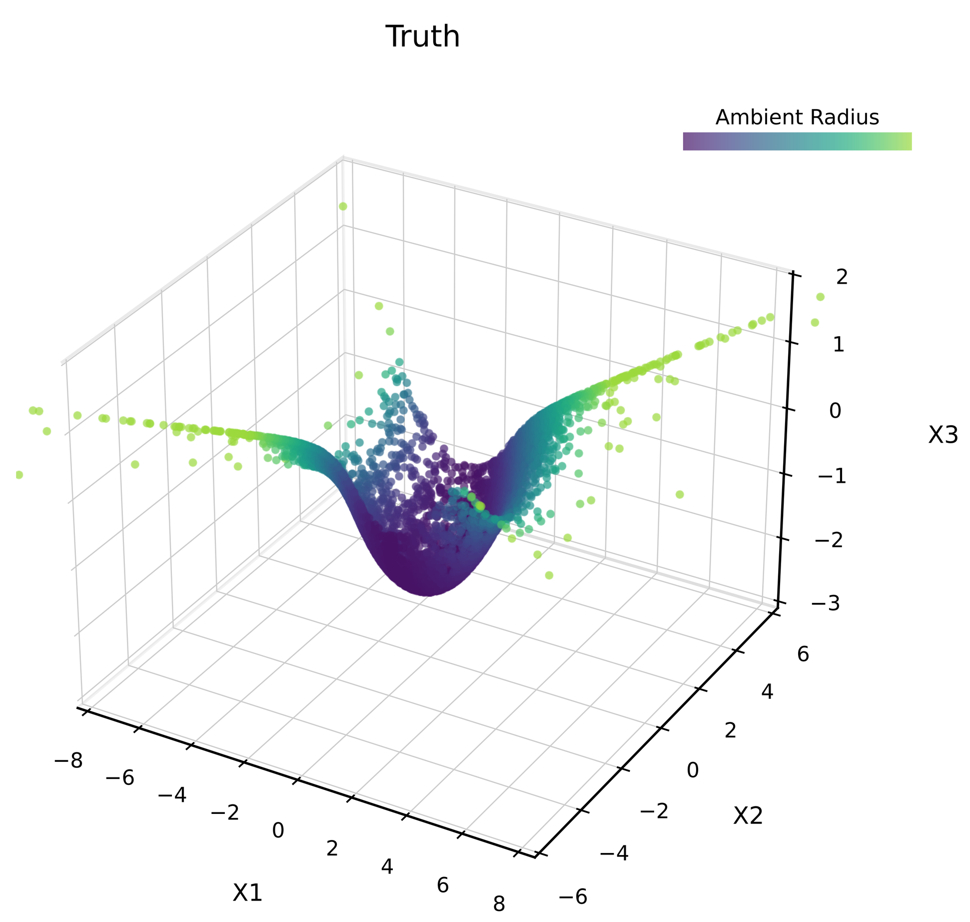


Figure 1: Heavy-tailed Student-t observations on a curved 2-surface in $\mathbb{R}^3$.

The events of greatest practical interest, such as climate disasters, market crashes and insurance losses, involve the most **extreme** observations in the tails of the distribution. A useful latent representation should transport the data to a lower dimensional space while retaining a notion of ambient space extremeness.

Standard autoencoders don't preserve magnitude

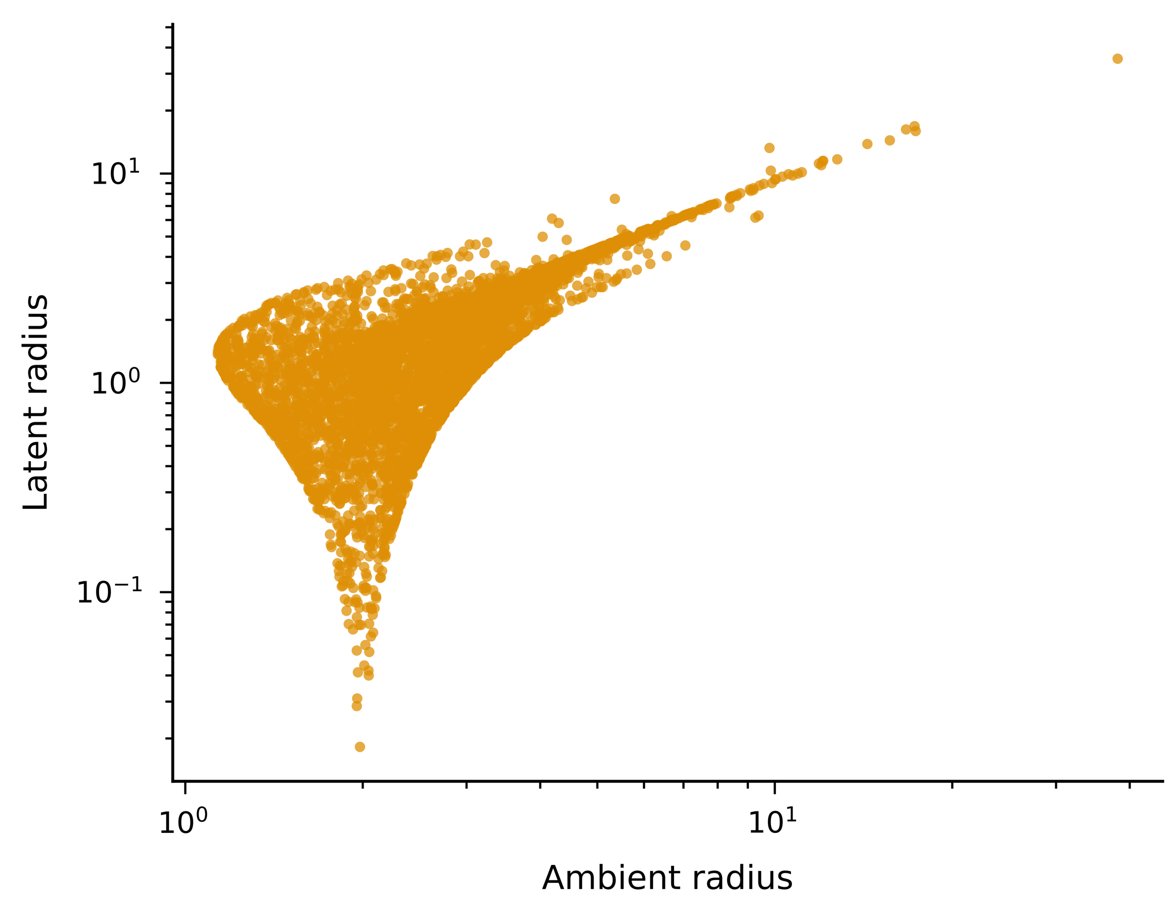


Figure 2: Latent $\|Z\|$ vs ambient $\|X\|$ radius for data encoded with a standard AE: $\|X\|$ does not have a clear correspondence to $\|Z\|$.

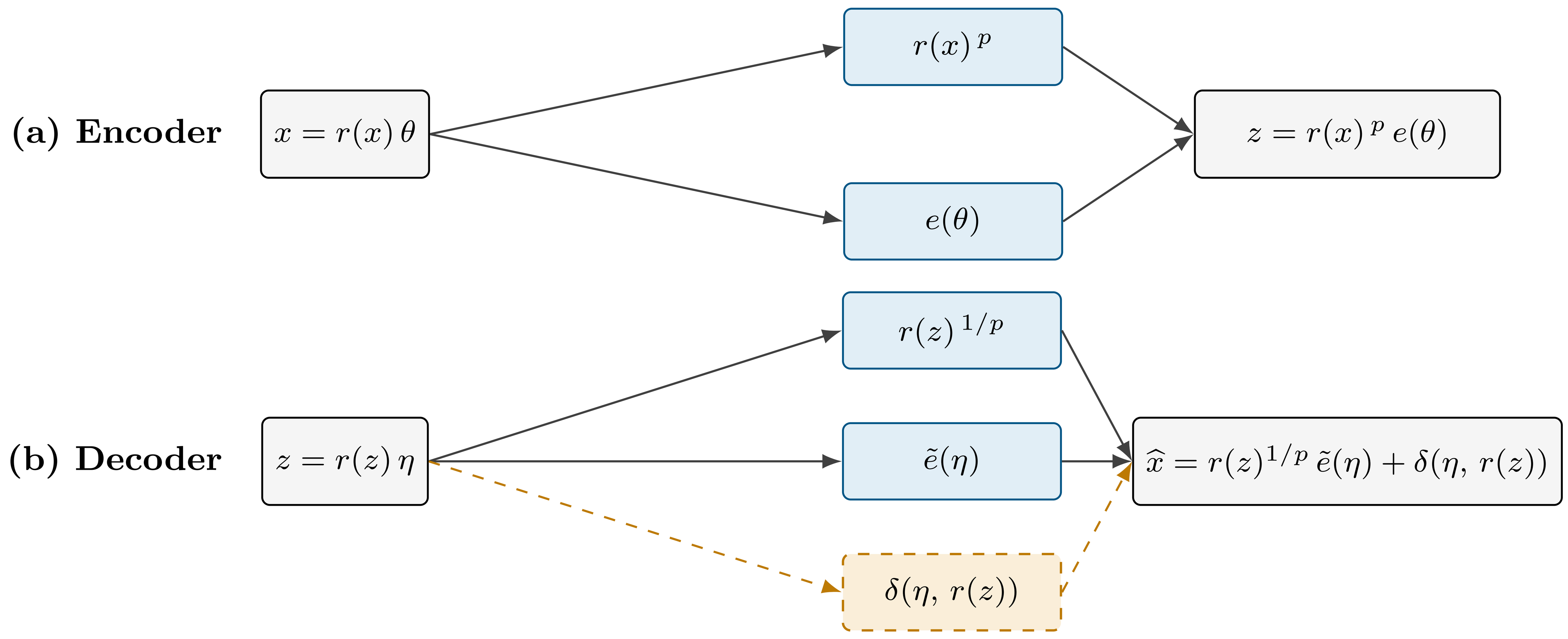
An unconstrained AE distorts radial scaling: $\|X\|$ is not reliably mapped to $\|Z\|$. This means that latent space radial diagnostics, such as latent tail index estimates, lose meaning with respect to the ambient data.

Homogeneous Autoencoder (HAE)

Decompose each ambient point as $x = r(x)\,\theta$ and transform only the angular component, leaving the radius preserved. We construct an **encoder** which is p-homogeneous and a **decoder** which is asymptotically $1/p$-homogeneous:

$$f(x) = r(x)^p\, e(\theta), \qquad h(z) = r(z)^{1/p}\, \tilde{e}(\eta) + \delta(\eta, r(z)),$$

with $\sup_\eta \|\delta(\eta, r(z))\| \to 0$ as $r(z) \to \infty$.
This construction guarantees preservation of the radius upon transport to and from latent space.



HAE with ERA5 atmospheric reanalysis data

ERA5 surface (10m) wind u_{10} over southwest England and Wales. $D{=}195$ grid cells, $m{=}8$. HAE versus a parameter-matched standard AE and PCA for comparison.

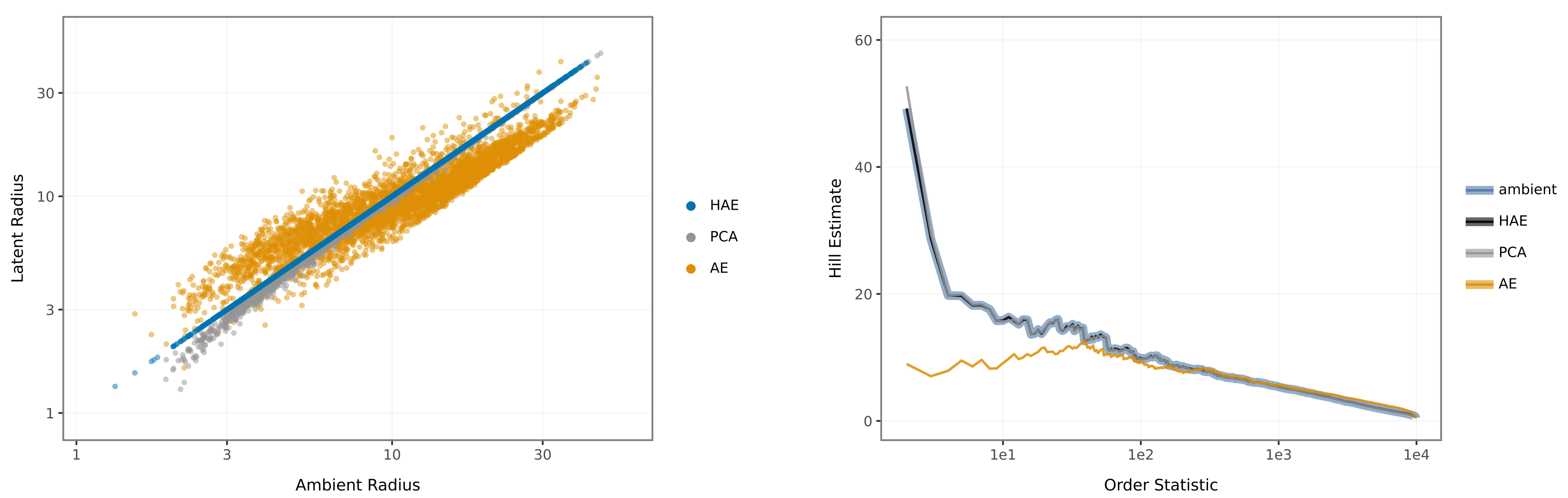


Figure 3: Left: latent vs ambient radius on u_{10} (AE drifts into a cloud). Right: latent Hill estimator (HAE matches ambient, AE drifts).

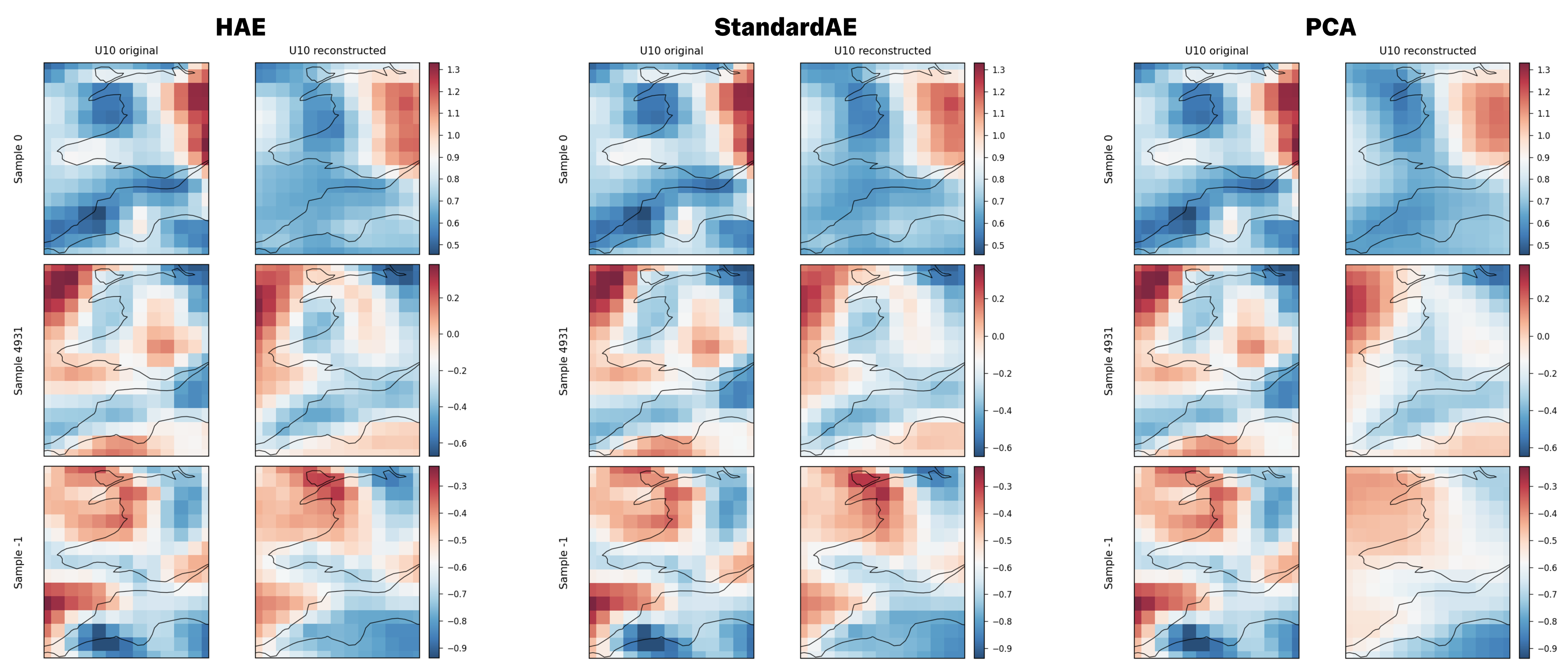


Figure 4: u_{10} wind-field reconstructions: original (left) vs reconstructed (right) per model.

Reconstruction MSE: HAE 9.2×10^{-3}, AE 9.0×10^{-3}, PCA 1.4×10^{-2}.
Latent Hill estimate drift $|\widehat{\alpha}_{\text{lat}} - \widehat{\alpha}_{\text{amb}}| \approx 0$ for HAE versus 0.26 for the AE.