HOMOGENEOUS AUTOENCODERS: MAGNITUDE-PRESERVING REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard nonlinear dimension-reduction methods do not in general guarantee preservation of observation magnitude, limiting their use in magnitude-sensitive tasks such as extreme-value analysis. We introduce the Homogeneous Autoencoder (HAE), whose encoder preserves a prescribed magnitude exactly while learning a map from ambient to latent directions. The encoder is positively homogeneous, so rescaling an observation rescales its latent point by the same factor. The decoder preserves the same magnitude while accommodating curved geometry. We identify conditions for homogeneous embeddings and provide a sufficient latent dimension bound for existence. For regularly varying (heavy-tailed) data, we characterise how encoding and decoding transform the limiting distribution of extremes and show that the radial tail index is unchanged. Common threshold-based models of extremes use approximate independence of magnitude and direction to extrapolate from moderate to larger events. HAE preserves this independence whenever it holds in the data, and we prove that, among magnitude-preserving encoders, homogeneity above the threshold is necessary and sufficient to preserve it for every input distribution. Experiments on a synthetic heavy-tailed curved surface and ERA5 precipitation fields show that HAE preserves latent magnitude exactly at reconstruction accuracy comparable to parameter-matched autoencoders, with the resulting latent representation supporting more accurate generation of held-out exceedances than a standard autoencoder.

1 INTRODUCTION

In climate science, finance, insurance and astronomy, observations may take the form of atmospheric variables over a spatial field (Hersbach et al., 2020), returns across an index of correlated stocks (Hofert & Oldford, 2018), portfolios of dependent risks (Cheng et al., 2024; Resnick, 1997) or emission rates for clusters of stars (Kollmeier et al., 2025). Modelling directly in the ambient space $\mathbb{R}^D$ can be computationally costly and statistically inefficient (Bellman, 1962; Chen, 2009). Under the *manifold hypothesis*, data exhibit lower-dimensional structure (Cayton, 2005; Bengio et al., 2013; Whiteley et al., 2026), which motivates latent representations adapted to that structure.

Standard autoencoders learn an encoder f and decoder h by optimising towards $h(f(x)) \approx x$ (Hinton & Salakhutdinov, 2006). Accurate reconstruction, however, does not identify the latent geometry since the encoder may distort an observation while the decoder compensates. In particular, for a chosen homogeneous radial functional r (§2), latent magnitude $r(f(x))$ need not reflect ambient magnitude $r(x)$ (Figure 1). This matters when magnitude is itself a quantity of interest, such as the Euclidean norm of an anomaly field or the $\|\cdot\|_1$ total of a nonnegative vector (Table C.1).

The issue is acute in the modelling of extremes, where the observations of greatest practical significance are rare and unusually large (Zscheischler & Seneviratne, 2017; Embrechts et al., 1997; Davison & Huser, 2015) and methods are built around quantities such as exceedance probabilities, tail indices and angular measures (Resnick, 2007; De Haan & Ferreira, 2006; Lindskog et al., 2014). A representation which distorts extreme values or radial ordering may therefore be unusable for tail-focused analysis in latent space (§2). Existing approaches typically either preserve magnitude or reconstruct curved geometry, not both. Linear maps such as principal component analysis (PCA) (Pearson, 1901; Jolliffe, 2002) are positively homogeneous, so regular variation is preserved whenever the projection retains tail mass (Resnick, 2007; Lindskog et al., 2014), but cannot reconstruct a generic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

curved manifold through a linear decoder. Nonlinear dimension reduction methods (Tenenbaum et al., 2000; Roweis & Saul, 2000; McInnes et al., 2020; Hinton & Salakhutdinov, 2006; Kingma & Welling, 2014) may learn curved manifold geometry effectively yet provide few, if any, guarantees relating ambient and latent magnitude.

We propose the *Homogeneous Autoencoder* (HAE), constructed in §3. The encoder is constrained to be positively homogeneous of degree one, $f(\lambda x) = \lambda f(x)$ for $\lambda > 0$, and parameterised so that $r(f(x)) = r(x)$ for every input. Latent magnitude equals ambient magnitude by construction while the learnt part of the encoder acts on the direction $\theta(x) = x/r(x)$. The decoder preserves the same radial coordinate and a radius-dependent angular correction lets its decoded direction vary with radius and vanishes uniformly at large radii. When magnitude is a meaningful quantity, such as a count, an amount or the size of an anomaly, latent radii may be read in ambient units, which a standard autoencoder does not provide (Proposition 5) and loss-based soft penalties incentivise only approximately. As well as this radial preservation, extreme-value analysis may benefit from further structure on the latent space. Common extreme-value approaches model exceedances over a threshold in the form of radial and angular components that become independent in the tail. Preserving the radius alone but leaving the latent direction free to vary with it may introduce dependence that this modelling approach rules out (Remark 12). The HAE encoder acts on radius and direction separately, so any radial–angular independence in the data survives encoding (Proposition 11). We show that reserving a latent coordinate for magnitude remains compatible with smooth embeddings under interpretable geometric conditions (§4). For regularly varying data the encoder transports the tail measure by pushforward (Proposition 9), the decoder correction is asymptotically negligible (Proposition 10) and the radial tail index is preserved throughout the round trip (Corollary E.4). These guarantees constrain the representation class and need not reduce reconstruction error. The experiments compare a standard autoencoder, a radial penalty, an autoencoder that preserves the radius exactly and HAE on a curved surface and ERA5 fields, and fit an extreme-value model to the surface's exceedances in each latent space to generate fields at a requested magnitude (§6).

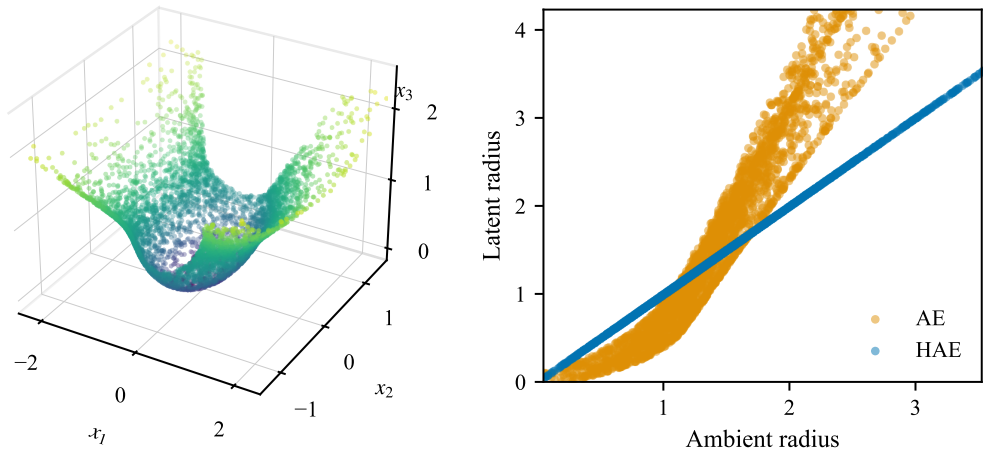


Figure 1: Data on a curved surface in $\mathbb{R}^3$ coloured by ambient radius (left). Latent against ambient radius for HAE and a standard autoencoder (right).

2 Preliminaries

Setup and notation. Magnitude is measured from the origin, so any centring of the data is preprocessing done beforehand. An observation is denoted $X \in \mathbb{R}^D$. Assume $\mathbb{P}(X \in \mathcal{M}) = 1$, where $\mathcal{M}$ is a smooth embedded m-dimensional manifold with $m < D$ (Cayton, 2005; Bengio et al., 2013; Whiteley et al., 2026). Near-manifold and noisy data are discussed in Appendix D.5. The latent dimension ℓ may differ from m (§4). We assume $\mathbb{P}(X = 0) = 0$ and work on $\mathbb{O}_D := \mathbb{R}^D \setminus \{0\}$ and $\mathbb{O}_\ell := \mathbb{R}^\ell \setminus \{0\}$.

Definition 1 (Conic set)**.** A set $A \subseteq \mathbb{R}^d$ is *conic* if $\lambda x \in A$ for every $x \in A$ and $\lambda > 0$.

Definition 2 (Positive homogeneity)**.** A map f defined on a conic set $A \subseteq \mathbb{R}^d$ is called *positively homogeneous* of degree $p > 0$ if $f(\lambda x) = \lambda^p f(x)$ for every $x \in A$ and every $\lambda > 0$.

Definition 3 (Radial functional)**.** For $d \in \mathbb{N}$, a *radial functional* $r : \mathbb{R}^d \to [0, \infty)$ is continuous, strictly positive (away from the origin) and positively homogeneous (Definition 2) of degree one. Its *unit shell* is the set $\mathbb{S}_r^{d-1} := \{x \in \mathbb{O}_d : r(x) = 1\}$.

Ambient and latent radial functionals are maps on $\mathbb{R}^D$ and $\mathbb{R}^\ell$ respectively, written r_D and r_ℓ where the distinction matters and r otherwise. $\|\cdot\|$ denotes the Euclidean norm, used for r in all experiments. Write $\theta(x) := x/r(x)$ for the normalisation of $x \in \mathbb{O}_d$ onto the unit shell, so that $x = r(x)\,\theta(x)$.

Proposition 4 (Radial–angular factorisation of positively homogeneous maps, proof in Appendix A.1)**.** *A map $f : \mathbb{O}_D \to \mathbb{R}^\ell$ is positively homogeneous (Definition 2) of degree $p > 0$ if and only if there exists a unique $g : \mathbb{S}_r^{D-1} \to \mathbb{R}^\ell$ such that $f(x) = r(x)^p\, g(\theta(x))$ for $x \in \mathbb{O}_D$.*

Reconstruction loss alone does not identify latent magnitude. Reconstruction loss depends on f and h only through $\hat{x} := h(f(x))$, so any change of latent magnitude that the decoder undoes leaves the loss unchanged.

Proposition 5 (Latent reparameterisation invariance, proof in Appendix A.2)**.** *Let $f : \mathbb{R}^D \to \mathbb{R}^\ell$ and $h : \mathbb{R}^\ell \to \mathbb{R}^D$ be an encoder and decoder and let ψ be a homeomorphism of $\mathbb{R}^\ell$ onto itself. Then $\tilde{f} := \psi \circ f$ and $\tilde{h} := h \circ \psi^{-1}$ satisfy $\tilde{h}(\tilde{f}(x)) = h(f(x))$ for every $x \in \mathbb{R}^D$.*

An encoder with bounded activations, such as tanh or sigmoid, saturates on large observations and maps them into a bounded region near the boundary of its range. Even a ReLU encoder (Nair & Hinton, 2010), which is piecewise linear and so admits a positively homogeneous limit $\lim_{t\to\infty} f(tx)/t$ along rays, is not constrained to be nonzero along the directions in which large observations occur. Positive homogeneity at finite magnitude instead requires the biases to be removed (Bamberger et al., 2025). Biases also accumulate across layers, so the encoder's response at the finite magnitudes where applied extreme-value methods operate need not resemble its limiting behaviour (Example A.2).

3 Homogeneous Autoencoders

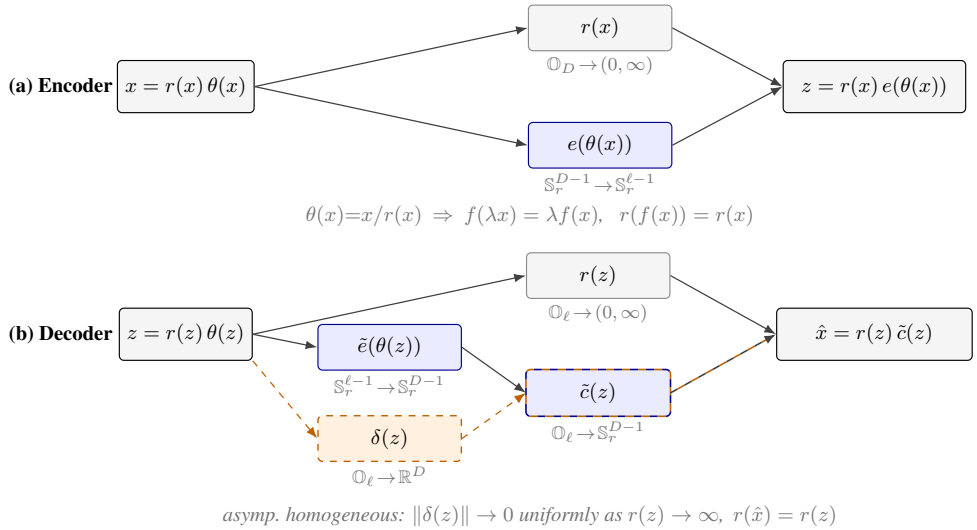


Figure 2: HAE architecture. (a) The encoder factorises into the radius $r(x)$ and an encoded direction $e(\theta(x))$ on the latent unit shell. (b) The decoder mirrors it with an angular correction δ that vanishes as $r(z) \to \infty$, renormalised onto the ambient unit shell.

Encoder. By Proposition 4, a positively homogeneous map is determined by its restriction to the unit shell, which we parameterise directly (Figure 2) by $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$,

$$f(x) = r(x)\,e(\theta(x)), \qquad r(f(x)) = r(x). \tag{3.1}$$

To preserve the radius exactly, we fix the homogeneity degree at one, with the general degree examined in Appendix E.6. The angular map is implemented by normalising a network $E : \mathbb{S}_r^{D-1} \to \mathbb{R}^\ell$ onto the latent unit shell, $e = E/r(E)$, defined wherever $E \neq 0$. Implementation details are given in Appendix C.1.

Decoder. A globally homogeneous decoder mirroring the encoder (3.1), $h_{\mathrm{hom}}(z) = r(z)\,\tilde{e}(\theta(z))$ with $\tilde{e} = \tilde{E}/r(\tilde{E})$ for a network $\tilde{E} : \mathbb{S}_r^{\ell-1} \to \mathbb{R}^D$ that is nonzero on the shell, similarly preserves magnitude. Every latent ray from the origin then maps to a single ambient ray (Figure C.2). To let the decoded direction vary with radius, introduce a correction term $\delta : \mathbb{O}_\ell \to \mathbb{R}^D$ on the latent space. Adding the correction to the network output and renormalising onto the unit shell gives

$$h(z) = r(z)\,\tilde{c}(z), \qquad \tilde{c}(z) := \frac{\tilde{E}(\theta(z)) + \delta(z)}{r\big(\tilde{E}(\theta(z)) + \delta(z)\big)} \in \mathbb{S}_r^{D-1}. \tag{3.2}$$

Since $\tilde{c}$ lies on the ambient unit shell, the identity $r(h(z)) = r(z)$ holds independently of δ, which only corrects the decoded angle. The normalisation in (3.2) requires $\tilde{E}(\theta(z)) + \delta(z) \neq 0$. This is assumed wherever $\tilde{c}$ is evaluated, holds at large radius by Lemma 6 and is enforced numerically (Appendix C.1). The correction is made to vanish at large radius by mapping the latent radius to $(1 + r(z))^{-1} \in (0, 1)$, as in Lafon et al. (2026), so that $r(z) \to \infty$ becomes the boundary point 0, taking $\Delta : \mathbb{S}_r^{\ell-1} \times [0, 1] \to \mathbb{R}^D$ continuous and subtracting its boundary value.

$$\delta(z) := \Delta\big(\theta(z), (1 + r(z))^{-1}\big) - \Delta(\theta(z), 0). \tag{3.3}$$

Lemma 6 (Uniform convergence of the corrected angular decoder, proof in Appendix B.1)**.** *Let Δ, δ and $\tilde{c}$ be as in (3.3) and (3.2). Then $\|\delta(z)\| \to 0$ uniformly in $\theta(z)$ as $r(z) \to \infty$. Moreover $\tilde{c}(z)$ is well defined whenever $r(z)$ is sufficiently large and $\|\tilde{c}(z) - \tilde{e}(\theta(z))\| \to 0$ uniformly in $\theta(z)$ as $r(z) \to \infty$.*

The decoder image is asymptotically conic (Corollary B.1), like the support of a regularly varying tail measure (Definition E.1), so accurate reconstruction of the largest observations requires the limiting decoder directions to match those of the data. On the curved surface the learnt correction vanishes as the radius grows (Figure C.4).

Radial identity and scale equivariance. Along each input ray the encoded direction is constant, so rescaling an input rescales its latent point by the same factor, and the equivariance error of Appendix F tests this.

Proposition 7 (Radial and scale identities, proof in Appendix B.2)**.** *Under well-defined shell normalisations in (3.1) and (3.2), for every $x \in \mathbb{O}_D$ and every $\lambda > 0$,*

$$f(\lambda x) = \lambda f(x), \qquad r_\ell(f(x)) = r_D(x), \qquad r_D\big(h(f(x))\big) = r_D(x).$$

Training objective. The angular maps e and $\tilde{e}$ and the correction Δ of (3.1), (3.2) and (3.3) are multilayer perceptrons (Rumelhart et al., 1986) and training minimises reconstruction loss with early stopping on a validation split (Appendices C.1 and C.2). The reconstruction loss can be chosen for the task, including angular losses for extremes (Drees & Sabourin, 2021), Fourier-space losses for weather fields (Yan et al., 2024; Subich et al., 2025) or likelihood-based objectives (Brehmer & Cranmer, 2020; Horvat & Pfister, 2021). The radial and scaling guarantees are architectural and hold for any loss or training procedure.

4 Existence of homogeneous radius-preserving embeddings

Reserving a coordinate for magnitude restricts the encoder. Since latent magnitude equals ambient magnitude, observations of different magnitudes are already separated, so the angular map only has to keep observations of equal magnitude distinct and preserve their directions of variation. At the minimal latent dimension $\ell = m$ an exact embedding imposes topological and geometric conditions on the fixed-magnitude slices, discussed in Appendix D.3. In applications m is generally unknown, so ℓ may be chosen as a conservative upper bound (§6.2). When $m \ll D$, choosing $m < \ell \ll D$ retains substantial compression while allowing greater representational freedom and radius-preserving embeddings then exist at a latent dimension determined by m, by the argument of the Whitney embedding theorem (Lee, 2012, Theorem 6.15).

Proposition 8 (Existence of overcomplete HAE embeddings, proof in Appendix D.2)**.** *Let $r_D : \mathbb{O}_D \to (0,\infty)$ and $r_\ell : \mathbb{O}_\ell \to (0,\infty)$ be smooth radial functionals (Definition 3), and let $U \subseteq \mathcal{M} \cap \mathbb{O}_D$ be an open subset of the data manifold of §2. If $2m+2 \le \ell < D$, then there exists a smooth angular map $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$ such that the HAE encoder (3.1), $f(x) = r_D(x)\, e(\theta(x))$, is an injective smooth immersion (Definition D.1) of U into $\mathbb{O}_\ell$ and satisfies $r_\ell(f(x)) = r_D(x)$ for every $x \in U$. Thus $f|_K$ is a smooth embedding (Definition D.2) for every compact embedded submanifold $K \subset U$.*

For $2m + 2 \le \ell < D$ exact radial preservation is compatible with an embedding of compact data regions. Sharper bounds and extensions to piecewise-smooth radial functionals such as $\|\cdot\|_1$ and $\|\cdot\|_\infty$ remain future work. Lower-dimensional sufficient conditions, singularity of the latent law with respect to Lebesgue measure when $\ell > m$ and near-manifold data are discussed in Appendix D.

5 Consequences for extreme-value analysis

Tail transport through the autoencoder. Regular variation is the standard framework for the heavy-tailed data considered here (Resnick, 2007; Lindskog et al., 2014), under which the radius has a power-law tail and the directions of the largest observations have a limiting distribution. Rescaling X by $b(t) \to \infty$ and weighting by t gives a sequence of measures $t\,\mathbb{P}(X/b(t) \in \cdot)$ converging to a nonzero limit μ. This limit is called the *ambient tail measure* and represents the directions in which extremes occur. The tail measure is homogeneous, $\mu(\lambda A) = \lambda^{-\alpha}\mu(A)$, and the *tail index* $\alpha > 0$ describes the heaviness of the tail. Definition E.1 states this precisely. Letting $T_\# \nu := \nu \circ T^{-1}$ denote the pushforward of a measure ν by a measurable map T, we establish the preservation of regular variation throughout the HAE round trip.

Proposition 9 (Regular variation under the HAE encoder, proof in Appendix E.1)**.** *Let X be regularly varying (Definition E.1) on $\mathbb{O}_D$ with scaling function b, index $\alpha > 0$ and tail measure μ. Let $f : \mathbb{O}_D \to \mathbb{O}_\ell$ be an HAE encoder (3.1) that is measurable and continuous μ-almost everywhere. Then $f(X)$ is regularly varying on $\mathbb{O}_\ell$ with the same scaling function b, the same tail index α and tail measure $f_\#\mu$.*

Two random vectors are tail-equivalent when they share a scaling function and a tail measure (Definition E.3). By Lemma 6 the corrected decoder deviates from its homogeneous leading term h_{hom} of §3 by decreasing amounts and is sufficient for tail equivalence. The round trip transports the tail measure with the radial index preserved (Corollary E.4). It is not constrained to return μ itself, which requires a further condition on the angular round trip (Appendix E.5).

Proposition 10 (Tail equivalence under asymptotically homogeneous decoding, proof in Appendix E.2)**.** *Let Z be regularly varying (Definition E.1) on $\mathbb{O}_\ell$ with scaling function b_Z, tail index $\beta > 0$ and tail measure μ_Z. Assume that $\tilde{e}$ is continuous and that the decoder h of (3.2) is well defined on $\mathbb{O}_\ell$ with $\|\tilde{c}(z) - \tilde{e}(\theta(z))\| \to 0$ uniformly in $\theta(z)$ as $r(z) \to \infty$ (Lemma 6). Then $h(Z)$ and $h_{\mathrm{hom}}(Z)$ are tail-equivalent with scaling function b_Z, common tail index β and common tail measure $(h_{\mathrm{hom}})_\#\mu_Z$.*

Radial–angular factorisation. Peaks-over-threshold generation of extremes involves fitting an angular law to the exceedances beyond a chosen level and assuming that this distribution does not depend on the size of the excess (Davison & Huser, 2015), as regular variation implies in the limit (Resnick, 2007).

Proposition 11 (Preservation of factorisation, proof in Appendix E.4)**.** *Fix $u > 0$. A measurable radius-preserving encoder $f : \mathbb{O}_D \to \mathbb{O}_\ell$ with $r_\ell(f(x)) = r_D(x)$ preserves independence of radius and direction conditional on $r(X) > u$, for every law of X on $\mathbb{O}_D$ with positive exceedance probability, if and only if $f(x) = r(x)\, e(\theta(x))$ whenever $r(x) > u$, for some measurable $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$. For such encoders, including HAE (3.1), the total-variation distance of the conditional radial–angular law from the product of its marginals cannot increase.*

Remark 12 (A radius-preserving, nonhomogeneous construction)**.** An alternative encoder $x \mapsto r(x)\, e_{\mathrm{rad}}(x)$ whose angular network $e_{\mathrm{rad}} : \mathbb{O}_D \to \mathbb{S}_r^{\ell-1}$ takes the full coordinate x rather than its projection to the angular shell $\theta(x)$ satisfies the two radial identities of Proposition 7. However, this construction is generally not scale equivariant and the encoded direction is free to vary with radius. Its latent radius is the ambient radius and keeps the tail index of the radius, but the latent vector may

not be regularly varying (Definition E.1), since e_{rad} need not converge along rays, so no latent tail measure or angular law is guaranteed. It is compared as AE + radius in §6.

6 Experiments

HAE is compared against a standard autoencoder (AE), an AE with a radial-consistency penalty (AE + penalty), the AE + radius construction of Remark 12 and uncentred PCA, all fitted on the training split at a shared latent dimension (Appendix C.2). Appendix F defines the models and every reported metric.

6.1 Curved surface

A synthetic $m{=}\ell{=}2$-dimensional surface is observed in $\mathbb{R}^3$ (Figure 1) and its radius is heavy tailed. Above the 0.85 quantile of the radius its direction approaches independence from the radius, up to a residual that a permutation test cannot detect. Below it the surface is displaced off the cone over its limiting directions, which a homogeneous decoder cannot represent and the correction has to absorb (details in Appendix G). The neural models reconstruct the surface with RMSE below one percent of the median test radius of 1.4 and PCA cannot represent the curvature (Table 1, reconstructed surfaces in Appendix G). HAE reconstructs the surface as accurately as the other neural models. AE's latent radii have no identifiable relation to the ambient radii (Figure 3), AE + penalty nearly preserves them and the radial preservation of AE + radius and HAE is exact by construction (Proposition 7), so only the equivariance error distinguishes those two.

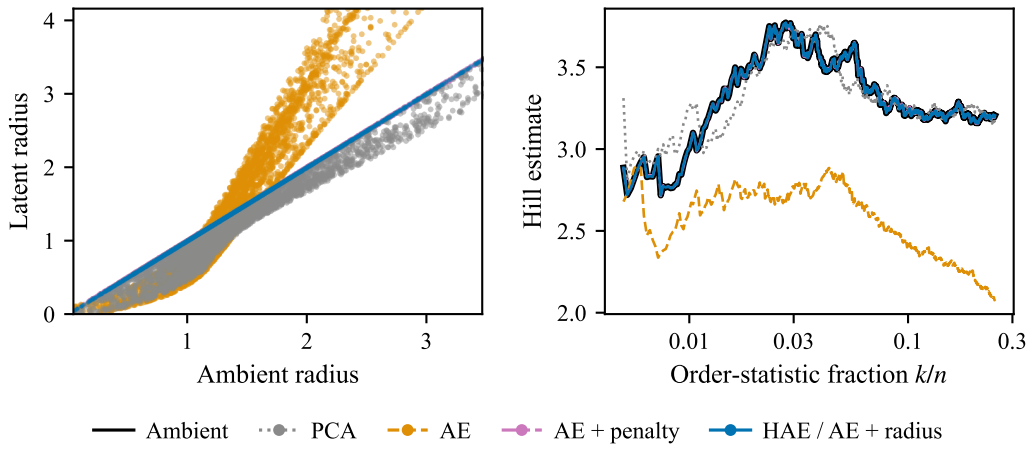


Figure 3: Curved surface ($D{=}3$, $m{=}\ell{=}2$), latent against ambient radius (left) and latent Hill-estimator curves (right).

Table 1: Curved surface ($D{=}3$, $m{=}\ell{=}2$), mean (standard deviation) across twenty-five seeds. † indicates the quantity is exact by construction. Reconstruction also provided as a median (median absolute deviation) row, since one extreme test observation dominates the mean.

	PCA	AE	AE + penalty	AE + radius	HAE
Reconstruction RMSE, mean	0.451 (0.018)	0.00264 (0.0037)	0.00180 (0.0011)	0.000734 (0.00049)	0.00175 (0.0035)
Reconstruction RMSE, median	0.447 (0.0067)	0.00160 (0.00033)	0.00140 (0.00022)	0.000552 (0.00012)	0.000814 (0.00037)
Equivariance error, scale 10	$0^{\dagger}$	1.48 (0.62)	0.0724 (0.023)	0.0875 (0.015)	$0^{\dagger}$
Mean absolute radial error	0.187 (0.0023)	1.50 (1.8)	0.00138 (0.00098)	$0^{\dagger}$	$0^{\dagger}$
Rank correlation of radii	0.976 (0.0024)	0.794 (0.23)	1.000 (0.000012)	$1^{\dagger}$	$1^{\dagger}$
Hill estimate difference	0.0700 (0.044)	0.491 (0.20)	0.00310 (0.0030)	$0^{\dagger}$	$0^{\dagger}$

Generating exceedances. A common workflow for tail events fits an angular law to moderate exceedances and generates larger ones from its product with a fitted univariate radial distribution

(Davison & Huser, 2015; Lafon et al., 2026; Engelke et al., 2026). For this experiment, we examine excess vectors $(r(x) - u)\,\theta(x)$ above the 0.85 radial quantile u. The models are then fitted to all training exceedances with latent dimension two and a generative angular law is fitted to only the lower half of excesses in this latent space. Testing how this latent angular law extrapolates avoids asking the decoder to leave its training range, since the decoder was trained on the full exceedance set. This allows the experiment to isolate generative extrapolation performance from pointwise decoder extrapolative ability. Over test exceedances the generation error is quantified as the energy distance between generated and observed directions. Resampling the observed lower-half directions in place of generating them, so that nothing passes through a model, gives the ambient generation error, the 'ambient' column of the tables. Angular reconstruction error, the distance correlation between latent radius and direction and its permutation test are defined in Appendix F, and the sampler in Appendix H. The exceedance set is drawn once, so the presented variance is across fits not dataset draws.

The difference in performance between HAE and AE + radius isolates what homogeneity adds beyond exact radius preservation (Table 2, latent angles in Figure H.1, generated points in Figure H.3). HAE's angular encoder uses ambient direction alone, whereas AE + radius permits radius-dependent latent directions. The permutation test does not reject independence for HAE on any fit but rejects several AE + radius fits. Generating excesses larger than the fitted half, HAE attains about two thirds the error of AE + radius under the fitted law and a little over half under the empirical law, the latter in all twenty-five paired fits. The separation is therefore between the representations rather than the angular models (Appendix H). AE and AE + penalty reconstruct accurately but exhibit substantially larger radial–angular dependence, which impairs the transfer of an angular law from moderate to larger excesses. PCA's generation error largely reflects its angular reconstruction error.

Table 2: Curved-surface exceedances, mean (standard deviation) over twenty-five fits. The fitted law is the von Mises–Fisher mixture of Appendix H and the empirical law resamples the training latent directions themselves, both using the same decoder. The 'ambient' column exists only for the empirical law as no dimension reduction takes place. Ambient and PCA methods are deterministic, so no standard deviations are given.

	Ambient	PCA	AE	AE + penalty	AE + radius	HAE
Generation error, fitted law		0.108	0.166 (0.088)	0.00875 (0.0056)	0.00323 (0.0015)	0.00203 (0.00061)
Generation error, empirical law	0.000808	0.107	0.167 (0.089)	0.00885 (0.0061)	0.00313 (0.0013)	0.00179 (0.00028)
Distance correlation	0.0433	0.0356	0.177 (0.066)	0.0765 (0.021)	0.0502 (0.011)	0.0399 (0.0074)
Fits rejecting independence	0/1	0/1	25/25	21/25	5/25	0/25
Reconstruction RMSE		0.431	0.00400 (0.0014)	0.00352 (0.00071)	0.00176 (0.00048)	0.000941 (0.00035)

6.2 ERA5 PRECIPITATION

The ERA5 data (Hersbach et al., 2020) covers southwest England and Wales ($53.5°$N–$50°$N, $5.5°$W–$2.5°$W) at $0.25°$ resolution and six-hourly intervals over 1979–2023. It contains $65{,}744$ fields with $D = 195$ cells, randomly split into training, validation and test sets in proportions 0.70, 0.15 and 0.15. For total precipitation tp, removing fully dry fields leaves $62{,}749$ observations, and each cell is then rescaled by its training-split standard deviation. Magnitude is the Euclidean norm of the standardised field, which weights each cell by its own variability and so measures field size relative to typical conditions rather than a total rainfall amount. A physically interpretable total would instead be the $\|\cdot\|_1$ functional of Table C.1, which the construction accommodates but which is not used here. The neural models reconstruct precipitation with nearly equal RMSE and all of them beat PCA, so fixing a magnitude coordinate does not discernibly cost accuracy on real fields (Table 3). The AE distorts the radius substantially and AE + penalty nearly preserves it, whereas AE + radius and HAE preserve it exactly by construction, and of those two only AE + radius departs from scale equivariance. The same pattern holds for wind, temperature and the five-variable joint tensor (Appendix I).

Table 3: ERA5 total precipitation (D=195, ℓ=44), means over three fits. Entries marked † are exact by construction.

	PCA	AE	AE + penalty	AE + radius	HAE
Reconstruction RMSE	0.164	0.135	0.133	0.137	0.138
Equivariance error, scale 10	$0^\dagger$	0.510	0.135	0.208	$0^\dagger$
Mean absolute radial error	0.179	4.56	0.0447	$0^\dagger$	$0^\dagger$
Rank correlation of radii	1.000	0.992	1.000	$1^\dagger$	$1^\dagger$
Hill estimate difference	0.0157	0.693	0.0103	$0^\dagger$	$0^\dagger$

Exceedances. Precipitation exceedances above the 0.70 radial quantile are represented as excess vectors with latent dimension 36 (Appendix J). Unlike the synthetic curved surface, directions *do* exhibit some dependence on the radial excess (ambient distance correlation 0.22), so the strict factorisation of the exceedance-generation experiment (§6.1) does not hold here and cannot be directly tested. This experiment examines whether encoding amplifies existing dependence or leaves it unchanged. Independence is correctly rejected in every fit, as it is in the ambient space itself. HAE and PCA remain closest to the ambient distance correlation, with AE + penalty and AE + radius somewhat larger and AE substantially larger (Table 4). Base reconstruction accuracy is again matched across the neural models, with PCA worse.

Permuting the observed radii against the angles within each split restores independence by construction while retaining the angular structure present in precipitation fields. Under this dataset, HAE and PCA empirically carry the independence through, whereas the other neural baselines exhibit larger distance correlation and reject independence in every fit (Table 5). AE + radius exhibits distance correlation more than three times the ambient value despite preserving the radius exactly. Under the fitted law HAE, AE + penalty and PCA all sit near the ambient generation error on the exceedances, whereas after permutation every neural model generates worse than PCA and than that error. Resampling training latent directions from their empirical distributions reduces HAE's error on the exceedances below both references, implicating the von Mises–Fisher generative fit rather than dependence on the excess (Appendices H and J).

Table 4: Precipitation exceedances, mean (standard deviation) over twenty-five fits. Columns and laws follow Table 2.

	Ambient	PCA	AE	AE + penalty	AE + radius	HAE
Generation error, fitted law		0.0217	0.188 (0.0084)	0.0204 (0.00073)	0.0372 (0.014)	0.0238 (0.00093)
Generation error, empirical law	0.0260	0.0192	0.172 (0.0089)	0.0137 (0.00060)	0.0194 (0.011)	0.0151 (0.00025)
Distance correlation	0.216	0.218	0.594 (0.023)	0.260 (0.013)	0.272 (0.038)	0.235 (0.011)
Fits rejecting independence	1/1	1/1	25/25	25/25	25/25	25/25
Reconstruction RMSE		0.233	0.218 (0.00088)	0.217 (0.00088)	0.221 (0.0011)	0.222 (0.00088)

Table 5: Precipitation exceedances with the observed radii permuted against the angles within each split and every model refitted, so that independence holds by construction. Columns and laws follow Table 2.

	Ambient	PCA	AE	AE + penalty	AE + radius	HAE
Generation error, fitted law		0.00321	0.151 (0.012)	0.0332 (0.0017)	0.0740 (0.013)	0.0364 (0.0017)
Generation error, empirical law	0.00120	0.00248	0.120 (0.012)	0.00515 (0.00045)	0.0331 (0.013)	0.00381 (0.00018)
Distance correlation	0.0792	0.0672	0.482 (0.031)	0.118 (0.015)	0.261 (0.050)	0.0793 (0.0038)
Fits rejecting independence	0/1	0/1	25/25	25/25	25/25	0/25
Reconstruction RMSE		0.355	0.319 (0.0022)	0.322 (0.0022)	0.329 (0.0038)	0.332 (0.0026)

7 Related work

Extreme-value theory in machine learning. Extreme-value theory has recently been combined with machine learning in several ways, often through the radial–angular decomposition used here (Engelke et al., 2026). Variational autoencoders (VAEs) (Kingma & Welling, 2014) have been

used to model extremes, both at full latent dimension (Lafon et al., 2026) and via structured lower-dimensional spatial factors (Zhang et al., 2026). Several authors apply generative techniques such as normalising flows (Rezende & Mohamed, 2015), adversarial methods (Goodfellow et al., 2014) or graphical approaches (Engelke & Ivanovs, 2021; Engelke et al., 2024) to multivariate data factored into marginal approximation and dependence modelling tasks (McDonald et al., 2022; Hu & Castro-Camilo, 2025; Amiri et al., 2022; Allouche et al., 2026; Boulaguiem et al., 2022; Lhaut et al., 2026; Peard & Hall, 2023; Monte et al., 2025; Wessel et al., 2025). Other work modifies generative models so that they remain compatible with extremes (Jaini et al., 2020; Laszkiewicz et al., 2022; Hickling & Prangle, 2025; Pandey et al., 2025). These methods operate in the ambient space, or in a latent space whose relation to observed magnitude is established and imposed separately. HAE provides a dimension-reduction layer that preserves ambient magnitude exactly, and the exceedance-generation experiment (§6.1) runs a generative model in the latent space.

Manifold learning and autoencoders. Nonlinear dimension reduction targeting bulk reconstruction or local geometric structure includes Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis & Saul, 2000), UMAP (McInnes et al., 2020), autoencoders (Hinton & Salakhutdinov, 2006) and variational autoencoders (Kingma & Welling, 2014), together with hyperspherical variants (Davidson et al., 2018) that place latent points on a sphere and discard magnitude altogether. Dimension reduction aimed at extremes estimates a low-dimensional angular support instead, through principal component analysis for extremes and kernel extensions (Drees & Sabourin, 2021; Medina et al., 2025). None of the nonlinear methods are designed to preserve magnitude and the flexibility that fits curved geometry also distorts magnitude and radial ordering in latent space, while the extremes-oriented techniques lack the expressivity to represent general curved manifolds. Where the tail support is close to a linear subspace those techniques already transport regular variation and reconstruct it accurately, so the contribution here is not the tail alone but data carrying a curved bulk and regularly varying extremes together. HAE imposes equivariance under the multiplicative action of $\mathbb{R}_+$, distinct from spatial scale equivariance in computer vision (Sosnovik et al., 2019; Worrall & Welling, 2019).

8 Discussion

HAE fixes magnitude and is scale equivariant while leaving the angular representation flexible and learnable. Radius-preserving embeddings exist under the geometric conditions of §4, and regularly varying tails retain their radial index throughout the round trip. On the curved surface and ERA5 fields, HAE removes latent radial distortion at comparable reconstruction accuracy. An angular law fitted to moderate latent exceedances gives the lowest generation error across the models for exceedances in curved-surface data. AE and AE + penalty exhibit substantially larger radial–angular dependence, which impairs the transfer of an angular law learnt on smaller observations to larger excesses. For precipitation, HAE tracks the observed dependence most closely, whereas the neural baselines exhibit a larger one and show detectable dependence on radius-permuted data. Thus, HAE supports magnitude-sensitive analyses where a low-dimensional representation is needed.

Limitations. HAE preserves the chosen magnitude, not its relevance to a task. An unsuitable origin, a noise-dominated radius or a fundamentally magnitude-invariant task can make this constraint inappropriate. The geometric results establish the existence of radius-preserving embeddings and not guarantees for optimisation, sample complexity or estimation of sparsely observed tails.

Outlook. Further work includes an in-depth exploration of end-to-end latent generative modelling for extremes (Engelke et al., 2026; Lafon et al., 2026; Zhang et al., 2026), objectives controlling angular tail structure (Drees & Sabourin, 2021), approximately invertible architectures (Brehmer & Cranmer, 2020; Caterini et al., 2021) to recover the full ambient tail measure (Appendix E.5) and separate radial coordinates for components with different scaling (Appendix I.4). Experiments with alternative radial functionals and near-manifold noise would show when exact magnitude preservation is beneficial.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs of mathematical claims, to provide feedback on research methodology and experiments, to implement methods, including data preprocessing pipelines, and to interpret experimental results. Additionally, we used generative AI tools to create and modify scientific figures, to create and edit software code, to draft and edit parts of the manuscript, to search for and summarise relevant literature, to format references and to suggest structure. We have reviewed all AI-assisted work: all mathematical claims and proofs were verified by the authors, AI-implemented experiment code and its results were reviewed and checked by the authors and all experimental design decisions are the authors' own. We take responsibility for the final content of this work, including text, claims and artefacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code and the full experiment configuration are provided in the supplementary material. One seed of every experiment takes about a day on a laptop CPU and the twenty-five-seed results took around 500 core-hours on a PBS cluster. Architecture, training and metric specifications are given in Appendices C.1, C.2 and F, the surface construction in Appendix G and extended ERA5 results in Appendix I.

The ERA5 reanalysis data supporting this study were produced by ECMWF and are distributed by the Copernicus Climate Change Service through the Climate Data Store (DOI 10.24381/cds.adbb2d47). The experiments use the hourly single-level product, subset as in §6.2, for 10 m zonal and meridional wind (u_{10}, v_{10}), 2 m temperature (t_{2m}), total precipitation (tp) and mean sea-level pressure. The data are available under the Creative Commons Attribution 4.0 licence and the retrieval script is included in the supplementary material.

REFERENCES

Michaël Allouche, Stéphane Girard, and Emmanuel Gobet. ExceedGAN: Simulation above extreme thresholds using Generative Adversarial Networks. *Extremes*, 2026.

Saba Amiri, Eric Nalisnick, Adam Belloum, Sander Klous, and Leon Gommans. Generating Heavy-Tailed Synthetic Data with Normalizing Flows. In *The 5th Workshop on Tractable Probabilistic Modeling*, 2022.

Stefan Bamberger, Reinhard Heckel, and Felix Krahmer. Approximating positive homogeneous functions with scale invariant neural networks. *Journal of Approximation Theory*, 311:106177, 2025.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(46): 1345–1382, 2005.

Richard Bellman. *Applied Dynamic Programming*. Princeton, N.J. : Princeton University Press, 1962.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Younes Boulaguiem, Jakob Zscheischler, Edoardo Vignotto, Karin van der Wiel, and Sebastian Engelke. Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environmental Data Science*, 1:e5, 2022.

Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 442–453. Curran Associates, Inc., 2020.

Anthony L Caterini, Gabriel Loaiza-Ganem, Geoff Pleiss, and John Cunningham. Rectangular Flows for Manifold Learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30228–30241. Curran Associates, Inc., 2021.

Lawrence Cayton. Algorithms for manifold learning. Technical Report CS2008-0923, University of California, San Diego, 2005.

Lei Chen. Curse of Dimensionality. In *Encyclopedia of Database Systems*, pp. 545–546. Springer, Boston, MA, 2009.

Ming Cheng, Dimitrios G. Konstantinides, and Dingcheng Wang. Multivariate regularly varying insurance and financial risks in multidimensional risk models. *Journal of Applied Probability*, 61 (4):1319–1342, 2024.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

Stuart G. Coles and Jonathan A. Tawn. Modelling Extreme Multivariate Events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):377–392, 1991.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical Variational Auto-Encoders. In *The Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*. arXiv, 2018.

A. C. Davison and R. Huser. Statistics of Extremes. *Annual Review of Statistics and Its Application*, 2(Volume 2, 2015):203–235, 2015.

Laurens De Haan and Ana Ferreira. *Extreme Value Theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, 2006.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Holger Drees and Anne Sabourin. Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943, 2021.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events: For Insurance and Finance*. Stochastic Modelling and Applied Probability, 33. Springer Berlin Heidelberg, Berlin, Heidelberg, corr. 4. print. edition, 1997.

Sebastian Engelke and Jevgenijs Ivanovs. Sparse Structures for Multivariate Extremes. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021):241–270, 2021.

Sebastian Engelke, Manuel Hentschel, Michaël Lalancette, and Frank Röttger. Graphical models for multivariate extremes. arXiv preprint arXiv:2402.02187, 2024.

Sebastian Engelke, Nicola Gnecco, and Anne Sabourin. Extrapolation in Statistical Learning with Extreme Value Theory. arXiv preprint arXiv:2605.01909, 2026.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Allen Hatcher. *Algebraic Topology*. University Press, Cambridge, 2001.

Hans Hersbach et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Tennessee Hickling and Dennis Prangle. Flexible Tails for Normalizing Flows. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 23155–23178. PMLR, 2025.

Bruce M. Hill. A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006.

Marius Hofert and Wayne Oldford. Visualizing dependence in high-dimensional data: An application to S&P 500 constituent data. *Econometrics and Statistics*, 8:161–183, 2018.

Christian Horvat and Jean-Pascal Pfister. Denoising Normalizing Flow. In *Advances in Neural Information Processing Systems*, 2021.

Chenglei Hu and Daniela Castro-Camilo. GPDFlow: Generative Multivariate Threshold Exceedance Modeling via Normalizing Flows. arXiv preprint arXiv:2503.11822, 2025.

Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of Lipschitz Triangular Flows. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4673–4681. PMLR, 2020.

Anders Hedegaard Jessen and Thomas Mikosch. Regularly Varying Functions. *Publications de l'Institut Mathématique*, 80(94)(100):171–192, 2006.

I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, New York, second edition. edition, 2002.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The 3rd International Conference for Learning Representations*. arXiv, 2015.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*, 2014.

Juna A. Kollmeier et al. Sloan Digital Sky Survey. V. Pioneering Panoptic Spectroscopy. *The Astronomical Journal*, 171(1):52, 2025.

Nicolas Lafon, Philippe Naveau, and Ronan Fablet. A VAE approach to sample multivariate extremes. *Journal of Statistical Computation and Simulation*, 0(0):1–37, 2026.

Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Marginal Tail-Adaptive Normalizing Flows. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12020–12048. PMLR, 2022.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2012.

Stéphane Lhaut, Holger Rootzén, and Johan Segers. Simulation of multivariate extremes: A Wasserstein–Aitchison GAN approach. *Extremes*, 2026.

Filip Lindskog, Sidney I. Resnick, and Joyjit Roy. Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Probability surveys*, 11(none):270–, 2014.

Andrew McDonald, Pang-Ning Tan, and Lifeng Luo. COMET Flows: Towards Generative Modeling of Multivariate Extremes and Tail Dependence. In *Thirty-First International Joint Conference on Artificial Intelligence*, volume 4, pp. 3328–3334, 2022.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426, 2020.

Marco Avella Medina, Richard A. Davis, and Gennady Samorodnitsky. Insights into Kernel PCA with Application to Multivariate Extremes. *SIAM Journal on Mathematics of Data Science*, 7(2): 777–801, 2025.

Lambert De Monte, Raphaël Huser, Ioannis Papastathopoulos, and Jordan Richards. Generative modelling of multivariate geometric extremes using normalising flows. arXiv preprint arXiv:2505.02957, 2025.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-Tailed Diffusion Models. In *International Conference on Learning Representations*, volume 2025, pp. 10880–10934, 2025.

Alison Peard and Jim Hall. Combining deep generative models with extreme value theory for synthetic hazard simulation: A multivariate and spatially coherent approach. arXiv preprint arXiv:2311.18521, 2023.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9:Article39, 2010.

Yury Polyanskiy. *Information Theory: From Coding to Learning*. University Press, Cambridge, 2025.

Sidney I. Resnick. Discussion of the Danish Data on Large Fire Insurance Losses. *ASTIN Bulletin: The Journal of the IAA*, 27(1):139–151, 1997.

Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, 1st ed. 2007. edition, 2007.

Danilo Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1530–1538. PMLR, 2015.

Chang-Han Rhee, Jose Blanchet, and Bert Zwart. Sample path large deviations for Lévy processes and random walks with regularly varying increments. *The Annals of Probability*, 47(6):3551–3605, 2019.

Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, second edition. edition, 1964.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Stephan Schonsheck, Jie Chen, and Rongjie Lai. Chart Auto-Encoders for Manifold Structured Data. 2019.

Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-Equivariant Steerable Networks. In *International Conference on Learning Representations*, 2019.

Christopher Subich, Syed Zahid Husain, Leo Separovic, and Jing Yang. Fixing the Double Penalty in Data-Driven Weather Forecasting Through a Modified Spherical Harmonic Loss Function. In *Forty-Second International Conference on Machine Learning*, 2025.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

Jakob Benjamin Wessel, Callum J. R. Murphy-Barltrop, and Emma S. Simpson. Generative Machine Learning for Multivariate Angular Simulation. *Extremes*, 2025.

Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy. Statistical exploration of the manifold hypothesis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(2): 353–385, 2026.

Daniel Worrall and Max Welling. Deep Scale-spaces: Equivariance Over Scale. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Chiu-Wai Yan, Shi Quan Foo, Van-Hoan Trinh, Dit-Yan Yeung, Ka-Hing Wong, and Wai-Kin Wong. Fourier Amplitude and Correlation Loss: Beyond Using L2 Loss for Skillful Precipitation Nowcasting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Likun Zhang, Xiaoyu Ma, Christopher K. Wikle, and Raphaël Huser. Fast and Flexible Emulation of Spatial Extremes Processes via Variational Autoencoders. *Journal of the American Statistical Association*, 0(0):1–14, 2026.

Jakob Zscheischler and Sonia I. Seneviratne. Dependence of drivers affects risks associated with compound events. *Science Advances*, 3(6):e1700263, 2017.

A PRELIMINARIES

Recall the notation of §2. The data lie on a smooth embedded m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$ with $m < D$, the latent dimension is ℓ, and $\mathbb{O}_d := \mathbb{R}^d \setminus \{0\}$ is the punctured space. A radial functional r (Definition 3) is continuous, strictly positive away from the origin and positively homogeneous of degree one, with unit shell $\mathbb{S}_r^{d-1} := \{x \in \mathbb{O}_d : r(x) = 1\}$, and $\theta(x) := x/r(x)$ normalises x onto that shell, so that $x = r(x)\,\theta(x)$. Ambient and latent radial functionals are written r_D and r_ℓ where the distinction matters and r otherwise.

Lemma A.1 (Norm equivalence of radial functionals)**.** *For every radial functional r (Definition 3) on $\mathbb{R}^d$ there are constants $0 < c_r \le C_r < \infty$ with*

$$c_r\|x\| \le r(x) \le C_r\|x\|, \qquad x \in \mathbb{R}^d.$$

In particular the unit shell $\mathbb{S}_r^{d-1}$ is compact and bounded away from the origin.

Proof. Set $c_r := \min_{\|\xi\|=1} r(\xi)$ and $C_r := \max_{\|\xi\|=1} r(\xi)$, attained and positive by continuity and strict positivity of r on the Euclidean unit sphere. Degree-one homogeneity extends the bounds from the sphere to $\mathbb{R}^d$. On the unit shell the bounds give $1/C_r \le \|x\| \le 1/c_r$, and the shell is closed by continuity of r, hence the unit shell is compact. □

A.1 RADIAL–ANGULAR FACTORISATION OF POSITIVELY HOMOGENEOUS MAPS

Proposition 4 (Radial–angular factorisation of positively homogeneous maps, proof in Appendix A.1)**.** *A map $f : \mathbb{O}_D \to \mathbb{R}^\ell$ is positively homogeneous (Definition 2) of degree $p > 0$ if and only if there exists a unique $g : \mathbb{S}_r^{D-1} \to \mathbb{R}^\ell$ such that $f(x) = r(x)^p\, g(\theta(x))$ for $x \in \mathbb{O}_D$.*

Proof of Proposition 4. If f is positively homogeneous of degree p, set $g = f|_{\mathbb{S}_r^{D-1}}$. Since $x = r(x)\theta(x)$,

$$f(x) = r(x)^p f(\theta(x)) = r(x)^p g(\theta(x)).$$

Conversely, this factorisation is p-homogeneous because $r(\lambda x) = \lambda r(x)$ and $\theta(\lambda x) = \theta(x)$. Restricting to $r(x) = 1$ forces $g = f|_{\mathbb{S}_r^{D-1}}$, proving uniqueness. □

A.2 LATENT REPARAMETERISATION INVARIANCE

Proposition 5 (Latent reparameterisation invariance, proof in Appendix A.2)**.** *Let $f : \mathbb{R}^D \to \mathbb{R}^\ell$ and $h : \mathbb{R}^\ell \to \mathbb{R}^D$ be an encoder and decoder and let ψ be a homeomorphism of $\mathbb{R}^\ell$ onto itself. Then $\tilde{f} := \psi \circ f$ and $\tilde{h} := h \circ \psi^{-1}$ satisfy $\tilde{h}(\tilde{f}(x)) = h(f(x))$ for every $x \in \mathbb{R}^D$.*

Proof of Proposition 5. For $x \in \mathbb{R}^D$,

$$\tilde{h}(\tilde{f}(x)) = h\big(\psi^{-1}(\psi(f(x)))\big) = h(f(x)).$$

□

A.3 Magnitude under unconstrained encoders

Example A.2 (Single-neuron ReLU encoder)**.** Consider a single-neuron ReLU encoder $f : \mathbb{R}^D \to \mathbb{R}$, $f(x) = \mathrm{ReLU}(w_0 + w^\top x)$, with $w \in \mathbb{R}^D$ and $w_0 > 0$. Observations with $w_0 + w^\top x \le 0$ are all mapped to $f(x) = 0$ and points with $w_0 + w^\top x > 0$ have encoded magnitudes distorted by the bias w_0.

B Homogeneous Autoencoders

Recall the architecture of §3. The encoder (3.1) is $f(x) = r(x)\, e(\theta(x))$ for an angular map $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$, implemented as $e = E/r(E)$ for a network E. The decoder (3.2) is $h(z) = r(z)\, \tilde{c}(z)$, where

$$\tilde{c}(z) = \frac{\tilde{E}(\theta(z)) + \delta(z)}{r\big(\tilde{E}(\theta(z)) + \delta(z)\big)}, \qquad \delta(z) = \Delta\big(\theta(z), (1 + r(z))^{-1}\big) - \Delta(\theta(z), 0),$$

for a network $\tilde{E} : \mathbb{S}_r^{\ell-1} \to \mathbb{R}^D$ that is nonzero on the shell and a continuous $\Delta : \mathbb{S}_r^{\ell-1} \times [0, 1] \to \mathbb{R}^D$, and $\tilde{e} := \tilde{E}/r(\tilde{E})$ is the normalised angular decoder. The homogeneous leading term of h is $h_{\mathrm{hom}}(z) := r(z)\, \tilde{e}(\theta(z))$.

B.1 Uniform convergence of the corrected angular decoder

Lemma 6 (Uniform convergence of the corrected angular decoder, proof in Appendix B.1)**.** *Let Δ, δ and $\tilde{c}$ be as in (3.3) and (3.2). Then $\|\delta(z)\| \to 0$ uniformly in $\theta(z)$ as $r(z) \to \infty$. Moreover $\tilde{c}(z)$ is well defined whenever $r(z)$ is sufficiently large and $\|\tilde{c}(z) - \tilde{e}(\theta(z))\| \to 0$ uniformly in $\theta(z)$ as $r(z) \to \infty$.*

Proof of Lemma 6. By Lemma A.1, $\mathbb{S}_r^{\ell-1} \times [0, 1]$ is compact. Hence Δ is uniformly continuous by Heine–Cantor (Rudin, 1964, Theorem 4.19), and

$$\sup_{r(z)=\rho} \|\delta(z)\| = \sup_{\xi \in \mathbb{S}_r^{\ell-1}} \|\Delta(\xi, (1 + \rho)^{-1}) - \Delta(\xi, 0)\| \longrightarrow 0.$$

The image of the shell under $\tilde{E}$ is compact and disjoint from the origin, so it has a compact neighbourhood disjoint from the origin, on which the normalisation θ is uniformly continuous. For sufficiently large ρ, every $\tilde{E}(\theta(z)) + \delta(z)$ with $r(z) = \rho$ lies in it, so $\tilde{c}$ is well defined and

$$\sup_{r(z)=\rho} \|\tilde{c}(z) - \tilde{e}(\theta(z))\| = \sup_{r(z)=\rho} \|\theta(\tilde{E}(\theta(z)) + \delta(z)) - \theta(\tilde{E}(\theta(z)))\| \longrightarrow 0.$$

□

Corollary B.1 (Asymptotically conic decoder image)**.** *With $\tilde{c}$ and $\tilde{e}$ as in (3.2), let $C_\rho := \{\tilde{c}(z) : r(z) = \rho\}$, for ρ large enough that $\tilde{c}$ is well defined there, and $C_\infty := \{\tilde{e}(\xi) : \xi \in \mathbb{S}_r^{\ell-1}\}$. The Hausdorff distance between C_ρ and C_∞ is at most $\sup_{r(z)=\rho} \|\tilde{c}(z) - \tilde{e}(\theta(z))\|$, which tends to zero as $\rho \to \infty$ by Lemma 6.*

Proof of Corollary B.1. Every z with $r(z) = \rho$ is $\rho\, \theta(z)$, so C_ρ and C_∞ are the images of the same shell $\mathbb{S}_r^{\ell-1}$ under $\xi \mapsto \tilde{c}(\rho\xi)$ and $\xi \mapsto \tilde{e}(\xi)$. At each ξ these maps differ by $\|\tilde{c}(\rho\xi) - \tilde{e}(\xi)\|$, which is at most $\sup_{r(z)=\rho} \|\tilde{c}(z) - \tilde{e}(\theta(z))\|$. A uniform pointwise bound between two maps bounds the Hausdorff distance between their images. □

B.2 Radial and scale identities

Proposition 7 (Radial and scale identities, proof in Appendix B.2)**.** *Under well-defined shell normalisations in (3.1) and (3.2), for every $x \in \mathbb{O}_D$ and every $\lambda > 0$,*

$$f(\lambda x) = \lambda f(x), \qquad r_\ell(f(x)) = r_D(x), \qquad r_D\big(h(f(x))\big) = r_D(x).$$

Proof of Proposition 7. The encoder has the factorised form of Proposition 4 with $p = 1$ and $g = e$, so it is positively homogeneous of degree one. For the radial identity, degree-one homogeneity of r gives

$$r(f(x)) = r\big(r(x)\, e(\theta(x))\big) = r(x)\, r\big(e(\theta(x))\big) = r(x),$$

since e takes values in $\mathbb{S}_r^{\ell-1}$, on which $r \equiv 1$. For the round trip, wherever the corrected decoder normalisation is defined, the decoder identity $r(h(z)) = r(z)$ is immediate from (3.2), and composing with the radial identity gives $r(h(f(x))) = r(f(x)) = r(x)$. □

C Implementation

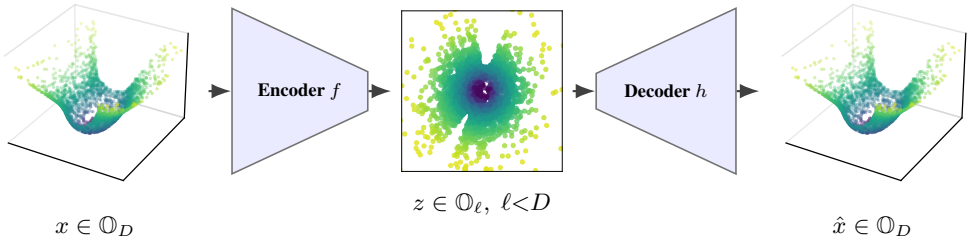


Figure C.1: Autoencoder architecture on the curved surface (D=3, m=2): input, HAE latent space and reconstruction.

C.1 Architecture and parameterisation

Figure C.1 demonstrates the architecture of Figure 2 on the curved surface.

Radial functionals. Table C.1 lists alternatives to the Euclidean norm used in the experiments.

Table C.1: Examples of homogeneous radial functionals and interpretations. Here, Σ represents a positive-definite covariance matrix.

Radial functional	Interpretation	Representative settings
$\|x\|_1$	total absolute amount	composition, calorimetry, intensities
$\|x\|_2$	Euclidean magnitude	waveforms, fields, anomalies
$\|x\|_\infty$	peak absolute magnitude	hazards, stresses, losses
weighted $\|\cdot\|_q$, $q \geq 1$	calibrated magnitude	multimodal data
$(x^\top \Sigma^{-1} x)^{1/2}$	covariance-adjusted magnitude	anomaly scoring

Encoder parameterisation. The encoder's angular map e of (3.1), which sends x to $f(x) = r(x)\, e(\theta(x))$, uses the normalised construction $e = E/r(E)$ of §3 for a network $E : \mathbb{S}_r^{D-1} \to \mathbb{R}^\ell$. The denominator $r(E)$ is clamped below by a small constant to keep it away from zero. Where the clamp is active, the radial identities hold only approximately and Proposition 7 is claimed on the domain of nondegenerate normalisations. The clamp was not active at any point of any trained model in this work. A parameterisation defined on the whole shell without a normalisation step, such as a spherical network (Cohen et al., 2018), is left to future work.

Decoder parameterisation. The decoder uses the normalised angular map $\tilde{e} = \tilde{E}/r(\tilde{E})$ of §3, for a network $\tilde{E} : \mathbb{S}_r^{\ell-1} \to \mathbb{R}^D$, together with the correction $\delta(z) = \Delta\big(\theta(z), (1+r(z))^{-1}\big) - \Delta(\theta(z), 0)$ of (3.3), where $\Delta : \mathbb{S}_r^{\ell-1} \times [0,1] \to \mathbb{R}^D$ is continuous, so that δ vanishes as $r(z) \to \infty$. It vanishes at rate $(1 + r(z))^{-1}$, since the ReLU network Δ (Appendix C.2) is Lipschitz in its scalar argument. The map $\tilde{e}$ is well defined wherever $\tilde{E}(\xi) \neq 0$ and its denominator is clamped below by the same constant as the encoder. For nonnegative fields (ERA5 precipitation) the coordinatewise absolute value is applied before normalisation for HAE and AE + radius, which both normalise onto a

shell. The absolute value is applied directly to the output for AE and AE + penalty, which do not. Since the coordinatewise absolute value is positively homogeneous of degree one and 1-Lipschitz, Lemma 6 and Propositions 7 and 10 apply unchanged, with the homogeneous leading term taken as $r(z)\,\theta(|\tilde{E}(\theta(z))|)$. Figures C.2 and C.3 illustrate the role of the correction by showing the surface's departure from its limiting directions at finite radius. A homogeneous decoder maps each latent ray to a single ambient ray and cannot represent that departure, whereas the correction admits it where accurate reconstruction requires it.

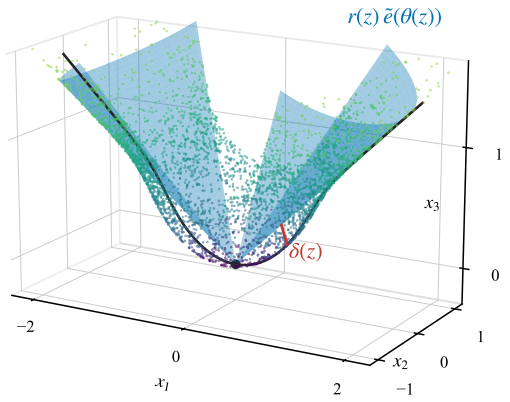


Figure C.2: The decoder correction visualised on the curved surface. The cone represents the surface's (conic) limiting angular support. The surface's cross section along $w_2 = 0$ is given in black for perspective.

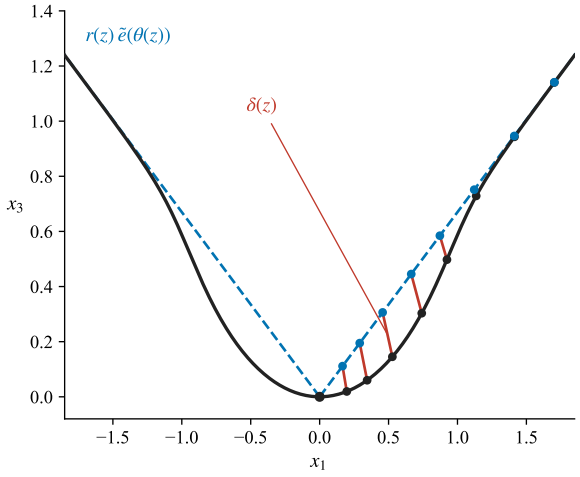


Figure C.3: The section of the curved surface along $x_2 = 0$ (black) against the cone over its limiting directions (dashed).

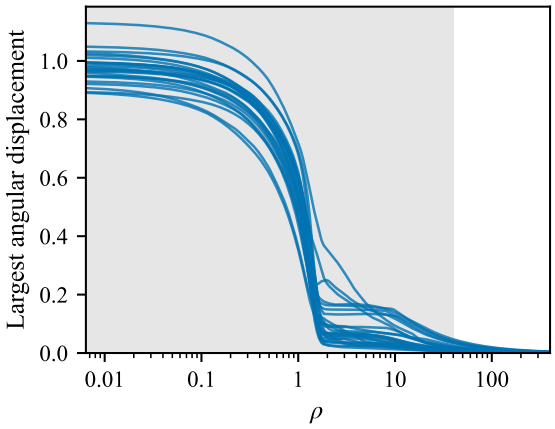


Figure C.4: Learnt HAE decoders on the curved surface, largest angular displacement against radius, one curve per seed, training radii shaded.

C.2 Training

Penalty coefficient. An AE + penalty fit was run on the curved surface data at each coefficient of Table C.2, showing reconstruction RMSE and the mean absolute radial error. Both fall as the coefficient rises until the penalty begins to dominate the reconstruction term. The experiments use 0.1 as a balance between the two terms.

Table C.2: Curved surface, AE + penalty under different penalty coefficients, one fit each.

Penalty coefficient	0.01	0.1	1
Reconstruction RMSE	0.00118	0.000937	0.00674
Mean radial error	0.00228	0.000609	0.00115

Latent dimension. Latent dimensions on the real data are the smallest explaining 0.975 of the training variance under uncentred PCA. The curved surface uses the intrinsic dimension two of its surface.

Optimisation and network sizes. All models are trained with the Adam optimiser (Kingma & Ba, 2015) at a learning rate of 3×10^{-4} and batch size 512 with early stopping on validation loss at a patience of 1,500 epochs. The angular and correction networks are standard MLPs with ReLU activations. The curved surface and its exceedances use hidden width 256 at depth 6 for the HAE networks and a matched width of 314 for the neural baselines, giving 991,752 and 992,873 parameters. The ERA5 runs use depth 3, hidden width 128 per variable and 256 for the stacked tensor, with baseline widths matched to the HAE parameter count per run. The precipitation exceedance runs use depth 2 at width 128. The networks are generally chosen to be larger than these datasets likely require, so that the parameter budget does not bind and the differences between models reflect their architectural constraints rather than their capacity. Consistent with this, the neural models reach nearly equal reconstruction error on every dataset.

D Existence of homogeneous radius-preserving embeddings

As a reminder, throughout this section $\mathcal{M} \subset \mathbb{R}^D$ is the smooth embedded m-dimensional data manifold of §2, r is a radial functional (Definition 3) with normalisation $\theta(x) = x/r(x)$ onto the unit shell $\mathbb{S}_r^{D-1} = \{x \in \mathbb{O}_D : r(x) = 1\}$, and the encoder being studied is the HAE encoder (3.1), $f(x) = r_D(x)\, e(\theta(x))$, for an angular map $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$.

D.1 FIXED-MAGNITUDE SLICES

Definition D.1 (Immersion)**.** An *immersion* is a smooth map whose differential is injective at every point, equivalently whose differential kernel is only the zero vector.

Definition D.2 (Embedding)**.** An *embedding* is an immersion (Definition D.1) that is a homeomorphism onto its image.

Definition D.3 (Diffeomorphic spaces)**.** Two spaces are *diffeomorphic* if there exists a smooth bijection with smooth inverse between them.

These definitions are standard and can be found in Lee (2012).

Let $V \subseteq \mathbb{O}_D$ be an open conic region (Definition 1) of ambient space on which r is smooth. V accommodates radial functionals that are not globally smooth and Lemma D.5 and Propositions D.7 and D.8 apply on any region where r is smooth. For the Euclidean norm used in all experiments, r is smooth on the whole of $\mathbb{O}_D$, so $V = \mathbb{O}_D$. V is relevant for piecewise-smooth choices such as $\|\cdot\|_1$ and $\|\cdot\|_\infty$ (Table C.1), allowing the following theoretical results to be interpreted in a similar piecewise fashion.

Definition D.4 (Radial regularity)**.** We call a region $U \subseteq \mathcal{M} \cap V$, open in $\mathcal{M}$, *radially regular* if the differential $d(r|_{\mathcal{M}})_x$ at x of the restriction of r to $\mathcal{M}$ is nonzero for every $x \in U$.

Radial regularity means that at every point of the region the chosen magnitude varies to first order along some tangent direction of the data manifold. Write $U_\rho = \{x \in U : r(x) = \rho\}$ for the fixed-magnitude slice of U at ρ.

Lemma D.5 (Fixed-magnitude slices and lossless normalisation)**.** *Let U be radially regular (Definition D.4). Then every nonempty U_ρ is a smooth embedded $(m-1)$-dimensional submanifold of $\mathcal{M}$ and for every $\rho \in r(U)$ the restricted normalisation*

$$\theta|_{U_\rho} : U_\rho \to V \cap \mathbb{S}_r^{D-1}$$

is a smooth embedding onto its image.

Proof of Lemma D.5. For the first conclusion, the restricted map $r|_U : U \to (0,\infty)$ is smooth and has nonzero differential at every point by radial regularity. Hence every $\rho \in r(U)$ is a regular value and the regular level set theorem (Lee, 2012, Corollary 5.14) gives that U_ρ is a smooth embedded $(m-1)$-dimensional submanifold of $\mathcal{M}$.

For the second, fix $\rho \in r(U)$. Every $x \in U_\rho$ has $r(x) = \rho$, so $\theta(x) = x/\rho$ on U_ρ and $\theta|_{U_\rho}$ is the restriction to U_ρ of the dilation $x \mapsto x/\rho$ of $\mathbb{R}^D$. That dilation is a linear diffeomorphism and U_ρ is an embedded submanifold by the first conclusion, so the restriction is a smooth embedding onto its image. The image lies in $V \cap \mathbb{S}_r^{D-1}$ because V is conic and $r(x/\rho) = r(x)/\rho = 1$. □

D.2 EXISTENCE OF OVERCOMPLETE HAE EMBEDDINGS

Proposition 8 (Existence of overcomplete HAE embeddings, proof in Appendix D.2)**.** *Let $r_D : \mathbb{O}_D \to (0,\infty)$ and $r_\ell : \mathbb{O}_\ell \to (0,\infty)$ be smooth radial functionals (Definition 3), and let $U \subseteq \mathcal{M} \cap \mathbb{O}_D$ be an open subset of the data manifold of §2. If $2m+2 \le \ell < D$, then there exists a smooth angular map $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$ such that the HAE encoder (3.1), $f(x) = r_D(x)\, e(\theta(x))$, is an injective smooth immersion (Definition D.1) of U into $\mathbb{O}_\ell$ and satisfies $r_\ell(f(x)) = r_D(x)$ for every $x \in U$. Thus $f|_K$ is a smooth embedding (Definition D.2) for every compact embedded submanifold $K \subset U$.*

Proof of Proposition 8. Since $\ell - 1 \ge 2m+1$, Whitney's linear-projection argument applied iteratively gives a linear map

$$J : \mathbb{R}^D \to \mathbb{R}^{\ell-1}$$

whose restriction to U is an injective immersion (Lee, 2012, Lemma 6.13). Throughout this proof, $[\cdot,\cdot]$ denotes concatenation of a vector in $\mathbb{R}^{\ell-1}$ with a scalar into a column vector of $\mathbb{R}^\ell$, so that $[J\xi, 1] = \big((J\xi)^\top, 1\big)^\top$. Define

$$e(\xi) = \frac{[J\xi, 1]}{r_\ell\big([J\xi, 1]\big)}, \qquad \xi \in \mathbb{S}_r^{D-1}.$$

Since the last coordinate of $[J\xi, 1]$ is nonzero, e is nondegenerate and thus smooth and takes values in $\mathbb{S}_r^{\ell-1}$. Hence

$$f(x) = r_D(x)e(\theta(x))$$

is smooth and satisfies $r_\ell(f(x)) = r_D(x)$ (Proposition 7).

To show that $f|_U$ is an injective immersion, note that by linearity $J\theta(x) = Jx/r_D(x)$, so

$$\big[J\theta(x), 1\big] = \frac{\big[Jx,\ r_D(x)\big]}{r_D(x)},$$

and substituting into $f(x) = r_D(x)\, e(\theta(x))$, we have

$$f(x) = r_D(x)\, \frac{\big[Jx,\ r_D(x)\big]}{r_\ell\big([Jx,\ r_D(x)]\big)}, \qquad x \in U.$$

Let the last vector component of $f(x)$ be written $f_\ell(x)$:

$$f_\ell(x) = \frac{r_D(x)^2}{r_\ell\big([Jx,\ r_D(x)]\big)} > 0,$$

then dividing $f(x)$ by it and using $r_\ell(f(x)) = r_D(x)$,

$$\frac{f(x)}{f_\ell(x)} = \frac{\big[Jx,\ r_D(x)\big]}{r_D(x)}, \qquad \text{that is} \qquad \big[Jx,\ r_D(x)\big] = r_\ell(f(x))\, \frac{f(x)}{f_\ell(x)}.$$

The right-hand side is smooth in $f(x)$, since r_ℓ is smooth on $\mathbb{O}_\ell$ and $f_\ell(x) > 0$. Hence $f(x) = f(y)$ gives $Jx = Jy$, and $df_x(v) = 0$ gives $Jv = 0$ by the chain rule. Since $J|_U$ is an injective immersion, this means that $f(x) = f(y)$ implies $x = y$ and $df_x(v) = 0$ implies $v = 0$, so $f|_U$ is an injective smooth immersion.

Finally, its restriction to any compact embedded submanifold $K \subset U$ is a continuous injection from a compact space into the Hausdorff space $\mathbb{O}_\ell$, hence a homeomorphism onto its image, and therefore a smooth embedding. □

D.3 Minimal latent dimension

Proposition D.6 (Topological requirements at the minimal latent dimension)**.** *Let $\ell = m$, let U be radially regular (Definition D.4) and let the HAE encoder (3.1), $f(x) = r(x)\, e(\theta(x))$, be a smooth embedding (Definition D.2) of U into $\mathbb{O}_m$. Then U is diffeomorphic (Definition D.3) to an open subset of $\mathbb{O}_m$, and for every $\rho \in r(U)$, the image $(e \circ \theta)(U_\rho)$ of the fixed-magnitude slice U_ρ is an open subset of the latent shell $\mathbb{S}_r^{m-1}$. Moreover, if $K \subseteq U_\rho$ is a nonempty compact, connected, boundaryless embedded $(m-1)$-submanifold, then $(e \circ \theta)(K)$ is a connected component of $\mathbb{S}_r^{m-1}$.*

Proof of Proposition D.6. Because f is a smooth embedding and $\ell = m$, invariance of domain (Hatcher, 2001, Theorem 2B.3) implies that $f(U)$ is open in $\mathbb{O}_\ell$. Thus f identifies U diffeomorphically with an open subset of the latent space.

For a fixed magnitude ρ, every $x \in U_\rho$ satisfies $f(x) = \rho\, e(\theta(x))$, so $f|_{U_\rho}$ is $e \circ \theta$ followed by the dilation $z \mapsto \rho z$, and $e \circ \theta$ is therefore an embedding of the $(m-1)$-dimensional slice into the $(\ell-1)$-dimensional shell $\mathbb{S}_r^{\ell-1}$. These dimensions agree, so invariance of domain again gives an open image.

Let $K \subseteq U_\rho$ be compact, connected and boundaryless, as in the proposition. Its image $e(\theta(K))$ is open in $\mathbb{S}_r^{\ell-1}$ by the preceding argument, and compact, hence closed, because K is compact and $\mathbb{S}_r^{\ell-1}$ is Hausdorff. A nonempty subset which is both open and closed is a union of connected components and is a single component because K is connected. □

The minimal latent dimension $\ell = m$ leaves the encoder the least topological freedom. For example, when $m = \ell = 2$, a slice consisting of two disjoint circular components cannot be embedded into a single circle. Restricting to a local region of the data or choosing a different origin may avoid this obstruction, while increasing ℓ gives the angular map more dimensions in which to separate each slice. When $\ell > m$ the latent law remains singular with respect to Lebesgue measure on $\mathbb{R}^\ell$, which may matter for downstream density modelling. The conclusion that U is diffeomorphic to an open subset of $\mathbb{O}_m$ is the HAE analogue of the requirement that a standard autoencoder at minimal latent dimension identify the data region with an open subset of $\mathbb{R}^m$ (Schonsheck et al., 2019).

D.4 Immersion and radial-fibre criteria

Proposition 8 constructs an admissible angular map at a sufficient latent dimension. The results below test how a given map interacts with manifold geometry, giving conditions under which the encoder is an immersion or an embedding. This makes precise the observation of §4 that, since the radius is already preserved, the angular map only has to separate observations of equal magnitude.

Proposition D.7 (Immersion criterion for HAE)**.** *Let $U \subseteq \mathcal{M} \cap V$ be open in $\mathcal{M}$, fix $x \in U$ and let e be differentiable at $\theta(x)$ and r_ℓ be differentiable at $f(x)$. The HAE encoder (3.1), $f(x) = r(x)\, e(\theta(x))$, is an immersion (Definition D.1) at x if and only if*

$$\ker d(r|_{\mathcal{M}})_x \cap \ker d(e \circ \theta)_x = \{0\}.$$

Proof of Proposition D.7. For $v \in T_x\mathcal{M}$, the product rule gives

$$df_x(v) = d(r_D|_{\mathcal{M}})_x(v)\, e(\theta(x)) + r_D(x)\, d(e \circ \theta)_x(v).$$

If $d(r_D|_{\mathcal{M}})_x(v) = 0$ and $d(e \circ \theta)_x(v) = 0$, both terms vanish and $df_x(v) = 0$. Conversely, differentiating the radial identity $r_\ell \circ f = r_D$ of Proposition 7 shows that $df_x(v) = 0$ implies $d(r_D|_{\mathcal{M}})_x(v) = 0$, leaving $r_D(x)\, d(e \circ \theta)_x(v) = 0$, so $d(e \circ \theta)_x(v) = 0$ since $r_D(x) > 0$. Thus $\ker df_x = \ker d(r_D|_{\mathcal{M}})_x \cap \ker d(e \circ \theta)_x$, which is the zero vector exactly when f is an immersion at x. □

On radially regular regions, this gives an equivalent criterion in terms of fixed-magnitude slices.

Proposition D.8 (Radial-fibre criterion)**.** *Let $U \subseteq \mathcal{M} \cap V$ be radially regular (Definition D.4), and let $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$ be smooth on $V \cap \mathbb{S}_r^{D-1}$. Then the HAE encoder (3.1), $f(x) = r(x)e(\theta(x))$, is an injective smooth immersion (Definition D.1) on U if and only if $e \circ \theta|_{U_\rho}$ is an injective immersion on every fixed-magnitude slice U_ρ, $\rho \in r(U)$. Such an f is a smooth embedding if it is proper, meaning preimages of compact sets are compact.*

Proof. The radial identity $r_\ell(f(x)) = r_D(x)$ of Proposition 7 separates points of different magnitudes, while $f|_{U_\rho} = \rho\,(e \circ \theta)|_{U_\rho}$. Thus injectivity is equivalent to injectivity on every slice. Radial regularity gives $T_x U_{r(x)} = \ker d(r|_{\mathcal{M}})_x$ (Lemma D.5), so Proposition D.7 gives the corresponding equivalence for immersion. The embedding conclusion follows because a proper injective immersion is a smooth embedding (Lee, 2012, Proposition 4.22). □

D.5 Near-manifold data

The manifold hypothesis of §2 is an idealised view of low-dimensional structure and is not considered to hold directly in practice. Write $X = X_{\mathcal{M}} + \varepsilon$ for an observation, with $X_{\mathcal{M}} \in \mathcal{M}$ and a perturbation ε carrying it off the manifold (Fefferman et al., 2016).

The architectural guarantees of HAE are unaffected. Proposition 7 holds pointwise for every input in $\mathbb{O}_D$, whether or not that input actually lies on $\mathcal{M}$, so $r_\ell(f(X)) = r_D(X)$ and $f(\lambda X) = \lambda f(X)$ continue to hold exactly. Noise changes the interpretation of magnitude rather than the identity. HAE preserves the observed magnitude $r(X)$, not the magnitude $r(X_{\mathcal{M}})$ of the underlying point on the manifold. The gap between $r(X)$ and $r(X_{\mathcal{M}})$ is greatest in the bulk and becomes negligible in the tail, which is where the extreme-value results operate. A noise-dominated radius, mentioned in the limitations of §8, is primarily a statement about the bulk of the data rather than about its extremes. The extreme-value results assume a regularly varying input, which is a statement about the underlying $X_{\mathcal{M}}$ rather than about what the encoder is given. If $\mathbb{P}(\|\varepsilon\| > x) = o\big(\mathbb{P}(\|X_{\mathcal{M}}\| > x)\big)$ as $x \to \infty$, then X inherits the regular variation (Definition E.1) of $X_{\mathcal{M}}$, with the same index and tail measure (Jessen & Mikosch, 2006, Lemma 3.12), and Propositions 9 and 10 and Corollary E.4 apply to the observations themselves.

E Consequences for extreme-value analysis

Definition E.1 (Regular variation)**.** Let $\mathbb{M}(\mathbb{O}_d)$ denote the space of Borel measures on $\mathbb{O}_d$ finite on sets bounded away from the origin, and convergence $\nu_t \to \nu$ in $\mathbb{M}(\mathbb{O}_d)$ mean $\nu_t(\chi) \to \nu(\chi)$ for every bounded continuous χ vanishing on a neighbourhood of the origin (Lindskog et al., 2014). Then an

$\mathbb{O}_d$-valued random vector X is *regularly varying* if there exist a scaling function $b : (0,\infty) \to (0,\infty)$ with $b(t) \to \infty$ and a nonzero measure $\mu \in \mathbb{M}(\mathbb{O}_d)$ such that

$$t\,\mathbb{P}\left(\frac{X}{b(t)} \in \cdot\right) \xrightarrow[t\to\infty]{} \mu(\cdot) \qquad \text{in } \mathbb{M}(\mathbb{O}_d).$$

The limit μ is the *tail measure*, called the *ambient tail measure* when $d = D$, and describes the limiting dependence structure of extreme events. It satisfies $\mu(\lambda A) = \lambda^{-\alpha}\mu(A)$ for some $\alpha > 0$ and every $\lambda > 0$ and measurable $A \subset \mathbb{O}_d$ bounded away from the origin (Resnick, 2007; Lindskog et al., 2014), and α is the *index* of regular variation.

E.1 Tail transport under the encoder

Proposition 9 is a case of the following, which applies to any positively homogeneous map and so covers both the encoder and the homogeneous leading term of the decoder.

Proposition E.2 (Tail transport under positively homogeneous maps)**.** *Let X be regularly varying (Definition E.1) on $\mathbb{O}_d$ with scaling function b, index $\alpha > 0$ and tail measure μ, and let $g : \mathbb{O}_d \to \mathbb{O}_{d'}$ be measurable, continuous μ-almost everywhere and positively homogeneous (Definition 2) of degree one, with $g^{-1}(A)$ bounded away from the origin of $\mathbb{O}_d$ for every measurable $A \subset \mathbb{O}_{d'}$ bounded away from the origin of $\mathbb{O}_{d'}$. Then $g(X)$ is regularly varying on $\mathbb{O}_{d'}$ with the same scaling function b, the same tail index α and tail measure $g_{\#}\mu$.*

Proposition 9 (Regular variation under the HAE encoder, proof in Appendix E.1)**.** *Let X be regularly varying (Definition E.1) on $\mathbb{O}_D$ with scaling function b, index $\alpha > 0$ and tail measure μ. Let $f : \mathbb{O}_D \to \mathbb{O}_\ell$ be an HAE encoder (3.1) that is measurable and continuous μ-almost everywhere. Then $f(X)$ is regularly varying on $\mathbb{O}_\ell$ with the same scaling function b, the same tail index α and tail measure $f_{\#}\mu$.*

Proof of Proposition E.2 and Proposition 9. Write $\nu_t = t\,\mathbb{P}(X/b(t) \in \cdot)$. g satisfies the hypotheses of the mapping theorem for $\mathbb{M}$-convergence (Lindskog et al., 2014, Theorem 2.3), under which $\nu \mapsto g_{\#}\nu$ is continuous at μ, so by homogeneity

$$t\,\mathbb{P}\left(\frac{g(X)}{b(t)} \in \cdot\right) = g_{\#}\nu_t \longrightarrow g_{\#}\mu \qquad \text{in } \mathbb{M}(\mathbb{O}_{d'}).$$

The pushforward is nonzero because g takes values in $\mathbb{O}_{d'}$ and $\mu \neq 0$. For any measurable A bounded away from the origin of $\mathbb{O}_{d'}$,

$$g^{-1}(\lambda A) = \lambda\, g^{-1}(A), \qquad (g_{\#}\mu)(\lambda A) = \lambda^{-\alpha}(g_{\#}\mu)(A).$$

The HAE encoder is the case $g = f$, where the bounded-preimage hypothesis follows from $r_\ell(f(x)) = r_D(x)$ of Proposition 7 and Lemma A.1. □

E.2 Tail equivalence under the decoder

Definition E.3 (Tail equivalence)**.** Random vectors Y and $\widetilde{Y}$ are *tail-equivalent* with scaling function b and tail measure μ_0 if both $t\,\mathbb{P}(Y/b(t) \in \cdot) \to \mu_0$ and $t\,\mathbb{P}(\widetilde{Y}/b(t) \in \cdot) \to \mu_0$ in $\mathbb{M}(\mathbb{O}_D)$.

In the proposition below, $h(z) = r(z)\,\tilde{c}(z)$ is the decoder (3.2) with corrected angular part $\tilde{c}$, and $\tilde{e} = \tilde{E}/r(\tilde{E})$ is its homogeneous angular map, so that $h_{\mathrm{hom}}(z) = r(z)\,\tilde{e}(\theta(z))$ is the homogeneous leading term.

Proposition 10 (Tail equivalence under asymptotically homogeneous decoding, proof in Appendix E.2)**.** *Let Z be regularly varying (Definition E.1) on $\mathbb{O}_\ell$ with scaling function b_Z, tail index $\beta > 0$ and tail measure μ_Z. Assume that $\tilde{e}$ is continuous and that the decoder h of (3.2) is well defined on $\mathbb{O}_\ell$ with $\|\tilde{c}(z) - \tilde{e}(\theta(z))\| \to 0$ uniformly in $\theta(z)$ as $r(z) \to \infty$ (Lemma 6). Then $h(Z)$ and $h_{\mathrm{hom}}(Z)$ are tail-equivalent with scaling function b_Z, common tail index β and common tail measure $(h_{\mathrm{hom}})_{\#}\mu_Z$.*

Proof of Proposition 10. h_{hom} is continuous, positively homogeneous of degree one and satisfies $r(h_{\mathrm{hom}}(z)) = r(z)$, so by the norm equivalence of Lemma A.1 and the transport of Proposition E.2, $h_{\mathrm{hom}}(Z)$ is regularly varying with scaling b_Z, index β and tail measure $(h_{\mathrm{hom}})_{\#}\mu_Z$.

By the asymptotic-equivalence result for $\mathbb{M}$-convergence (Rhee et al., 2019, Definition 1 and Corollary 2.1), it suffices to show $t\,\mathbb{P}(F_t) \to 0$ for every $\gamma > 0$, where F_t is the event

$$F_t := \{\|h(Z) - h_{\hom}(Z)\| > \gamma b_Z(t)\}.$$

The ambient shell is compact by Lemma A.1, so $\sup_{\xi,\xi' \in \mathbb{S}_r^{D-1}} \|\xi - \xi'\| < \infty$. Write $\bar{\delta}(\rho) := \sup_{r(z)=\rho} \|\tilde{c}(z) - \tilde{e}(\theta(z))\|$ for the largest angular displacement at radius ρ. Then (3.2) gives

$$\|h(z) - h_{\hom}(z)\| \le r(z)\,\bar{\delta}(r(z)), \qquad \bar{\delta}(\rho) \le \sup_{\xi,\xi' \in \mathbb{S}_r^{D-1}} \|\xi - \xi'\|.$$

Fix $\varepsilon > 0$ and choose ρ_0 such that $\bar{\delta}(\rho) \le \varepsilon$ for $\rho \ge \rho_0$ (Lemma 6). On $\{r(Z) < \rho_0\}$, since $b_Z(t) \to \infty$,

$$\|h(Z) - h_{\hom}(Z)\| \le \rho_0 \left(\sup_{\xi,\xi' \in \mathbb{S}_r^{D-1}} \|\xi - \xi'\| \right) \le \gamma\, b_Z(t) \qquad \text{for all large } t,$$

so $F_t \cap \{r(Z) < \rho_0\} = \emptyset$. On $F_t \cap \{r(Z) \ge \rho_0\}$,

$$\gamma\, b_Z(t) < \|h(Z) - h_{\hom}(Z)\| \le r(Z)\,\bar{\delta}(r(Z)) \le \varepsilon\, r(Z).$$

Hence, for all large t, homogeneity of r gives

$$F_t \subseteq \left\{ r(Z) > \frac{\gamma}{\varepsilon} b_Z(t) \right\} = \left\{ r\left(\frac{Z}{b_Z(t)} \right) > \frac{\gamma}{\varepsilon} \right\},$$

and therefore

$$\begin{aligned} t\,\mathbb{P}(F_t) &\le t\,\mathbb{P}\left(r\left(\frac{Z}{b_Z(t)} \right) > \frac{\gamma}{\varepsilon} \right) \\ &= t\,\mathbb{P}\left(\frac{Z}{b_Z(t)} \in r^{-1}\left(\left(\frac{\gamma}{\varepsilon}, \infty \right) \right) \right) \\ &= t\,\mathbb{P}\left(\frac{Z}{b_Z(t)} \in \left\{ z : r(z) > \frac{\gamma}{\varepsilon} \right\} \right) \\ &\le t\,\mathbb{P}\left(\frac{Z}{b_Z(t)} \in \left\{ z : r(z) \ge \frac{\gamma}{\varepsilon} \right\} \right). \end{aligned}$$

By the regular variation of Z, $t\,\mathbb{P}(Z/b_Z(t) \in \cdot) \to \mu_Z$ in $\mathbb{M}(\mathbb{O}_\ell)$ (Definition E.1), and $\{z : r(z) \ge \gamma/\varepsilon\}$ is closed and bounded away from the origin (Lemma A.1), so the Portmanteau upper bound (Lindskog et al., 2014, Theorem 2.1) and homogeneity of μ_Z give

$$\limsup_{t\to\infty} t\,\mathbb{P}(F_t) \le \mu_Z\left(\left\{ z : r(z) \ge \frac{\gamma}{\varepsilon} \right\} \right) = \left(\frac{\varepsilon}{\gamma} \right)^{\beta} \mu_Z(\{z : r(z) \ge 1\}) < \infty.$$

Letting $\varepsilon \downarrow 0$ gives $t\,\mathbb{P}(F_t) \to 0$ for every $\gamma > 0$. □

E.3 Round-trip tail transport

Corollary E.4 (Round-trip tail transport). *Let X be regularly varying (Definition E.1) on $\mathbb{O}_D$ with scaling function b, tail index $\alpha > 0$ and tail measure μ. Let f be an HAE encoder (3.1) that is measurable and continuous μ-almost everywhere, as in Proposition 9, and let the decoder h of (3.2) satisfy Proposition 10 for $Z = f(X)$, with homogeneous leading term $h_{\hom}(z) = r(z)\,\tilde{e}(\theta(z))$. Then $h(f(X))$ is regularly varying on $\mathbb{O}_D$ with scaling function b, tail index α and tail measure $(h_{\hom} \circ f)_{\#}\mu$.*

Proof of Corollary E.4. By Proposition 9, $Z = f(X)$ has scaling b, tail index α and tail measure $f_{\#}\mu$. Applying Proposition 10 gives the same scaling and index for $h(Z)$, with tail measure

$$(h_{\hom})_{\#}(f_{\#}\mu) = (h_{\hom} \circ f)_{\#}\mu.$$

□

E.4 PRESERVATION OF RADIAL–ANGULAR FACTORISATION

Proposition 11 (Preservation of factorisation, proof in Appendix E.4)**.** *Fix $u > 0$. A measurable radius-preserving encoder $f : \mathbb{O}_D \to \mathbb{O}_\ell$ with $r_\ell(f(x)) = r_D(x)$ preserves independence of radius and direction conditional on $r(X) > u$, for every law of X on $\mathbb{O}_D$ with positive exceedance probability, if and only if $f(x) = r(x)\, e(\theta(x))$ whenever $r(x) > u$, for some measurable $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{\ell-1}$. For such encoders, including HAE (3.1), the total-variation distance of the conditional radial–angular law from the product of its marginals cannot increase.*

Proof of Proposition 11. Let X be any $\mathbb{O}_D$-valued random vector with $\mathbb{P}(r(X) > u) > 0$ and write P for the conditional law of $(r(X), \theta(X))$ given $r(X) > u$, with marginals P_r and P_θ, so that $r(X)$ and $\theta(X)$ are conditionally independent exactly when $P = P_r \otimes P_\theta$. Under the representation $f(x) = r(x)\, e(\theta(x))$, which is the HAE encoder (3.1) above the threshold, radius preservation $r(f(x)) = r(x)$ (Proposition 7) gives

$$\theta(f(x)) = \frac{f(x)}{r(f(x))} = \frac{r(x)\, e(\theta(x))}{r(x)} = e(\theta(x)),$$

so $\{r(f(X)) > u\} = \{r(X) > u\}$ and, defining $T(\rho, \xi) := \big(\rho, e(\xi)\big)$,

$$\big(r(f(X)), \theta(f(X))\big) = T\big(r(X), \theta(X)\big).$$

The encoded conditional law is $T_\# P$. Since $T = \mathrm{id} \times e$,

$$T_\# P\big(\cdot \times \mathbb{S}_r^{\ell-1}\big) = P_r, \qquad T_\# P\big((u, \infty) \times \cdot\big) = e_\# P_\theta,$$
$$T_\#\big(P_r \otimes P_\theta\big) = P_r \otimes e_\# P_\theta.$$

If $P = P_r \otimes P_\theta$ then $T_\# P = P_r \otimes e_\# P_\theta$, so f preserves independence of radius and direction.

Total-variation distance does not increase under a common measurable map (Polyanskiy, 2025, Theorem 7.4), so

$$\sup_A \big|T_\# P(A) - \big(P_r \otimes e_\# P_\theta\big)(A)\big| \le \sup_A \big|P(A) - \big(P_r \otimes P_\theta\big)(A)\big|$$

over measurable sets A.

Conversely, suppose that for every law of X on $\mathbb{O}_D$ with $\mathbb{P}(r(X) > u) > 0$, if $r(X)$ and $\theta(X)$ are independent conditional on $r(X) > u$ then so are $r(f(X))$ and $\theta(f(X))$. For $\xi \in \mathbb{S}_r^{D-1}$ and $\rho > u$, degree-one homogeneity of r (Definition 3) gives $r(\rho\xi) = \rho$, so radius preservation gives $r(f(\rho\xi)) = \rho$ and

$$f(\rho\xi) = r\big(f(\rho\xi)\big)\, \theta\big(f(\rho\xi)\big) = \rho\, \theta\big(f(\rho\xi)\big).$$

It is therefore enough to show that $\theta(f(\rho\xi))$ takes the same value at every $\rho > u$. Fix $\xi \in \mathbb{S}_r^{D-1}$ and distinct $\rho_1, \rho_2 > u$, and let

$$\mathbb{P}(X = \rho_1\xi) = \mathbb{P}(X = \rho_2\xi) = \frac{1}{2}.$$

Since here $\mathbb{P}(r(X) > u) = 1$, $\theta(X) = \xi$ is constant, $r(X)$ and $\theta(X)$ are independent, and the hypothesis then makes $r(f(X))$ and $\theta(f(X))$ independent.

Suppose $\theta(f(\rho_1\xi)) \neq \theta(f(\rho_2\xi))$. Radius preservation gives $r(f(X)) = r(X)$, so $r(f(X))$ takes the values ρ_1 and ρ_2 at the two points, while $\theta(f(X))$ takes $\theta(f(\rho_1\xi))$ and $\theta(f(\rho_2\xi))$. Both pairs are distinct, so the following events are equal

$$\big\{r(f(X)) = \rho_1\big\} = \{X = \rho_1\xi\} = \big\{\theta(f(X)) = \theta(f(\rho_1\xi))\big\}.$$

Independence of $r(f(X))$ and $\theta(f(X))$ gives

$$\mathbb{P}\big(r(f(X)) = \rho_1,\ \theta(f(X)) = \theta(f(\rho_1\xi))\big) = \mathbb{P}\big(r(f(X)) = \rho_1\big)\, \mathbb{P}\big(\theta(f(X)) = \theta(f(\rho_1\xi))\big).$$

Each of these three events is $\{X = \rho_1\xi\}$, so

$$\mathbb{P}(X = \rho_1\xi) = \mathbb{P}(X = \rho_1\xi)^2,$$

which clearly fails plugging in $\mathbb{P}(X = \rho_1\xi) = 1/2$. Hence $\theta(f(\rho_1\xi)) = \theta(f(\rho_2\xi))$.

Since ξ, ρ_1 and ρ_2 were arbitrary, $\theta(f(\rho\xi))$ takes the same value at every $\rho > u$, on every ray. Define

$$e(\xi) := \theta\big(f((u+1)\xi)\big),$$

measurable as a composition of measurable maps and valued in $\mathbb{S}_r^{\ell-1}$. For $r(x) > u$, radius preservation and $x = r(x)\,\theta(x)$ give

$$f(x) = r(x)\,\theta(f(x)) = r(x)\,e(\theta(x)).$$

□

E.5 Tail-index versus full tail-measure recovery

Full recovery of the ambient tail measure would require the round trip to satisfy $(h_{\mathrm{hom}} \circ f)_{\#}\mu = \mu$, flagged after Corollary E.4. One way to recover the tail measure exactly is to combine the radial–angular factorisation with an approximately invertible architecture. Normalising flows (Rezende & Mohamed, 2015) and other bijective architectures are globally invertible and therefore incompatible with dimension reduction. Alternative constructions such as manifold-flow and related architectures (Brehmer & Cranmer, 2020; Caterini et al., 2021) relax global invertibility and examine invertibility over restriction to a learnt manifold and concentrate density there, providing a promising route towards full tail-measure recovery.

E.6 Degree-p extension

The arguments in this paper carry over to a general homogeneity degree. For $p > 0$, define

$$f_p(x) = r(x)^p\,e(\theta(x)), \qquad h_p(z) = r(z)^{1/p}\,\tilde{c}(z),$$

so that $r(f_p(x)) = r(x)^p$ and $r(h_p(f_p(x))) = r(x)$. If X is regularly varying (Definition E.1) with tail index α and scaling b, then $f_p(X)$ has index α/p and scaling b^p, and the degree-$1/p$ decoder restores index α under the round trip. Exact magnitude preservation is the case $p = 1$, where the index is left unchanged throughout.

F Models and metrics

HAE is the homogeneous autoencoder of §3. AE is a standard autoencoder (Rumelhart et al., 1986; Hinton & Salakhutdinov, 2006). AE + penalty adds the term $0.1\,\mathbb{E}[(r(f(X)) - r(X))^2]$ to the AE loss, so it incentivises but does not enforce preservation of magnitude. AE + radius is the construction of Remark 12, $x \mapsto r(x)\,e_{\mathrm{rad}}(x)$, whose angular network takes the full coordinate x rather than only its direction $\theta(x)$ as HAE does. AE + radius preserves the radius exactly but lets the latent direction vary with it and is thus not constrained to be homogeneous. PCA (Jolliffe, 2002) is used uncentred, hence is positively homogeneous (Definition 2). All share the latent dimension and the neural models are parameter-matched. Throughout this section $\{x_i\}_{i=1}^n$ is a held-out test split of the data, f and h are the encoder and decoder of the relevant model, $z_i := f(x_i)$ is the latent representation of the ith observation and $\hat{x}_i := h(f(x_i))$ its reconstruction.

Reconstruction RMSE.

$$\mathrm{RMSE} = \left(\frac{1}{nD}\sum_{i=1}^{n}\|x_i - \hat{x}_i\|^2\right)^{1/2},$$

averaging over observations and coordinates. Training optimises the mean squared error.

Hill estimator and threshold choice. Write $Y_{(1)} \geq \cdots \geq Y_{(n)}$ for the descending order statistics of the n radii of a split, either the ambient $r(x_i)$ or the latent $r(z_i)$. The Hill estimate of the tail index (Hill, 1975) from the largest k of them is

$$\widehat{\alpha}_k = \left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{Y_{(i)}}{Y_{(k+1)}}\right)^{-1}, \qquad 1 \leq k < n.$$

We use $k = \lfloor 0.05n \rfloor$ to compare models and spaces at a common tail fraction, and plot the curves over k/n from 0.005 to 0.25 to show the Hill estimator's behaviour over a range of k. $k = \lfloor 0.05n \rfloor$ is chosen as a heuristic tail fraction and is not claimed to be the optimal threshold for tail estimation in all cases. The purpose of presenting the Hill tail estimator is not to directly provide tail diagnostics, but to demonstrate how a standard autoencoder's radial distortion may distort the *behaviour* of magnitude-dependent analysis in latent space.

Hill estimate difference. The absolute difference $|\widehat{\alpha}_{\mathrm{lat}} - \widehat{\alpha}_{\mathrm{amb}}|$ between the Hill estimates from latent and ambient radii at the same k and n.

Mean absolute radial error. The mean absolute difference between latent and ambient radii, $\frac{1}{n}\sum_i |r(z_i) - r(x_i)|$.

Rank correlation. Let R_i^x be the rank of $r(x_i)$ within the ambient radii $\{r(x_j)\}_{j=1}^n$ and R_i^z the rank of $r(z_i)$ within the latent radii $\{r(z_j)\}_{j=1}^n$. The Spearman rank correlation is the Pearson correlation between these ranks,

$$\frac{\sum_{i=1}^n (R_i^x - \bar{R}^x)(R_i^z - \bar{R}^z)}{\left(\sum_{i=1}^n (R_i^x - \bar{R}^x)^2\right)^{1/2}\left(\sum_{i=1}^n (R_i^z - \bar{R}^z)^2\right)^{1/2}}, \qquad \bar{R}^x = \bar{R}^z = \frac{n+1}{2},$$

which is equal to one when the encoder preserves the ordering of the radii.

Equivariance error. The relative encoder equivariance error at scale λ is

$$\frac{1}{n}\sum_{i=1}^n \frac{\|f(\lambda x_i) - \lambda f(x_i)\|}{\lambda \|f(x_i)\|},$$

evaluated over a fixed-seed subsample of 2,048 held-out test data.

Energy distance. Let F and G be probability distributions on $\mathbb{R}^d$ with finite first moments, and let $U, U' \overset{\text{iid}}{\sim} F$ and $V, V' \overset{\text{iid}}{\sim} G$. The energy distance (Székely & Rizzo, 2013) between F and G is

$$\mathcal{E}(F, G) = 2\,\mathbb{E}\|U - V\| - \mathbb{E}\|U - U'\| - \mathbb{E}\|V - V'\|,$$

which is zero when $F = G$. For finite samples $\{a_i\}_{i=1}^{n_a}$ and $\{b_j\}_{j=1}^{n_b}$ of points in $\mathbb{R}^d$, $\mathcal{E}$ is estimated by

$$\widehat{\mathcal{E}}\big(\{a_i\}_{i=1}^{n_a}, \{b_j\}_{j=1}^{n_b}\big) = \frac{2}{n_a n_b}\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\|a_i - b_j\| - \frac{1}{n_a^2}\sum_{i,j=1}^{n_a}\|a_i - a_j\| - \frac{1}{n_b^2}\sum_{i,j=1}^{n_b}\|b_i - b_j\|.$$

the estimator's arguments are subsampled without replacement to at most 2,000 points (dataset size permitting). $\widehat{\mathcal{E}}$ is biased upwards at finite sample size, so values are read against the within-sample level, the mean of $\widehat{\mathcal{E}}$ between the two halves of a sample over twenty random splits.

Generation error. Here $\{x_i\}_{i=1}^n$ are the held-out test points in the top half of the exceedance set and $\tilde{\theta}_i \in \mathbb{S}_r^{\ell-1}$ is drawn from an angular law. The generation error is $\widehat{\mathcal{E}}\big(\{\hat{\theta}_i\}_{i=1}^n, \{\theta(x_i)\}_{i=1}^n\big)$ for

$$\hat{\theta}_i := \theta\big(h(r(z_i)\,\tilde{\theta}_i)\big),$$

averaged over twenty draws of $\{\tilde{\theta}_i\}_{i=1}^n$, and is reported in Tables 2, 4 and 5. The angular law is either the von Mises–Fisher mixture of Appendix H (*fitted law*) or the lower-half training latent directions resampled with replacement (*empirical law*). The *ambient generation error*, the 'ambient' column of those tables, replaces $\hat{\theta}_i$ by the ambient directions of the lower-half training exceedances, and the *angular reconstruction error* replaces $\hat{\theta}_i$ by $\theta(h(f(x_i)))$.

Distance correlation and independence test. Radius and direction are paired as $r(z_i)$ and $\theta(z_i)$ over $n = 2{,}000$ exceedances drawn without replacement. With $a_{ij} := |r(z_i) - r(z_j)|$ and $b_{ij} := \|\theta(z_i) - \theta(z_j)\|$, the empirical distance covariance and correlation (Székely et al., 2007) are

$$A_{ij} = a_{ij} - \frac{1}{n}\sum_k a_{ik} - \frac{1}{n}\sum_k a_{kj} + \frac{1}{n^2}\sum_{k,l} a_{kl},$$
$$\mathcal{V}^2 = \frac{1}{n^2}\sum_{i,j} A_{ij}B_{ij}, \qquad \mathcal{R}^2 = \frac{\sum_{i,j} A_{ij}B_{ij}}{\left(\sum_{i,j} A_{ij}^2\right)^{1/2}\left(\sum_{i,j} B_{ij}^2\right)^{1/2}},$$

with B_{ij} built from b_{ij} the same way and $\mathcal{R} \in [0, 1]$. The ambient value is estimated from a subsample of the same size as the latent ones and carries comparable sampling variability, with a standard deviation near 0.01 on the precipitation exceedances. Writing $\mathcal{V}_\pi^2$ for the statistic after permuting the rows and columns of A by π, over 200 uniform permutations, we reject at the 5% level on the p-value (Phipson & Smyth, 2010)

$$p = \frac{1 + \#\{\pi : \mathcal{V}_\pi^2 \geq \mathcal{V}^2\}}{201}.$$

G CURVED SURFACE

The surface of §6.1 is generated from independent w_1 and w_2, where w_1 has a Student-t law with 2.8 degrees of freedom and unit scale and $w_2 \sim \mathcal{N}(0, 0.9^2)$. The pair $q := (0.67 \tanh 2w_1,\ 0.70 \tanh w_2)$ is mapped to $\mathbb{S}^2$ by the inverse stereographic projection and lifted onto the unit sphere through the origin, giving

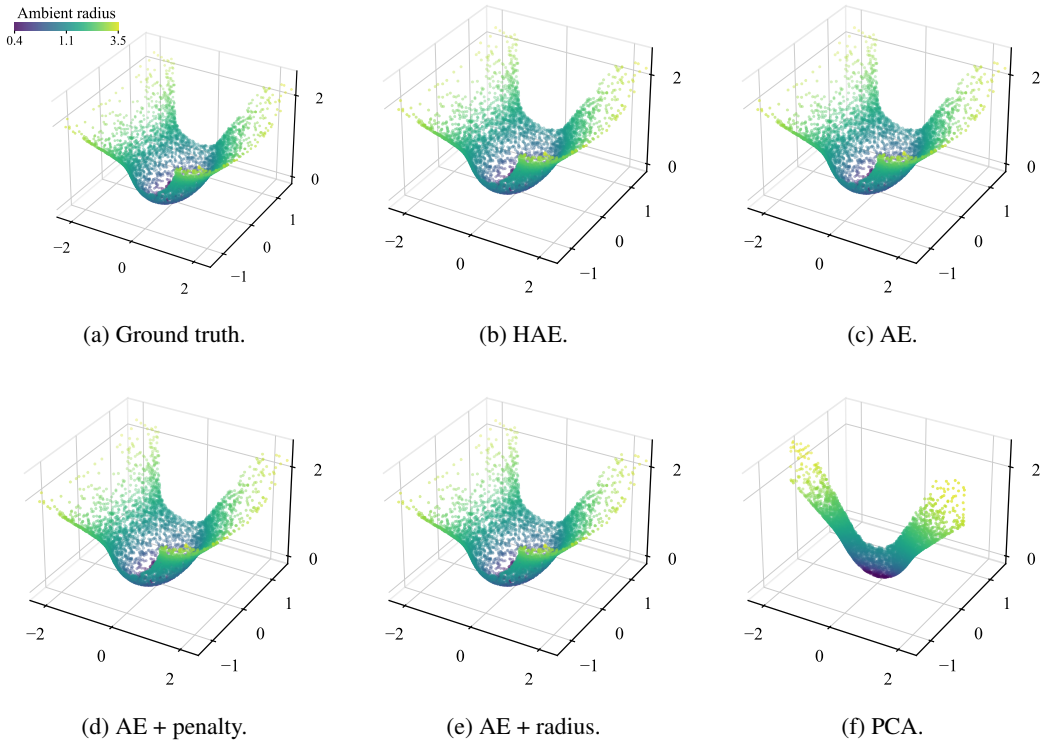
$$y := \frac{1}{\|q\|^2 + 1}\left(2q_1,\ 2q_2,\ \|q\|^2 - 1\right) + (0, 0, 1) \in \mathbb{R}^3,$$

and the surface point is

$$x = \left(\|y\| + 0.7\left(\sqrt{1 + w_1^2} - 1\right)\right)\frac{y}{\|y\|} \in \mathbb{R}^3.$$

the tail index (Definition E.1) of $\|x\|$ is $\alpha = 2.8$, inherited from w_1. The numerical parameter values used for the surface construction were chosen to provide visualisable yet interesting manifold geometry.

Sampling. The surface experiments draw $25{,}000$ training, $5{,}000$ validation and $5{,}000$ test points per seed. The exceedance set is drawn once from $70{,}000$ points split in the proportions 0.70, 0.15 and 0.15, giving $7{,}321$, $1{,}594$ and $1{,}585$ exceedances for the exceedance-generation experiment (§6.1). Reconstructions are shown in Figure G.1



(a) Ground truth. (b) HAE. (c) AE.

(d) AE + penalty. (e) AE + radius. (f) PCA.

Figure G.1: Curved surface test split and its reconstruction by each model, coloured by ambient radius.

H GENERATING CURVED SURFACE EXCEEDANCES

For regularly varying X (Definition E.1) with index α, the excess and the direction become independent in the limit,

$$\mathbb{P}\big(r(X) > \lambda u,\ \theta(X) \in B \mid r(X) > u\big) \longrightarrow \lambda^{-\alpha} S(B) \qquad (u \to \infty),$$

for $\lambda \geq 1$, a limiting angular law S on $\mathbb{S}_r^{D-1}$ and suitable Borel B (Resnick, 2007; De Haan & Ferreira, 2006; Lindskog et al., 2014).

Models are fitted at latent dimension two to the excess vectors $(r(x) - u)\,\theta(x)$ of the exceedance-generation experiment (§6.1) (Coles & Tawn, 1991). The angular law is a von Mises–Fisher mixture of eight components (Banerjee et al., 2005) fitted to the training latent directions below the median excess by 100 expectation-maximisation (Dempster et al., 1977) iterations. The energy distance between 4,000 draws of that mixture and the directions it was fitted to, divided by their within-sample level, has median below 0.9 on the surface.

Generation is evaluated on held-out observations above the median excess across twenty-five independent-seed fits with the mixture refitted each time. The radius-permuted exceedances keep every direction while permuting the radii within each split. Only the angular law differs between the fitted and empirical laws, so the difference in their generation errors is the cost of the fitted law rather than of the representation (Table 2). On the curved surface, Figure H.1 shows the latent directions against the excess, Figure H.2 the angular densities either side of the median excess and Figure H.3 the generated exceedances.

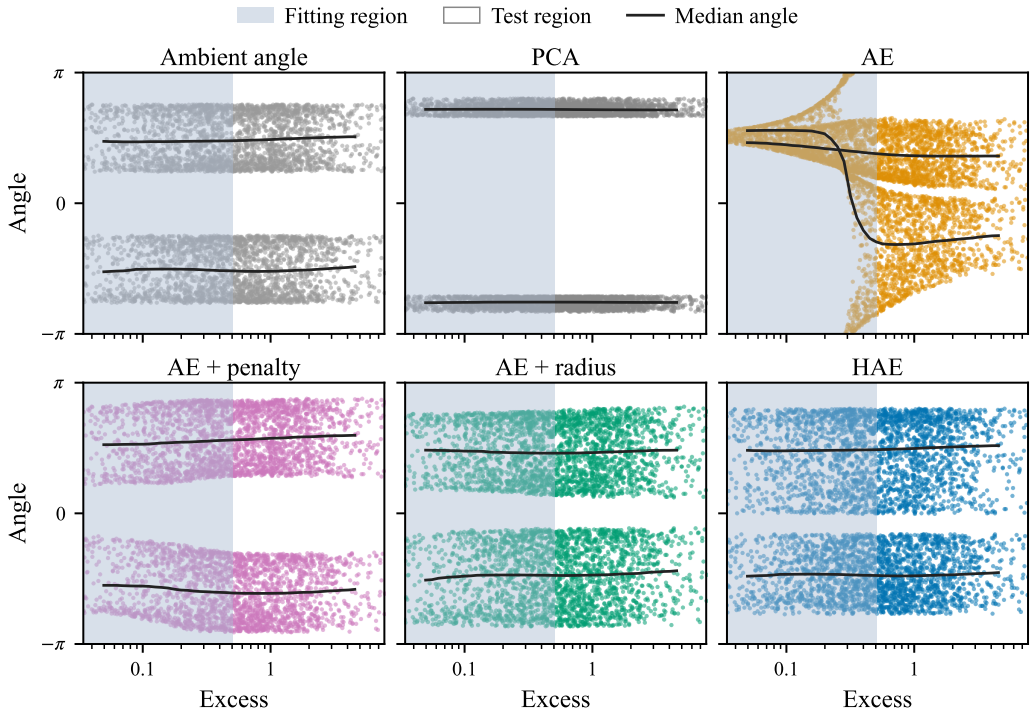


Figure H.1: Curved-surface exceedances, direction against the excess, each model's latent direction rotated and reflected onto the ambient. Solid curves are the median angle towards each tail, flat under independence.

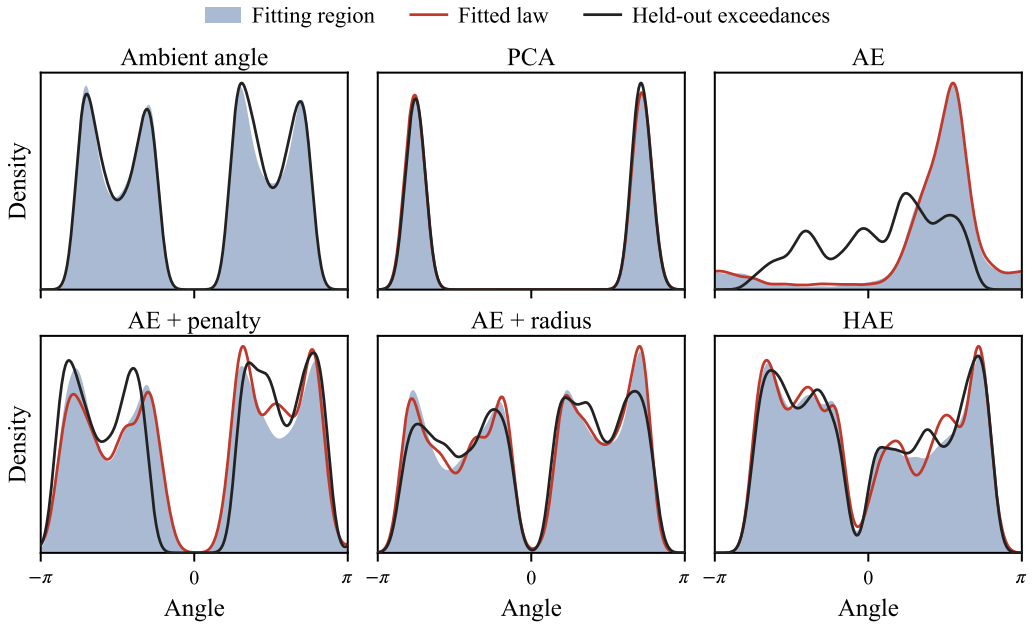


Figure H.2: Curved-surface angular densities on the frame of Figure H.1, for the fitting region (shaded), the fitted mixture (red) and the held-out exceedances above the median excess (black).

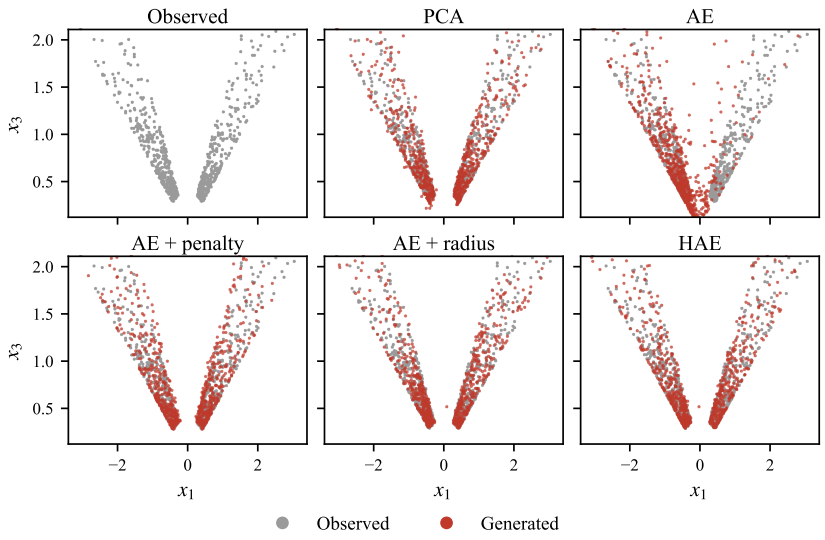


Figure H.3: Curved-surface exceedances above the median excess, generated from the law fitted below it (red) against the observed (grey).

I ERA5 FIELDS

Each cell of wind and temperature is standardised by its training-split mean and standard deviation, precipitation is divided by its standard deviation alone so a dry field keeps zero magnitude. The joint tensor stacks the variables after the same per-variable treatment. Table I.1 reports all four, with precipitation headlined in the body (Table 3). The radial diagnostics of Figures I.1, I.3 and I.5 show latent against ambient radius (log–log, left) and latent Hill-estimator curves (right), and Figures I.2, I.4 and I.6 the largest-magnitude test field with its reconstructions on the colour scale of the original.

Table I.1: ERA5 univariate fields (D=195) and the joint tensor (D=975), means over three fits. Entries marked † are exact by construction.

		PCA	AE	AE + penalty	AE + radius	HAE
u_{10} wind	Reconstruction RMSE	0.155	0.122	0.128	0.126	0.129
	Equivariance error, scale 10	$0^\dagger$	0.311	0.0500	0.175	$0^\dagger$
	Mean absolute radial error	0.276	9.68	0.0327	$0^\dagger$	$0^\dagger$
	Rank correlation of radii	0.999	0.879	1.000	$1^\dagger$	$1^\dagger$
	Hill estimate difference	0.0461	1.05	0.0188	$0^\dagger$	$0^\dagger$
t_{2m} temperature	Reconstruction RMSE	0.127	0.109	0.114	0.111	0.110
	Equivariance error, scale 10	$0^\dagger$	0.284	0.0306	0.108	$0^\dagger$
	Mean absolute radial error	0.173	10.6	0.0261	$0^\dagger$	$0^\dagger$
	Rank correlation of radii	1.000	0.700	1.000	$1^\dagger$	$1^\dagger$
	Hill estimate difference	0.00551	1.54	0.00783	$0^\dagger$	$0^\dagger$
tp precipitation	Reconstruction RMSE	0.164	0.135	0.133	0.137	0.138
	Equivariance error, scale 10	$0^\dagger$	0.510	0.135	0.208	$0^\dagger$
	Mean absolute radial error	0.179	4.56	0.0447	$0^\dagger$	$0^\dagger$
	Rank correlation of radii	1.000	0.992	1.000	$1^\dagger$	$1^\dagger$
	Hill estimate difference	0.0157	0.693	0.0103	$0^\dagger$	$0^\dagger$
Joint tensor	Reconstruction RMSE	0.157	0.124	0.130	0.126	0.127
	Equivariance error, scale 10	$0^\dagger$	0.182	0.0400	0.127	$0^\dagger$
	Mean absolute radial error	0.362	14.9	0.106	$0^\dagger$	$0^\dagger$
	Rank correlation of radii	0.999	0.976	1.000	$1^\dagger$	$1^\dagger$
	Hill estimate difference	0.00257	1.31	0.0138	$0^\dagger$	$0^\dagger$

I.1 u_{10} SURFACE WIND

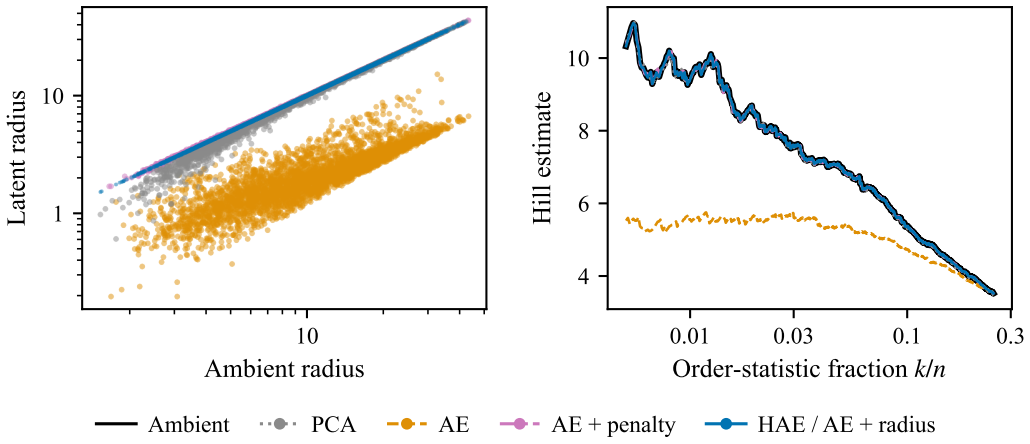


Figure I.1: ERA5 u_{10} (D=195, ℓ=5) radial diagnostics.

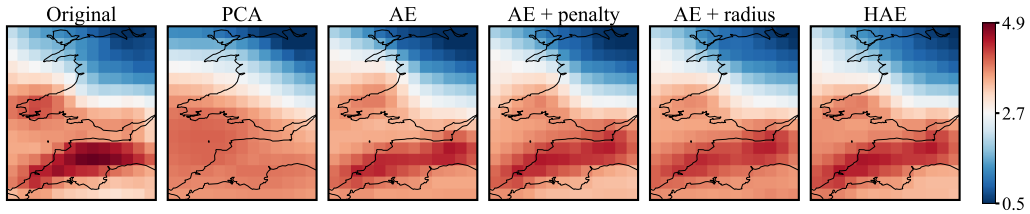


Figure I.2: ERA5 u_{10}, largest-magnitude test field and its reconstructions.

I.2 t_{2m} 2-METRE TEMPERATURE

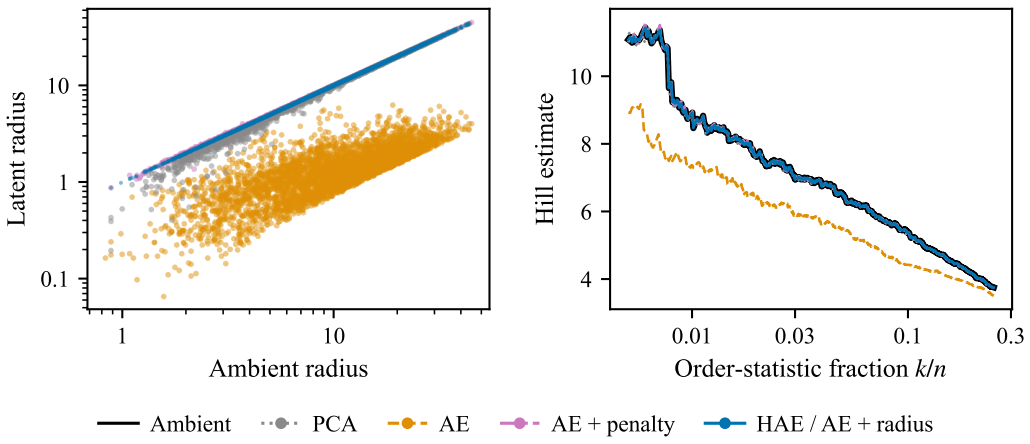


Figure I.3: ERA5 t_{2m} (D=195, ℓ=3) radial diagnostics. PCA's Hill curve lies on top of the ambient and HAE curves on this variable.

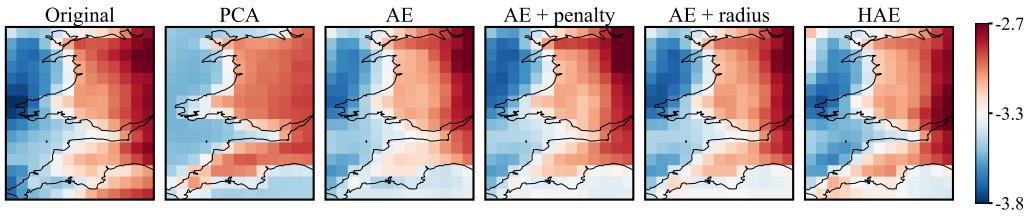


Figure I.4: ERA5 t_{2m}, largest-magnitude test field and its reconstructions.

I.3 tp TOTAL PRECIPITATION

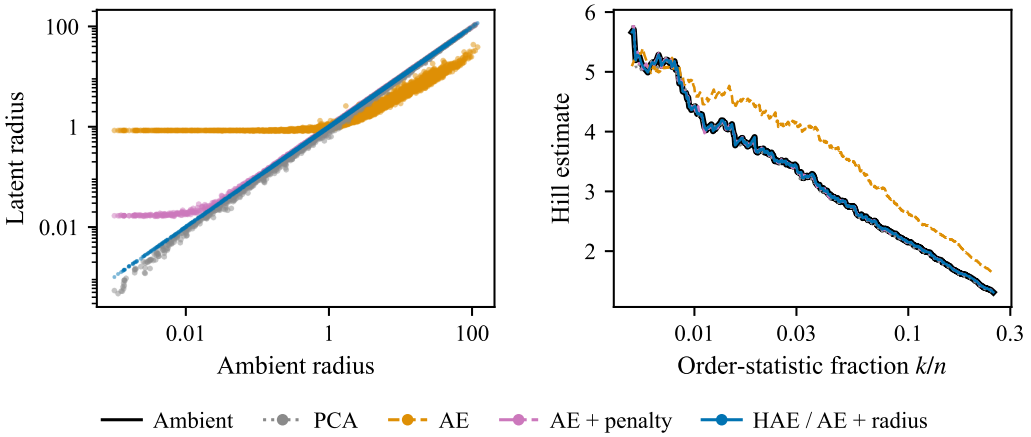


Figure I.5: ERA5 tp (D=195, ℓ=44) radial diagnostics.

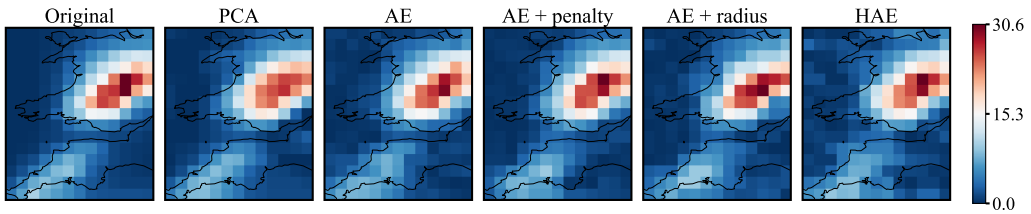


Figure I.6: ERA5 tp, largest-magnitude test field and its reconstructions.

I.4 JOINT TENSOR OF FIVE VARIABLES

Stacking five surface variables u_{10}, v_{10}, t_{2m}, tp and mean sea-level pressure into a single tensor gives D=975 ambient dimensions, compressed to ℓ=35. Figure I.7 shows the radial diagnostics. HAE preserves the joint radius but does not separately preserve each component variable's magnitude or tail index. A separate radial coordinate or degree per component would preserve each.

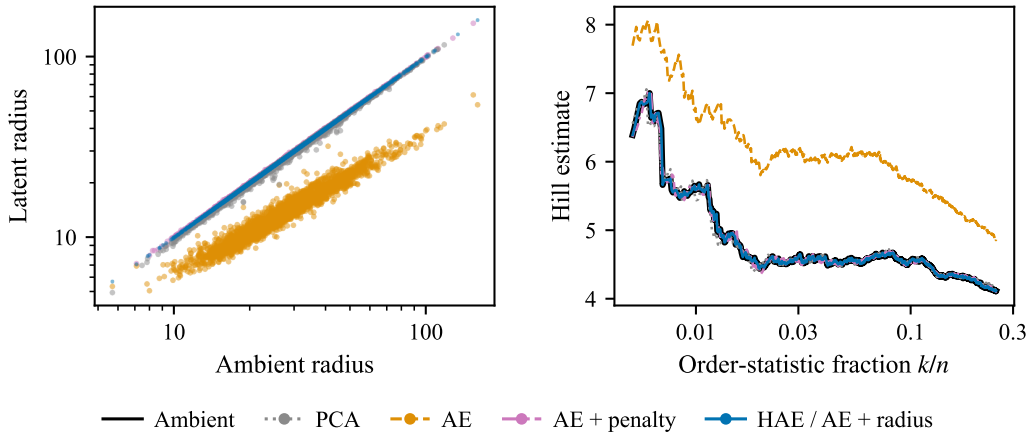


Figure I.7: ERA5 joint tensor (D=975, ℓ=35) radial diagnostics.

J PRECIPITATION EXCEEDANCES

A precipitation field of §6.2 is an exceedance when its magnitude exceeds the empirical 0.70 quantile u and is represented by its excess vector $(r(x) - u)\,\theta(x)$. The split gives 13,210 training, 2,838 validation and 2,777 test exceedances at latent dimension 36 with two hidden layers (Appendix C.2).

Generation. The independence assumed by the generation workflow of Appendix H fails at every threshold quantile, not only at the 0.70 threshold used here (Table J.1). Under the fitted law each neural model's generation error changes by less than a factor of two between the exceedances and the radius-permuted exceedances, whereas PCA's falls by nearly an order of magnitude and the ambient generation error, which exists only under the empirical law, by more than one (Tables 4 and 5). The dependence therefore accounts for most of PCA's generation error and of the ambient generation error, but little of the neural models'. Replacing the fitted law by the empirical law cuts their generation error on the radius-permuted exceedances by roughly an order of magnitude for HAE and AE + penalty. On the exceedances the generation error under the fitted law is more than an order of magnitude above the within-sample level while the angular reconstruction error stays below three times it, so the angular law rather than the round trip through the representation is what limits generation. The fitted law's draws already sit at two to three times the within-sample level of the latent directions they were fitted to.

Threshold sensitivity. Raising the threshold quantile from 0.70 to 0.97 halves the ambient distance correlation (Table J.1). HAE and PCA stay within 0.02 of the ambient value at every threshold quantile and AE + penalty and AE + radius within 0.05. The AE sits at 2.7 and 1.7 times the ambient value at the 0.70 and 0.76 quantiles, coming within 0.04 of it only from the 0.82 quantile upwards. A threshold chosen in the AE latent space would therefore be higher than one chosen in the ambient space.

Table J.1: Precipitation, distance correlation between radius and direction above each threshold quantile, each space using its own radius, means over twenty-five fits.

Threshold quantile	Exceedances	Ambient	PCA	AE	AE + penalty	AE + radius	HAE
0.70	18,825	0.223	0.229	0.596	0.259	0.269	0.237
0.76	15,060	0.220	0.227	0.383	0.227	0.202	0.226
0.82	11,295	0.209	0.192	0.238	0.215	0.184	0.208
0.85	9,412	0.188	0.181	0.201	0.202	0.177	0.198
0.88	7,530	0.171	0.171	0.179	0.192	0.172	0.186
0.91	5,648	0.160	0.165	0.163	0.179	0.162	0.179
0.94	3,765	0.160	0.150	0.139	0.161	0.147	0.157
0.97	1,883	0.113	0.111	0.144	0.115	0.114	0.117