Supplementary Information for *TerraNova*: A Foundation Model for the Anthropocene

Carlos Rodriguez-Pardo carlos.rodriguezpardo.jimenez@gmail.com
Department of Management, Economics and Industrial Engineering, Politecnico di Milano
RFF-CMCC European Institute on Economics and the Environment (EIEE)
Euro-Mediterranean Center on Climate Change (CMCC)

Massimo Tavoni massimo.tavoni@polimi.it
Department of Management, Economics and Industrial Engineering, Politecnico di Milano
RFF-CMCC European Institute on Economics and the Environment (EIEE)
Euro-Mediterranean Center on Climate Change (CMCC)

Abstract

This document is the Supplementary Information for *TerraNova: A Foundation Model for the Anthropocene*. It contains the full methodological specification of the model and training corpus (App. A), the complete forward and backward ablation programme (App. B), extended results (App. C) and the computational-cost analysis (App. D). The main paper and further resources are available at the project page: https://carlosrodriguezpardo.es/projects/TerraNova/. Section, figure and table numbers cited as "main text" refer to the main paper.

Appendix A. Full Methodological Specification

Part A. Formal setup

This Part fixes the notation used throughout this appendix (§A.1), states the multi-geometry learning problem that motivates the entire architecture (§A.2), and defines the task taxonomy over which the model is trained (§A.3). Readers familiar with the main text may skim §A.2 but should note the embedding-dimension table (Table 1) and the radius-normalisation convention (Eq. N), both of which recur in every subsequent section.

A.1 Notation

Vectors are columns; $\odot$ is the Hadamard (element-wise) product; $\sigma(\cdot)$ is the logistic sigmoid; $\hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$ is the ℓ_2-normalised $\mathbf{v}$; $\mathrm{sg}[\cdot]$ is the stop-gradient operator (identity on the forward pass, zero Jacobian on the backward pass); $\langle\cdot\rangle$ denotes an empirical mean over a batch; and Φ is the standard-normal CDF. Linear maps are written $W_\bullet$ with biases absorbed unless stated. LN is LayerNorm (Ba et al., 2016), GELU the Gaussian-error linear unit (Hendrycks and Gimpel, 2016), and $\mathrm{SiLU}(x) = x\,\sigma(x)$.

A convention used by every encoder is to rescale its output to a fixed ℓ_2 radius,

$$\mathrm{N}_d(\mathbf{v}) \;\coloneqq\; \sqrt{d}\,\frac{\mathbf{v}}{\|\mathbf{v}\|_2}, \tag{N}$$

so that all embeddings live on a sphere of radius $\sqrt{d}$. Fixing the radius (rather than leaving the norm free) keeps cosine similarities, residual gates, and cross-attention logits at a stable scale across encoders of very different internal magnitude, and prevents any one modality from dominating a fused representation simply by having a larger norm. The dimensions of the production model are collected in Table 1.

Table 1: Production embedding dimensions. The run scales the selected components up relative to those selected in the ablation study (ablation appendix).

symbol	object	dim	source field
$\mathbf{e}_\ell, \mathbf{s}$	location / spatial bus	256	`location.embedding_dim`
$\mathbf{e}_c$	country	64	`arch.country_embedding_dim`
$\boldsymbol{\tau}$	time	32	`arch.time_embedding_dim`
$\mathbf{z}_{\mathrm{st}}$	spatiotemporal	256	`spatiotemporal_fusion.output_dim`
$\mathbf{e}_\tau$	task	256	`arch.task_embedding_dim`
$\mathbf{c}$	fused task·ST conditioning	512	`task_st_fusion_output_dim`
$\mathbf{h}$	hypernetwork hidden	512	`hypernet_hidden_dim`
H	generated decoder hidden	32	`one_layer_hidden_dim`

A.2 Problem formulation

The defining difficulty of modelling the Anthropocene jointly is not a shortage of data but a mismatch of *geometry*. Environmental processes are observed as continuous fields over the surface of the Earth: temperature, precipitation, vegetation, emissions and soil moisture vary smoothly over space with strong, multi-scale geographic dependence, and are naturally indexed by coordinates. Human systems, by contrast, are observed through *boundaries*: GDP, governance, health, inequality and institutional capacity are reported for countries and other administrative units whose borders are historical and political, not physical. Italy is not a coordinate, and a grid cell in the Alps is not an economy. Yet climate impacts occur locally while many responses are organised nationally, so any model of climate–society interaction must connect continuous geographic variation with discrete administrative structure.

TerraNova treats this as a multi-geometry representation-learning problem rather than collapsing one geometry into the other. Formally, a *query* is a triple $q = (\tau, u, t)$ in which τ indexes a prediction *task*, t is time (a calendar year, optionally refined to month/day), and u is a geometry-dependent spatial argument:

- a coordinate $u = x = (\lambda, \phi) \in [-180, 180) \times [-90, 90]$ (longitude, latitude in degrees) for a *gridded task*;
- an administrative unit $u = c \in \{1, \ldots, n_c\}$ (an ISO3 country code) for a *country task*.

The model is a single map

$$f_\theta : \; (\tau, u, t) \;\longmapsto\; (\mu, \nu, \alpha, \beta) \in \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>1} \times \mathbb{R}_{>0}, \tag{1}$$

returning the four hyperparameters of a Normal–Inverse–Gamma (NIG) predictive distribution over the standardised target y (Part C). Two consequences of the z-scoring (§A.14) are worth stating now, because they are exploited repeatedly downstream: the marginal target variance is ≈ 1, and a predictive standard deviation beyond a few units is therefore *a priori* implausible, a fact the evidential head and loss use to make the aleatoric/epistemic split well-posed (§A.7.2, §A.8).

Design rationale. The two standard ways to force these geometries together are both imperfect. Averaging fields to country level (the route taken by most economic models) discards within-country exposure patterns: precisely the local structure that determines who is affected. Gridding socioeconomic variables may assume a spatial precision that was never observed and treats a national aggregate as if it were a measurement at every cell. Either way the coupling between local environment and national institutions is at some level lost before learning begins. TerraNova instead keeps each native geometry and learns a shared latent space in which a location inside a country and that country are *representationally* adjacent(§A.9.3); knowledge then transfers across geometries through the representation rather than through a lossy coordinate transform. This is a central premise of the model and the reason every component below is built to operate on a common 256-dimensional latent vector regardless of input geometry.

A.3 Task taxonomy

The model is trained on $T = 1{,}024$ *reconstruction* tasks: 512 gridded fields (§A.11) and 512 country variables (§A.12): each a regression target predicted from its native geometry and supervised by the evidential loss of §A.8. Beyond reconstruction, three families of *auxiliary* objectives shape the shared representation without themselves being prediction targets:

(i) *temporal transport* (§A.9.1), which regularises the temporal geometry so that moving a query in time is a smooth, composable operation;

(ii) *external geospatial alignment* (§A.9.2), which couples the spatiotemporal representation to pretrained satellite/ground-image embedding spaces; and

(iii) *internal country–location alignment* (§A.9.3), which is the mechanism that makes the two native geometries mutually informative.

A single backbone serves every task and objective. The only input-dependent routing is the spatial-input geometry (Eq. 7) and the task embedding (§A.4.6); the encoders, fusion, decoder and head are shared. Each optimiser step draws exactly one task/objective from a fixed mixed schedule (§A.17), so the total objective (Eq. 35) is optimised in expectation. Table 2 summarises the five step families and their roles.

Part B. Model architecture

The architecture is a composition of specialised encoders (§A.4), two cross-modal fusion blocks (§A.6), and a task-conditioned decoder with an evidential head (§A.7), with embedding dropout, applied at every representation boundary (§A.5) shared across all three. The guiding principle is that each input modality is processed by an encoder matched to its geometry and then

Table 2: The five training step families. "Geometry" is the spatial input used; "supervision" is the loss; "#/epoch" is the per-epoch step count (§A.17).

family	geometry	supervision	#/epoch
reconstruction	coordinate / country	evidential NIG (27)	32768
transport	coordinate / country	NIG + latent consistency (32)	512
geo_alignment	coordinate	sampled-negative InfoNCE (33)	256
country_alignment	country ↔ coords	sampled-negative InfoNCE (34)	256
country_consistency	country ↔ coords	MSE (disabled, w=0)	0

projected onto a common 256*-dimensional bus (this number is significantly lower than any domain's dimensionality, thus making Terranova's representations a compression of the Earth System) on which a uniform set of fusion and decoding operations acts regardless of where the query came from. The forward pass is the three-stage pipeline* `encode` → `fuse` → `decode` *of Algorithm 1, with the dataflow of Figure 1. Throughout this Part we give the exact operations, the tensor shapes, the initialisation that matters for training stability, and the study that selected each block.*

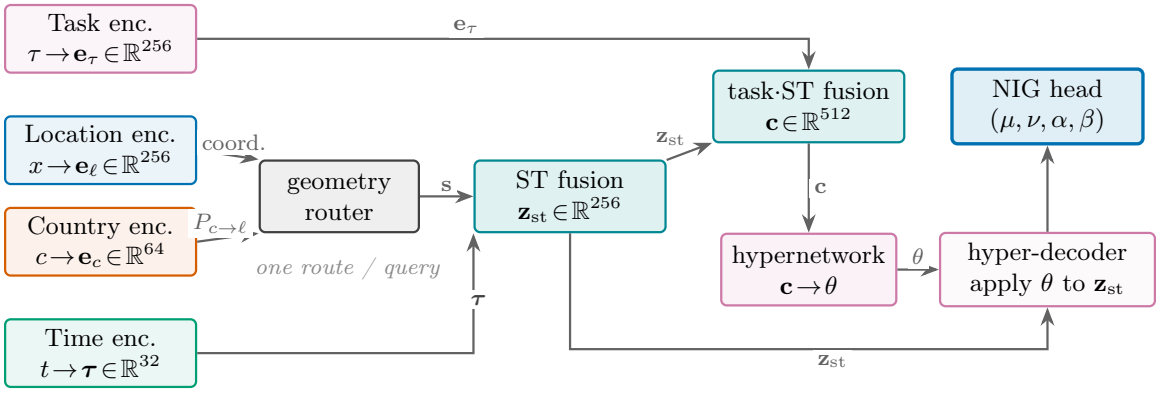

Figure 1: Model dataflow. Modality-specific encoders (left column, task on top) map each input to the shared bus. Because every task is purely gridded or purely country-level, the *geometry router* selects exactly one spatial route per query (the location embedding $\mathbf{e}_\ell$ for a coordinate, or the country embedding projected into the location space, $\mathrm{N}(P_{c\to\ell}\mathbf{e}_c)$, for a country) producing the shared 256-D bus $\mathbf{s}$ (the additive hybrid route exists in code but is unused). ST fusion folds time $\boldsymbol{\tau}$ into $\mathbf{s}$ to give the spatiotemporal embedding $\mathbf{z}_{\text{st}}$; task·ST fusion conditions it on the task to give $\mathbf{c}$; the hypernetwork turns $\mathbf{c}$ into the weights θ of a small decoder, and that generated decoder is applied to $\mathbf{z}_{\text{st}}$ to emit the four NIG parameters.

A.4 Encoders

TerraNova reads four heterogeneous inputs: a year, a coordinate, a country identifier, and a task identifier, and maps each through a dedicated encoder to a unit-norm embedding on the sphere of radius $\sqrt{d}$ (Eq. N). Figure 2 shows the four encoders as per-module dataflows; the remainder of this section specifies each in turn.

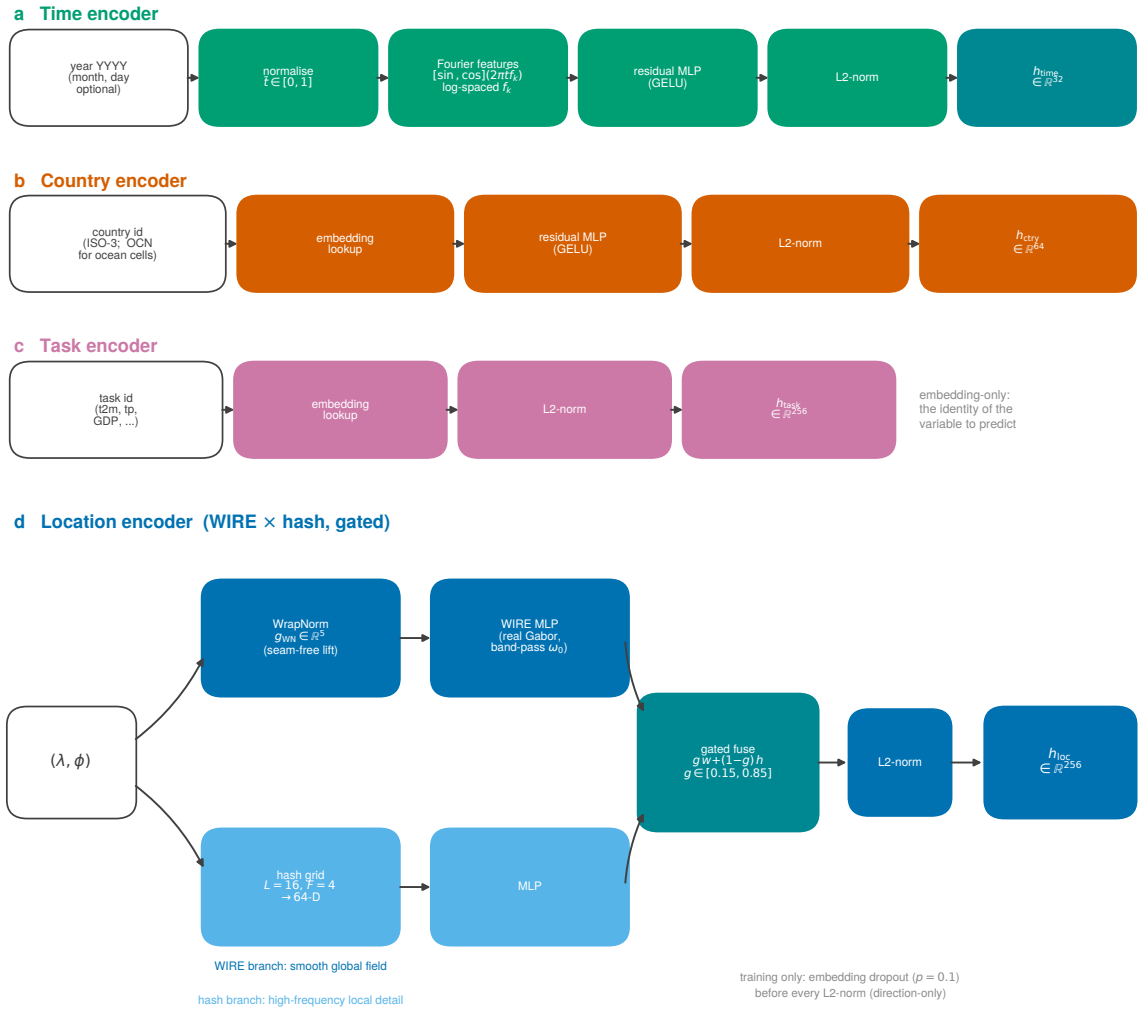

Figure 2: The four input encoders as per-module dataflows. The **time** encoder normalises the year, applies a log-spaced Fourier feature map (Eq. 8) and a residual MLP. The **country** and **task** encoders are embedding lookups (the country encoder adds a residual MLP; the synthetic ocean country `OCN` of §A.15 is registered dynamically by the samplers). The **location** encoder runs two complementary branches in parallel: a seam-free WrapNorm lift feeding a WIRE U-Net (smooth, low-frequency structure) and a multi-resolution hash grid (sharp, high-frequency detail); and blends them per latent dimension with a learned gate (Eq. 5). Every encoder ends in an L^2 normalisation onto the radius-$\sqrt{d}$ sphere; during training, embedding dropout (p_e=0.15) is applied to each output *before* that normalisation, regularising the embedding's direction rather than its scale (§A.5).

Algorithm 1 TerraNova forward pass $f_\theta(\tau, u, t)$ (single query; batched in practice)

Require: task id τ, spatial arg u (coordinate x or country c), time t

// **Encode (independent, parallel)**
$\boldsymbol{\tau} \leftarrow \mathrm{N}_{32}(\mathrm{TimeEnc}(t)) \in \mathbb{R}^{32}$; $\mathbf{e}_\tau \leftarrow \mathrm{N}_{256}(E_\tau[\tau]) \in \mathbb{R}^{256}$ ▷ Eqs. 8,11
if u is a coordinate x **then**
  $\mathbf{e}_\ell \leftarrow \mathrm{LocationEnc}(x) \in \mathbb{R}^{256}$; $\mathbf{s} \leftarrow \mathbf{e}_\ell$ ▷ Eq. 5
else if u is a country c **then**
  $\mathbf{e}_c \leftarrow \mathrm{CountryEnc}(c) \in \mathbb{R}^{64}$; $\mathbf{s} \leftarrow \mathrm{N}_{256}(P_{c\to\ell}\mathbf{e}_c) \in \mathbb{R}^{256}$ ▷ Eqs. 6,7
// **Fuse**
$\mathbf{s}' \leftarrow W_s\mathbf{s}$; $\mathbf{z}_{\mathrm{st}} \leftarrow \mathrm{N}_{256}\big(\mathbf{s}' + g\,\boldsymbol{\delta}(\mathbf{s}', \boldsymbol{\tau})\big) \in \mathbb{R}^{256}$ ▷ spatial queries time, Eq. 15
$\mathbf{c} \leftarrow \mathrm{TaskSTFusion}(\mathbf{z}_{\mathrm{st}}, \mathbf{e}_\tau) \in \mathbb{R}^{512}$ ▷ ST queries task, Eq. 16
// **Decode**
$\mathbf{h} \leftarrow \mathrm{GELU}(W_h\mathbf{c} + b_h) \in \mathbb{R}^{512}$ ▷ hypernetwork core
$(W_1, b_1, W_2, b_2) \leftarrow$ row-normalised generator heads on $\mathbf{h}$ ▷ Eq. 17
$\mathbf{a} \leftarrow \big(\mathrm{GELU}(\mathbf{z}_{\mathrm{st}}^\top W_1 + b_1)\big)W_2 + b_2 \in \mathbb{R}^4$ ▷ Eq. 18
return $(\mu, \nu, \alpha, \beta) \leftarrow \mathrm{NIGactivate}(\mathbf{a})$ ▷ Eqs. 19,20

A.4.1 Location encoder: gated WIRE U-Net × hash grid

The location encoder is the hybrid `HybridWireHash`, by a wide margin the largest of the four encoders at 50.2 million parameters and the one the ablation study treated as most consequential. Its input is the raw coordinate $x = (\lambda, \phi)$ in degrees and its output is the location embedding $\mathbf{e}_\ell \in \mathbb{R}^{256}$. The design problem it solves is that a single coordinate network cannot be simultaneously good at the two regimes Earth-system fields present: *smooth* fields (temperature, pressure) whose spatial spectrum is concentrated at low frequencies, and *sparse* fields (population density, built surface, burned area) whose spectrum has heavy high-frequency content and sharp discontinuities at coastlines and city edges. A network biased toward smoothness blurs the latter; a network with enough local capacity for the latter injects noise into the former. `HybridWireHash` resolves this by running two complementary encodings of the *same* coordinate in parallel: a continuous implicit branch with a controllable, smoothness-favouring spectral prior, and a lookup branch with a local, high-capacity prior; nd blending them *per latent dimension* with a learned gate, so that each of the 256 output dimensions is supplied by whichever branch carries the structure that dimension needs. The two branches, the gate, and the coordinate preprocessing they share are described in turn below; Table 3 summarises the sub-blocks and their roles.

Coordinate preprocessing (WrapNorm). Raw degrees are first lifted to a discontinuity-free 5-dimensional spherical feature,

$$g_{\mathrm{WN}}(\lambda, \phi) = \Big(\cos\phi\cos\lambda,\ \cos\phi\sin\lambda,\ \sin\phi,\ \cos\phi\cos(\lambda{+}\tfrac{\pi}{2}),\ \cos\phi\sin(\lambda{+}\tfrac{\pi}{2})\Big) \in \mathbb{R}^5. \quad (2)$$

The first three components $(\cos\phi\cos\lambda, \cos\phi\sin\lambda, \sin\phi)$ are exactly the Cartesian unit vector of the point on the sphere, so they encode position continuously and isometrically (Euclidean distance in this 3-D embedding is a monotone function of great-circle distance) and are automatically periodic in longitude and bounded in latitude. The last two repeat the

Table 3: Sub-blocks of the location encoder (`HybridWireHash`). The continuous and lookup branches each output $\mathbb{R}^{256}$ and are blended by the per-dimension gate.

sub-block	operation	role
WrapNorm lift	$(\lambda, \phi) \to \mathbb{R}^5$ (Eq. 2)	seam-free spherical features
WIRE U-Net	$\mathbb{R}^5 \to \mathbb{R}^{256}$, Gabor activations (Eq. 3)	smooth / low-frequency structure
hash grid	$(\lambda, \phi) \to \mathbb{R}^{64}$ multi-res. lookup (Eq. 4)	sharp / high-frequency detail
hash decoder	$\mathbb{R}^{64} \to \mathbb{R}^{256}$, GELU/LN residual MLP	decode dense hash features
gate + combine	per-dim blend (Eq. 5)	route each dim to a branch

longitude encoding with a $\frac{\pi}{2}$ phase shift; because $\cos(\lambda+\frac{\pi}{2}) = -\sin\lambda$ and $\sin(\lambda+\frac{\pi}{2}) = \cos\lambda$, they are, up to sign, linear functions of the first pair (Fig. 3), a redundant, symmetric parameterisation of longitude rather than new information, which we retain so that the WIRE branch's first nonlinearity reads a balanced pair of phases. The periodic cos / sin lift is moreover seam-free at the $\pm 180°$ antimeridian, where a naive longitude/180 scaling would give geographic neighbours nearly opposite codes. The lift is fixed (non-learnable, with no parameters) and is shared by the WIRE branch only; the hash branch (§A.4.1, "lookup branch") consumes the raw degrees directly. Feeding a smooth, seam-free, bounded function of position into the WIRE branch is what lets that branch's smoothness prior act on *geographic* structure rather than on artefacts of the coordinate parameterisation (Fig. 3).

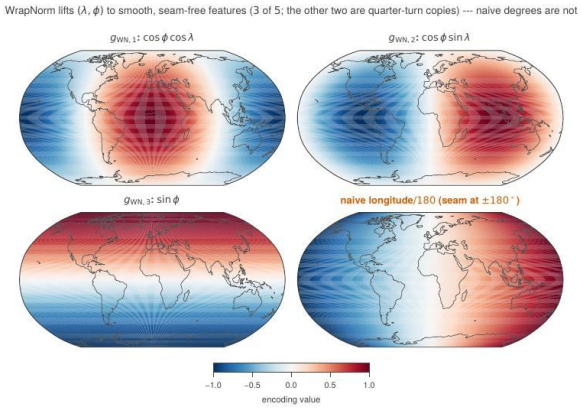

Figure 3: The WrapNorm features (Eq. 2). The first three components $(g_{\mathrm{WN},1}, g_{\mathrm{WN},2}, g_{\mathrm{WN},3})$ are the Cartesian unit vector of the surface point and are shown here; the remaining two repeat the longitude encoding with a $\pi/2$ phase shift and, because $\cos(\lambda+\frac{\pi}{2}) = -\sin\lambda$ and $\sin(\lambda+\frac{\pi}{2}) = \cos\lambda$, equal $-g_{\mathrm{WN},2}$ and $g_{\mathrm{WN},1}$ exactly, identical maps up to sign, so they are omitted to avoid redundancy (they guarantee a non-degenerate longitude derivative without adding new information). All are smooth and seam-free across the $\pm 180°$ antimeridian, whereas a naive longitude/180 scaling (bottom right) is discontinuous there: geographic neighbours astride the antimeridian would otherwise receive nearly opposite codes.

Continuous branch (WIRE U-Net). The smooth branch is a U-shaped multilayer perceptron with hidden widths $[512, 1024, 2048, 4096, 2048, 1024, 512]$. The widths expand then contract around the central 4096-unit layer, giving the branch the multiscale structure

of a U-Net: the early (narrow-to-wide) layers lift the coordinate feature into a progressively richer description of position, and the later (wide-to-narrow) layers compress it back to the output width. Symmetric skip connections join encoder- and decoder-side layers of equal width as pre-activation residual additions, so gradients reach the early layers directly and the expanding side can reuse the contracting side's features at matched resolution (Ronneberger et al., 2015; He et al., 2016b); this multiscale structure is what lets a single coordinate network represent both planetary-scale gradients and regional structure. Every unit uses the *real* Gabor wavelet activation (Saragadam et al., 2023)

$$\sigma_{\mathrm{WIRE}}(z) = \exp\Big(-\tfrac{1}{2}s_0^2 z^2\Big)\cos(\omega_0 z), \qquad \omega_0 = 10,\ s_0 = 10,\ \omega_0^{\mathrm{first}} = 30, \tag{3}$$

mapping $g_{\mathrm{WN}}(x) \mapsto \mathbf{h}_{\mathrm{wire}} \in \mathbb{R}^{256}$. The two constants have distinct roles: ω_0 is the carrier frequency that sets how fast each unit oscillates, and s_0 is the inverse envelope width that sets how tightly the Gaussian $\exp(-\frac{1}{2}s_0^2 z^2)$ localises that oscillation. A Gabor wavelet is jointly compact in space and frequency (it saturates the uncertainty bound between the two), so each unit is a tunable band-pass detector: this is precisely the representation needed for fields that are smooth almost everywhere but carry localised structure, and it is the reason WIRE encodes high-frequency content with better conditioning than a pure-sine SIREN, whose units are global, so a single frequency scale ω_0 governs the whole field (Sitzmann et al., 2020; Saragadam et al., 2023). The first layer uses a higher carrier ($\omega_0^{\mathrm{first}} = 30$) so the network injects high-frequency content immediately, while the recurrent layers use $\omega_0 = 10$ to refine it. Weights follow the SIREN initialisation philosophy (Sitzmann et al., 2020), adapted to the Gabor carrier: the first layer is drawn uniformly in $[-1/d_{\mathrm{in}}, 1/d_{\mathrm{in}}]$ to keep pre-activations in the active (non-saturated) Gabor regime, and hidden layers in $[-\sqrt{6/d_{\mathrm{in}}}/\omega_0, \sqrt{6/d_{\mathrm{in}}}/\omega_0]$ so that the pre-activation variance is preserved with depth and the network neither collapses to a constant nor explodes. This branch alone is a competitive coordinate encoder; hybrid uses it for the dimensions where smoothness is the right prior.

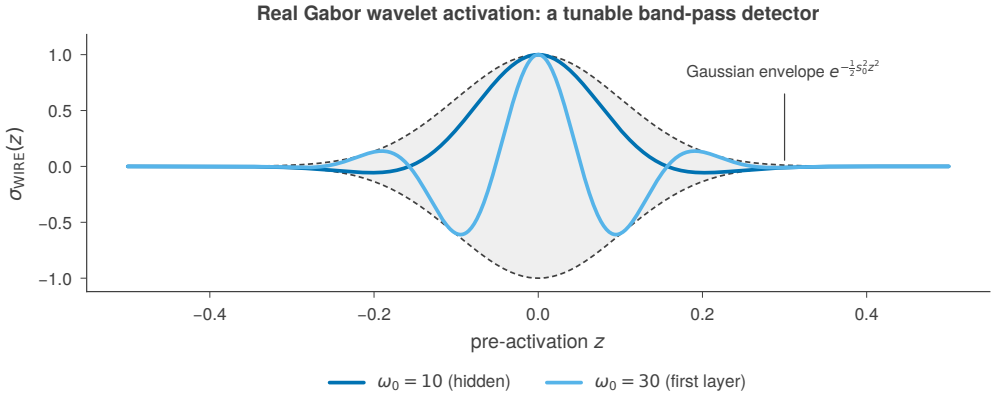

Figure 4: The real Gabor wavelet activation (Eq. 3), $\sigma_{\mathrm{WIRE}}(z) = e^{-\frac{1}{2}s_0^2 z^2}\cos(\omega_0 z)$, shown for the recurrent carrier ω_0=10 and the higher first-layer carrier ω_0=30. The Gaussian envelope localises the oscillation, making each unit a tunable band-pass detector that is jointly compact in space and frequency.

Lookup branch (multi-resolution hash grid). The high-frequency branch follows the multi-resolution hash encoding of Müller et al. (2022), with our settings $L = 16$ resolution levels, each with a hash table of 2^{16} rows storing $F = 4$ trainable features, with geometric per-level resolutions

$$N_\ell = \lfloor N_{\min}\, b^\ell \rfloor, \quad b = \left(N_{\max}/N_{\min}\right)^{1/(L-1)}, \quad N_{\min} = 16,\ N_{\max} = 1024,\ \ell = 0, \ldots, L{-}1. \tag{4}$$

At each level ℓ the coordinate is mapped to a regular grid of resolution N_ℓ; the integer corners of the enclosing cell are mapped to table rows by a spatial hash, and the level's contribution is the bilinear interpolation of the four corner feature vectors by the coordinate's fractional position within the cell. Hashing (rather than a dense grid) is what makes the high-resolution levels affordable: a dense grid at $N_{\max} = 1024$ would need $\sim 10^6$ cells per level, whereas each hash table is fixed at $2^{16} = 65{,}536$ rows. At coarse levels ($N_\ell^2 \le 2^{16}$) the map is collision-free and the level behaves like a true dense grid; at fine levels distinct cells inevitably collide into the same row, but the collisions are not resolved at lookup: they are left for training to disambiguate, because the gradients of colliding entries average, and the samples that matter most to the reconstruction dominate that average, and the other (non-colliding) levels provide the context that separates them. The $L = 16$ levels are geometrically spaced (Eq. 4) so that, taken together, the concatenated $LF = 64$-D feature is an explicitly *multi-scale* description of the neighbourhood of the query, from $N_{\min} = 16$ (continental) to $N_{\max} = 1024$ (roughly 0.18°, sub-regional). Because each level's lookup is local and the levels alias independently, the branch can represent sharp, locally discontinuous structure with no smoothness penalty: the regime in which sparse socioeconomic rasters (population, built surface, impervious area) live, and the regime the WIRE branch handles poorly (Fig. 5).

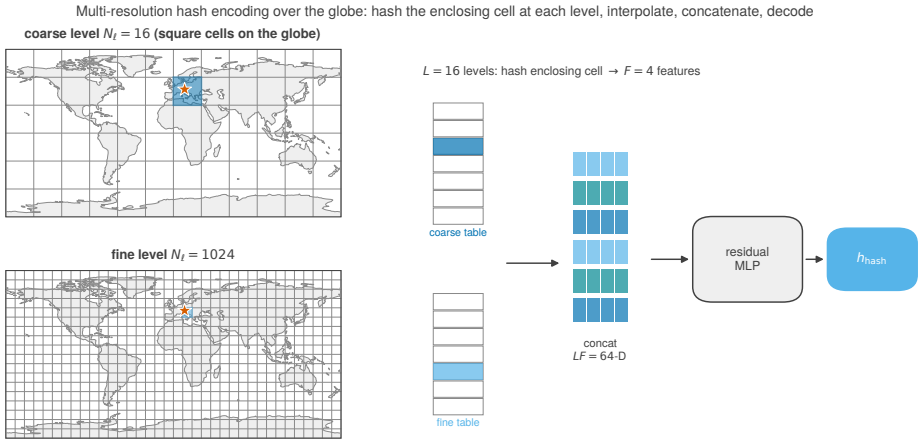

Figure 5: Multi-resolution hash encoding (Eq. 4). The query coordinate (star) falls inside one cell at each of L=16 geometrically spaced resolutions; two are shown, a coarse level (N_ℓ=16) and a fine level (N_ℓ=1024), as square graticule cells on the globe. The enclosing cell is mapped to a hash-table row, the four corners are bilinearly interpolated, and the per-level F=4 features are concatenated into an LF=64-D vector and decoded by a residual MLP into $\mathbf{h}_{\text{hash}}$.

Hash decoder. The 64-D multi-resolution feature is a dense learned vector, so it is decoded by a conventional network rather than a coordinate network: a four-block pre-

norm residual MLP of width 256 with GELU activations and LayerNorm (He et al., 2016a; Ba et al., 2016; Hendrycks and Gimpel, 2016), with an orthogonally initialised output projection, producing $\mathbf{h}_{\text{hash}} \in \mathbb{R}^{256}$. Using a residual MLP here (and not a second WIRE) is deliberate: SIREN/WIRE activations are tuned to map raw low-dimensional coordinates to high-frequency outputs, whereas the hash features are already high-frequency and merely need a smooth nonlinear read-out, for which GELU/LayerNorm residual blocks are better-conditioned.

Per-dimension gate. The two branches are blended by a learnable per-dimension gate $\boldsymbol{\alpha}_g \in \mathbb{R}^{256}$ (initialised to 1.386, so $\sigma(1.386) \approx 0.8$, biasing the blend toward the smooth branch at the start of training):

$$\mathbf{g} = \text{clamp}\big(\sigma(\boldsymbol{\alpha}_g),\, 0.15,\, 0.85\big), \qquad \mathbf{e}_\ell = \text{N}_{256}\Big(W_{\text{comb}}\big(\mathbf{g} \odot \mathbf{h}_{\text{wire}} + (1 - \mathbf{g}) \odot \mathbf{h}_{\text{hash}}\big)\Big), \quad (5)$$

with W_{comb} identity-initialised so the module starts as a plain convex blend and learns a final linear remixing on top. The gate is *per-dimension* rather than a single scalar because the two priors are not uniformly better or worse: a smooth field wants nearly all of its 256 dimensions from the WIRE branch, a sparse raster wants many from the hash branch, and most fields want a mixture: a scalar gate could not express this, whereas 256 independent gates let the encoder dedicate different latent dimensions to different spectral regimes and serve all tasks from one shared embedding. The sigmoid constrains each gate to $(0, 1)$ so the blend is always a convex combination (no dimension can be amplified beyond the two branches' range), and the bias initialisation $\boldsymbol{\alpha}_g = 1.386$ ($\sigma \approx 0.8$) starts the encoder at 80% WIRE, reflecting the prior that most Earth-system fields are smoother than they are sparse and giving the high-capacity hash branch a gentle entry. The clamp to $[0.15, 0.85]$ keeps both branches gradient-alive throughout training (a fully closed gate would strand a branch's parameters with zero gradient and prevent recovery). After training the gate is also a *diagnostic*: the per-dimension mean $\langle \mathbf{g} \rangle$ measures how much of the final embedding the continuous branch contributes, and its spread measures how specialised the dimensions are. The continuous branch ends up carrying the smooth, globally organised zonal gradients while the hash branch carries the sparse, high-frequency local detail, and the gated combination inherits both.

Design rationale. The ablation study isolates the location encoder on a 51-field WorldTensor reconstruction benchmark and frames the central tension precisely: the spatial pathway must reconstruct fields whose spatial irregularity spans several orders of magnitude, from zonally smooth sea-level pressure to sparse high-frequency rasters such as population density and wildfire area. The two families of coordinate encoders in the literature sit at opposite ends of this spectrum. Continuous implicit representations like SIREN (Sitzmann et al., 2020), WIRE (Saragadam et al., 2023), Fourier-feature networks (Tancik et al., 2020), or spherical harmonics (Rußwurm et al., 2024) fix their bandwidth a priori through the frequency parameterisation rather than adapting it locally, so one global setting either blurs sharp detail or injects noise into smooth fields. Lookup encodings (Müller et al., 2022) impose a local, high-capacity prior that excels on discontinuities but leaks noise on smooth targets. The study's key finding is that this trade-off is *architecturally mixable*: a symmetric, identity-initialised gate over LayerNormed sub-embeddings learns to route each latent dimension to the branch whose inductive bias fits the signal it carries, and the resulting hybrid strictly outperforms both named parents on the majority of fields while sitting on the parameter–accuracy Pareto frontier. We adopt the concrete WIRE×Hash-L pairing and biases the initial gate toward the continuous branch, following the assumption that most fields are smoother than they are sparse.

A.4.2 Country encoder

A country code carries no geometric structure to exploit, so the country encoder is a learned table followed by a small nonlinear refinement. The table $E_c \in \mathbb{R}^{n_c \times 64}$ is initialised $\sim \mathcal{N}(0, 0.02^2)$ with extra reserved rows so that new countries can be registered without reallocating the parameter (which would otherwise break optimiser state); a residual MLP R_ψ of four pre-norm residual blocks (width 256, SiLU, LayerNorm) then refines the lookup:

$$\mathbf{e}_c = \mathrm{N}_{64}\big(R_\psi(E_c[c])\big) \in \mathbb{R}^{64}. \tag{6}$$

Design rationale. The ablation study reframes the country encoder away from the expressivity question that governs the location encoder and toward representational *sufficiency*: how much embedding capacity is needed to encode heterogeneous national attributes, and whether a learned nonlinearity helps beyond a bare table. We use the table-plus-residual-MLP form; the 64-D width is the operating point inside the saturating regime, chosen to enable compressed latents, and to balance the country pathway against the much higher-capacity spatial bus rather than to maximise country reconstruction in isolation.

A.4.3 Geometry routing: the shared spatial embedding

Every query, of either geometry, enters the temporal and decoding stack through a single 256-dimensional *spatial embedding* $\mathbf{s}$. tasks are purely one geometry or the other: a task is either gridded (a coordinate x) or country-level (a country id c), never both, so $\mathbf{s}$ is formed by *selecting* a route, not by combining the two encoders. The country vector is lifted to the location dimension by $P_{c\to\ell} : \mathbb{R}^{64} \to \mathbb{R}^{256}$, implemented in the model as a two-layer projector

$$\mathbb{R}^{64} \xrightarrow{\text{Linear}} \mathbb{R}^{256} \xrightarrow{\text{LN+SiLU+dropout}} \mathbb{R}^{256} \xrightarrow{\text{Linear+LN}} \mathbb{R}^{256},$$

and the query geometry selects which branch defines the bus:

$$\mathbf{s} = \begin{cases} \mathbf{e}_\ell, & \text{coordinate task,} \\ \mathrm{N}_{256}\big(P_{c\to\ell}(\mathbf{e}_c)\big), & \text{country task.} \end{cases} \tag{7}$$

Routing rather than mixing is what lets a single temporal-fusion and decoding stack serve both native geometries: whichever branch is active, everything downstream sees one 256-D vector on a common bus and behaves identically for gridded and country queries. The country branch is the one projected (the location embedding already defines the bus at its native width), so the country signal is mapped *into* the spatial space rather than reshaping it. The code additionally exposes an additive *hybrid* route $\mathrm{N}_{256}(\mathbf{e}_\ell + P_{c\to\ell}\mathbf{e}_c)$ for queries that carry both a coordinate and a containing country, but no reconstruction task is hybrid, so this path is never exercised in the runs reported here.

A.4.4 Time encoder

Time enters as a Fourier-feature lift followed by a residual MLP. The year is normalised to $[0,1]$ via $\bar{t} = (t - t_{\min})/(t_{\max} - t_{\min})$ with $t_{\min} = 1900$, $t_{\max} = 2035$, and lifted with $K = 32$ log-spaced frequencies $f_k \in [1, K/2]$ and a learnable scalar scale s (Tancik et al., 2020; Mildenhall et al., 2022):

$$\gamma(t) = \big[\sin(2\pi\,\bar{t}\,s\,f_k),\ \cos(2\pi\,\bar{t}\,s\,f_k)\big]_{k=1}^{K} \in \mathbb{R}^{64}, \qquad \boldsymbol{\tau} = \mathrm{N}_{32}\big(\mathrm{MLP}_{\mathrm{res}}(\gamma(t))\big) \in \mathbb{R}^{32}. \tag{8}$$

$\mathrm{MLP}_{\mathrm{res}}$ is a width-256 linear lift, four residual blocks, and a linear head. The log-spaced frequencies give the lift sensitivity to both decadal trends (low f_k) and year-to-year structure (high f_k); s lets the model learn the overall temporal bandwidth. The encoder also supports optional zero-mean cyclic month/day features, which are disabled in the annual configuration.

Design rationale. The ablation study analyses the temporal encoder under a deliberately tight 8-dimensional bottleneck across 15 variables spanning smooth long-term trends and episodic series (e.g. burned area). Eight families are compared, from a raw-scalar projection to RBF bases and gated Fourier/RBF hybrids. The Fourier basis followed by a small residual MLP is the winner on both fit quality and temporal-dynamics fidelity: the explicit oscillatory basis organises long-term structure cleanly, while the residual MLP supplies the nonlinear refinement that basis selection alone cannot: a raw scalar is too weak, and localised RBF encoders behave like high-capacity lookups that do not generalise across heterogeneous temporal regimes. Notably, hybridising the bases did *not* help here, in contrast to the location encoder, indicating that the temporal signal is well captured by a single smooth basis plus a small correction.

A.4.5 Temporal transport operator

A learned operator $T(\boldsymbol{\tau}, \Delta)$ advances a time embedding by a signed offset Δ (in years); it is used by the transport objective (§A.9.1) and is the only module that explicitly models temporal *evolution* in latent space. The offset is normalised $\bar{\Delta} = \mathrm{clamp}(\Delta/\Delta_{\max}, -1, 1)$, $\Delta_{\max} = 4$, and encoded by a gated combination of a signed-Fourier basis γ_Δ and a signed-RBF basis ρ_Δ (centres on $[-1,1]$):

$$\mathbf{h}_\Delta = \sigma(\boldsymbol{\alpha}_\Delta) \odot W_F\,\gamma_\Delta(\bar{\Delta}) + \big(1 - \sigma(\boldsymbol{\alpha}_\Delta)\big) \odot W_R\,\rho_\Delta(\bar{\Delta}), \qquad \rho_{\Delta,k}(\bar{\Delta}) = e^{-\frac{1}{2}((\bar{\Delta} - c_k)/\varsigma)^2}, \tag{9}$$

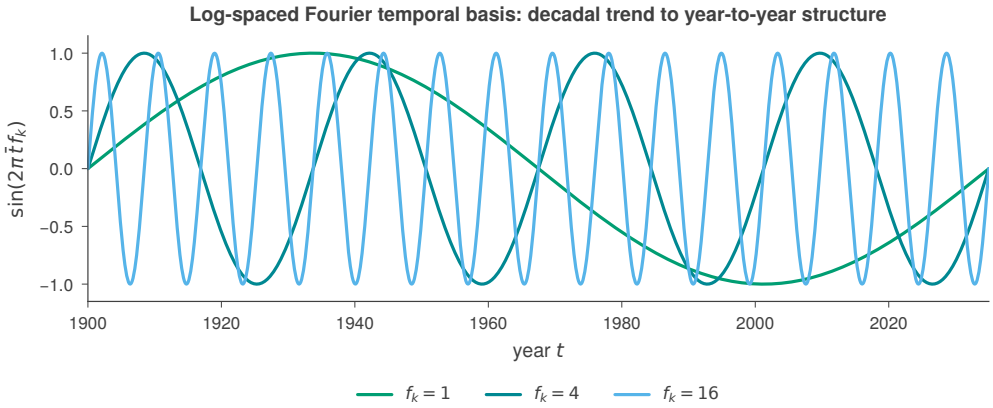

Figure 6: Three components of the log-spaced Fourier temporal basis (Eq. 8) over the calendar range $[1900, 2035]$. Low frequencies carry decadal trend; high frequencies carry year-to-year structure. The learnable scale s lets the model choose the overall temporal bandwidth, and the residual MLP supplies the nonlinear refinement the basis alone cannot.

with ς learnable (init 0.15). Transport then applies a gated residual update to the reference embedding (the gate $\mathbf{g}_T = \sigma(\mathbf{r})$ is initialised at $\mathbf{r} = -2$, $\sigma(\mathbf{r}) \approx 0.12$, so T starts as a small, near-identity perturbation), *without* output renormalisation:

$$T(\boldsymbol{\tau}, \Delta) = \boldsymbol{\tau} + \mathbf{g}_T \odot \mathrm{Trunk}\Big(\mathrm{act}\big(\mathrm{LN}(W_{\mathrm{ref}}\boldsymbol{\tau} + \mathbf{h}_\Delta)\big)\Big). \tag{10}$$

Output renormalisation is deliberately omitted here so that the identity ($\Delta = 0$) and semigroup constraints (§A.9.1) are exactly, not approximately, satisfiable.

Design rationale. In the ablation study, we ask whether an explicit temporal operator is still useful once strong time and location encoders exist, or whether direct prediction already suffices. Five formulations are compared, from "direct-only" (no transport) to "transport + full consistency" (predictive supervision, latent matching, an identity constraint at zero shift, and semigroup consistency). The clean but non-trivial result is that the best overall held-out model and the best explicit transported rollout are *different* configurations: matching-plus-identity yields the sharpest rollout, but full consistency yields the strongest overall model and the smoothest, most predictable time-embedding geometry, because semigroup consistency regularises the operator globally even where it does not improve the rollout path. The key negative result is equally important: predictive supervision alone does not beat the no-transport baseline, so transport helps only when constrained by latent-consistency objectives.

A.4.6 Task encoder

Task identity is a pure lookup, $E_\tau \in \mathbb{R}^{N_\tau \times 256}$ (entries $\sim \mathcal{N}(0, 0.02^2)$, with reserved rows for dynamic task registration):

$$\mathbf{e}_\tau = \mathrm{N}_{256}\big(E_\tau[\tau]\big) \in \mathbb{R}^{256}. \tag{11}$$

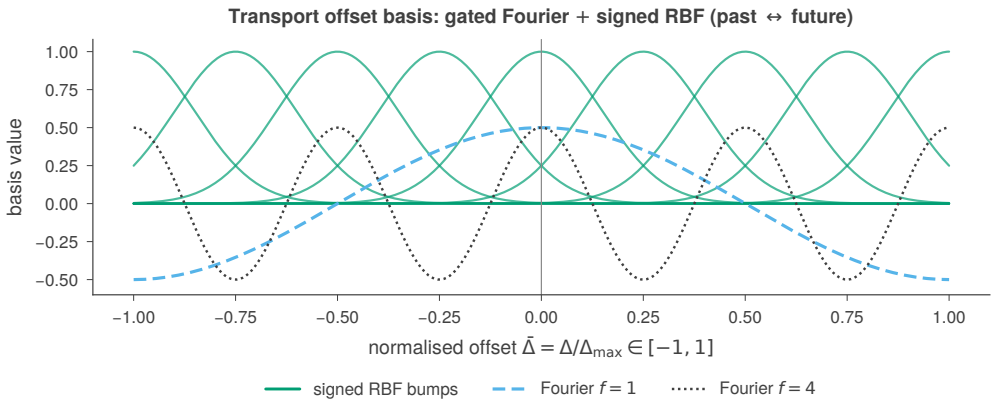

Figure 7: The transport offset basis (Eq. 9): a learnable gated combination of a signed-Fourier basis (smooth, global in the offset) and a signed-RBF basis (localised bumps at centres on $[-1, 1]$). Encoding the *signed* normalised offset $\bar{\Delta} = \Delta/\Delta_{\max}$ symmetrically lets one operator advance and rewind the time embedding, which is what makes the identity and semigroup constraints (§A.9.1) expressible.

Design rationale. In the ablation study, we compare an embedding-only task encoder against an embedding-plus-residual-MLP variant and finds the extra nonlinearity provides no benefit once task identity conditions the decoder through the hypernetwork; the embedding-only form is selected as the simpler model with equal performance. The expressive task→prediction mapping is deliberately concentrated in the task·ST fusion and the hypernetwork (§A.6.2, §A.7), where it can interact with spatial and temporal context, rather than being spent transforming a context-free task identifier. Note that we choose the task embedding size to be lower to the total number of training tasks, thereby encouraging compression and information transfer.

A.5 Embedding dropout and regularisation

Our model uses two distinct dropout mechanisms. Ordinary internal dropout (Srivastava et al., 2014) ($p_{\text{gen}} = 0.15$, from `arch.general_dropout`) appears inside MLPs, attention blocks, and projector layers. Separately, the `data.embedding_dropout`=0.15 knob applies dropout at representation boundaries, after an encoder or fusion block has produced an embedding but before the final fixed-radius normalisation:

$$\mathcal{D}_p(\mathbf{v}) = \frac{\mathbf{m} \odot \mathbf{v}}{1-p}, \qquad m_j \sim \text{Bernoulli}(1-p), \quad p = 0.1, \qquad \tilde{\mathbf{v}} = \mathrm{N}_d\big(\mathcal{D}_p(\mathbf{v})\big). \tag{12}$$

This is applied during training to the location, time, task, and country encoder outputs and to the spatiotemporal fusion output $\mathbf{z}_{\text{st}}$ (§A.6.1, specified next). For these radius-normalised modules, dropout changes the *direction* of the embedding while the subsequent N_d restores the radius $\sqrt{d}$, so the regulariser discourages reliance on single latent coordinates without making attention logits or cosine losses fluctuate simply because the embedding norm changed. The same knob is wired to the original task-time fusion when that block is enabled; in the evidential model the original task-time block is disabled and replaced by Task·ST fusion

(§A.6.2). Temporal transport also receives the knob, but it is configured without output renormalisation (§A.4.5), so transported-time dropout masks the transported embedding directly during training and is absent at evaluation.

Design rationale. The fixed-radius convention (Eq. N) makes encoder outputs comparable across modalities, but it would be undermined by post-normal dropout, which changes the norm seen by every downstream dot product. Applying dropout before radius restoration keeps the training-time scale convention intact while forcing the model to distribute spatial, temporal, country, and task information across multiple latent dimensions. This matters especially because the decoder is hypernetwork-conditioned: without boundary-level dropout, a task could over-specialise a small subset of embedding coordinates that the generated readout then amplifies.

A.6 Cross-modal fusion

The encoder outputs are combined and decoded by the model's prediction head, shown end-to-end in Fig. 8: two cross-modal fusions produce the conditioning vector, a one-layer hypernetwork (§A.7.1) turns it into decoder weights, and the resulting decoder maps the spatiotemporal embedding to the evidential output (§A.7). This section specifies the two fusion blocks; the decoder and head follow in §A.7.

Both fusion blocks are cross-modal Transformers (Vaswani et al., 2017) in which one modality, treated as a single-token query, attends to another, treated as key/value memory. Suppressing the batch axis, $\mathbf{q}, \mathbf{m} \in \mathbb{R}^{1\times w}$. We write multi-head attention as $\mathrm{Attn}(\mathbf{q}, \mathbf{m}, \mathbf{m}) = W^o[\mathrm{Attn}_1; \ldots; \mathrm{Attn}_H]$, where, for head h of dimension $d_h = w/H$,

$$\mathrm{Attn}_h(\mathbf{q}, \mathbf{m}) = \mathrm{softmax}\left(\frac{(\mathbf{q}W_h^q)(\mathbf{m}W_h^k)^\top}{\sqrt{d_h}}\right)(\mathbf{m}W_h^v). \tag{13}$$

The length-one memory makes the softmax in Eq. (13) identically one. Thus these blocks do *not* use attention to select among tokens: each head learns a value subspace in which to inject the memory, while the query survives through the residual path. The subsequent feed-forward network acts on their sum, so stacking blocks repeatedly alternates cross-modal value injection with nonlinear joint refinement. Attention direction still matters because it determines which modality is the persistent residual state and which is the repeatedly injected memory. The two stacks additionally differ in depth, normalisation placement, and residual gating, as detailed below.

A.6.1 Spatiotemporal fusion (Tr-A, post-norm, α-gated residual)

Here a projected *spatial* token queries the *time* token. Width $w = 1024$, 16 heads, depth $D_{\mathrm{st}} = 8$. The spatial embedding is first projected onto the bus, $\mathbf{s}' = W_s\mathbf{s} \in \mathbb{R}^{256}$; the cross-modal stack runs in the width-1024 space, $\mathbf{q}_0 = W_q\mathbf{s}'$, $\mathbf{kv} = W_{kv}\boldsymbol{\tau}$, with each of the D_{st} *post-norm* blocks

$$\mathbf{q} \leftarrow \mathrm{LN}\big(\mathbf{q} + \mathrm{Attn}(\mathbf{q}, \mathbf{kv}, \mathbf{kv})\big), \qquad \mathbf{q} \leftarrow \mathrm{LN}\big(\mathbf{q} + \mathrm{FFN}(\mathbf{q})\big). \tag{14}$$

Here $\mathbf{q}_0, \mathbf{kv} \in \mathbb{R}^{B\times 1\times 1024}$ and $\mathrm{FFN}(z) = W_2\,\mathrm{SiLU}(W_1 z)$ with $W_1 : \mathbb{R}^{1024} \to \mathbb{R}^{4096}$ and $W_2 : \mathbb{R}^{4096} \to \mathbb{R}^{1024}$. The implementation applies dropout on the attention and feed-forward

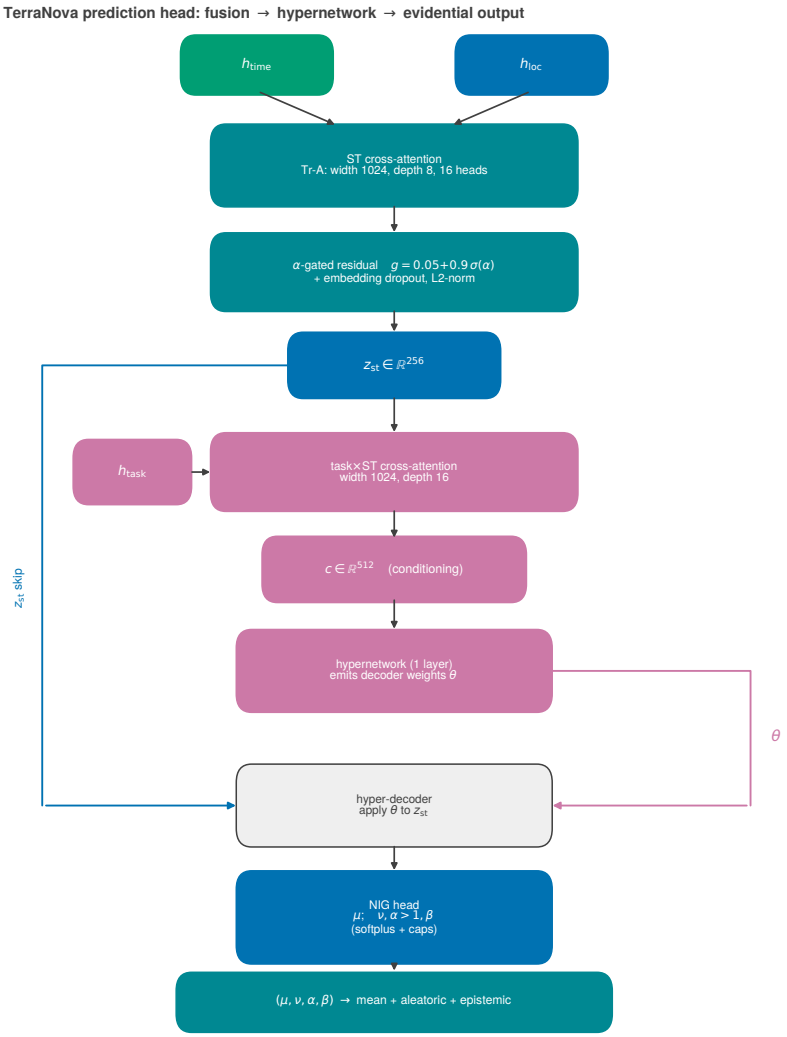

Figure 8: The TerraNova prediction dataflow. The spatiotemporal fusion (§A.6.1) cross-attends the time and location tokens and applies an α-gated residual to produce $\mathbf{z}_{\text{st}} \in \mathbb{R}^{256}$ (with embedding dropout before its normalisation, §A.5). The task·ST fusion (§A.6.2) conditions on the task token to produce $\mathbf{c} \in \mathbb{R}^{512}$, from which a one-layer hypernetwork emits the decoder weights θ. The decoder then applies θ to $\mathbf{z}_{\text{st}}$ (passed in by a skip, left rail) and the Normal–Inverse–Gamma head reads out $(\mu, \nu, \alpha, \beta)$, giving the predictive mean together with aleatoric and epistemic uncertainty (§A.7.3).

residual branches (omitted from Eq. 14 for readability). A zero-initialised head produces the temporal modulation $\boldsymbol{\delta} = W_o \mathbf{q}_{D_{\text{st}}} \in \mathbb{R}^{256}$, which is combined with the spatial embedding through a scalar α-gated residual (α init -1.1, giving gate $g \approx 0.275$, so the block starts as a small perturbation of the spatial path):

$$g = 0.05 + 0.9\,\sigma(\alpha), \qquad \mathbf{z}_{\text{st}} = \mathrm{N}_{256}\big(\mathbf{s}' + g\,\boldsymbol{\delta}\big). \tag{15}$$

Design rationale. In the ablation study, we compare 48 variants: (concatenation-MLP, three FiLM forms, gated FiLM, residual gates, four attention directions, joint self-attention, and two transformer families) are swept over hidden width and depth on a benchmark that combines gridded held-out years and 150 time-varying country variables. Three conclusions drive the choice. First and obviously, explicit spatiotemporal fusion is necessary: a spatial-only baseline reconstructs static structure but fails on held-out country years. Second, the *direction* of interaction matters: letting the spatial state query time and inject temporal information as conditioning memory is consistently useful, whereas the reverse direction (time querying space, then reconstructing) is nearly indistinguishable from the spatial-only baseline. Third, and most importantly, the best mechanism is a *repeated* cross-modal transformer rather than a single modulation: FiLM, gated FiLM, residual gates and concat-MLPs all behave like single-step corrections and saturate in suboptimal errors, while the transformer alternates cross-modal attention and nonlinear refinement and keeps improving with width and depth. The selected spatial-queries-time transformer (Tr-A) at $w512, d8$ was both the lowest-error model and the efficient high-capacity endpoint of the Pareto frontier, and it won the harder country modality without sacrificing gridded accuracy. Our model runs Tr-A at $w1024$ with 16 heads and wraps it in the zero-initialised delta and α-gated residual of (15) so the fused state is a controlled, identity-initialised perturbation of the spatial embedding.

A.6.2 Task·spatiotemporal fusion (Tr-A, pre-norm, learned gates)

A second cross-modal Transformer lets the *spatiotemporal* state query the *task* token, producing the conditioning vector $\mathbf{c}$ that drives the decoder. Width 1024, 16 heads ($d_h = 64$), depth $D_f = 16$, output $d_f = 512$. With $\mathbf{q}_0 = W'_s \mathbf{z}_{\text{st}}$, $\mathbf{m} = W'_t \mathbf{e}_\tau$ (both in $\mathbb{R}^{B\times 1\times 1024}$), each block d is *pre-norm* with its own learned scalar gates $g_{a,d} = 0.05 + 0.9\,\sigma(\alpha_{a,d})$ and $g_{f,d} = 0.05 + 0.9\,\sigma(\alpha_{f,d})$ (all $\alpha_\bullet$ init -1.1):

$$\begin{aligned} \hat{\mathbf{q}}_d &= \mathbf{q}_{d-1} + g_{a,d}\ \mathrm{Attn}\big(\mathrm{LN}(\mathbf{q}_{d-1}), \mathbf{m}, \mathbf{m}\big), \qquad && d = 1, \dots, D_f, \\ \mathbf{q}_d &= \hat{\mathbf{q}}_d + g_{f,d}\ \mathrm{FFN}_d\big(\mathrm{LN}(\hat{\mathbf{q}}_d)\big), \qquad && \mathbf{c} = W'_o\, \mathbf{q}_{D_f} \in \mathbb{R}^{512}. \end{aligned} \tag{16}$$

Each FFN_d again has a $1024 \to 4096 \to 1024$ SiLU structure; no additional dropout is applied inside this task-conditioning stack. The pre-norm placement and per-block learned gates (rather than the post-norm, fixed-residual form of §A.6.1) suit the much greater depth here ($D_f = 16$): pre-norm keeps the residual stream well-conditioned at depth, and the gates let the network grow its effective depth gradually from an identity-like initialisation.

Design rationale. Through our ablation study, we establish two facts about task conditioning that together specify this block. First, conditioning must happen *before* the decoder: task-only conditioning at the head level is insufficient, because the head then has no access to the spatiotemporal state. Second, the interaction must be *repeated and cross-modal*: a concat-MLP fusion of task and spatiotemporal embeddings is not competitive with cross-modal attention, and the task-level winners are distributed across the transformer family rather than concentrated in any simpler mechanism. This block is that interaction; crucially it queries the already-fused $\mathbf{z}_{\text{st}}$ (rather than the raw spatial embedding), so the conditioning vector $\mathbf{c}$ is aware of *when* as well as *where* the query is observed.

A.7 Decoder and evidential head

A.7.1 One-layer hypernetwork

Rather than feeding the conditioning vector and the spatiotemporal state into a shared MLP, TerraNova uses $\mathbf{c}$ to *generate* a small, per-query decoder which is then applied to $\mathbf{z}_{\text{st}}$. A hypernetwork core h_ϕ (Linear $512 \to 512$ followed by GELU; depth 1, i.e. no residual blocks) maps $\mathbf{c} \mapsto \mathbf{h} \in \mathbb{R}^{512}$, and four linear generator heads emit the parameters of a one-hidden-layer decoder of hidden width $H = 32$ and output 4:

$$W_1 = \text{rownorm}\big(g_{W_1}(\mathbf{h})\big) \in \mathbb{R}^{256\times 32},\; b_1 = g_{b_1}(\mathbf{h}),\; W_2 = \text{rownorm}\big(g_{W_2}(\mathbf{h})\big) \in \mathbb{R}^{32\times 4},\; b_2 = g_{b_2}(\mathbf{h}). \tag{17}$$

The row-normalisation $\text{rownorm}(W)_i = s_i \hat{W}_i$ (a learned per-row scale s_i times the unit-normalised i-th row) bounds the scale of the generated layers, which is important because generated weights have no weight decay or fixed initialisation of their own and would otherwise be free to explode. The generated decoder is then applied to the spatiotemporal state:

$$\mathbf{a} = \Big(\text{GELU}\big(\mathbf{z}_{\text{st}}^\top W_1 + b_1\big)\Big) W_2 + b_2 \;=\; (a_\mu, a_\nu, a_\alpha, a_\beta) \in \mathbb{R}^4. \tag{18}$$

A single shared output projection emits all four NIG parameters.

Design rationale. The hypernetwork is the mechanism by which a single shared backbone adapts its computation to over a thousand heterogeneous tasks. Because the generated weights act on the full spatiotemporal state $\mathbf{z}_{\text{st}}$ rather than replacing it, spatial and temporal information survive intact to the readout while each task contributes a genuinely different local decoding function, a stronger form of task conditioning than concatenation or FiLM, which only shift or scale a shared computation. Our ablation selects this single-head one-layer hypernetwork: it is statistically indistinguishable from the otherwise identical four-head variant on held-out RMSE (the four-head model is the marginal raw-RMSE winner but the difference is not significant over the matched seeds), and the single-head form removes an unnecessary output-head over-parameterisation while keeping the decisive architectural ingredients. The row-normalisation helps making the generated weights stable.

A.7.2 Normal–Inverse–Gamma head

The four raw scalars are mapped to NIG hyperparameters (Amini et al., 2020),

$$\mu = a_\mu, \quad \nu = \text{softplus}(a_\nu) + \tfrac{1}{2}, \quad \alpha = \text{softplus}(a_\alpha) + 1 + 10^{-4}, \quad \beta = \text{softplus}(a_\beta) + 10^{-4}. \tag{19}$$

The offsets guarantee the parameter constraints ($\nu > 0$, $\alpha > 1$, $\beta > 0$); the $\nu \geq \frac{1}{2}$ floor in particular prevents an epistemic blow-up as $\nu \to 0$. Because targets are z-scored (unit marginal variance), a predictive standard deviation above $\sigma_{\max} = 3$ is implausible, and two differentiable caps enforce this without hard-clamping gradients on the active branch:

$$\beta \leftarrow \min\big(\beta,\ \sigma_{\max}^2(\alpha - 1)\big), \qquad \nu \leftarrow \max\Big(\nu,\ \tfrac{\beta}{\sigma_{\max}^2(\alpha-1)}\Big), \qquad \sigma_{\max} = 3. \tag{20}$$

The first cap bounds the aleatoric variance $\beta/(\alpha - 1) \leq \sigma_{\max}^2$; the second then bounds the epistemic variance $\beta/(\nu(\alpha - 1)) \leq \sigma_{\max}^2$. The min / max keep the gradient on the active branch, so the model can still lower β or raise ν when it wants smaller uncertainty; the caps only block runaway widths.

A.7.3 Predictive distribution and uncertainty read-outs

Placing the conjugate NIG prior on the unknown mean and variance of a Gaussian observation,

$$y \sim \mathcal{N}(\gamma, \sigma^2), \qquad \gamma \sim \mathcal{N}\big(\mu, \sigma^2/\nu\big), \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \tag{21}$$

and marginalising γ and σ^2 yields a Student-t posterior predictive,

$$p(y \mid \mu, \nu, \alpha, \beta) = \mathrm{St}\Big(y;\ \underbrace{\mu}_{\text{loc.}},\ \underbrace{\tfrac{\beta(1+\nu)}{\nu\alpha}}_{\text{scale}^2},\ \underbrace{2\alpha}_{\text{d.o.f.}}\Big). \tag{22}$$

We report both the identifiable Student-t width (Meinert et al., 2023) and the classic decomposition (Amini et al., 2020),

$$w_{\mathrm{St}} = \sqrt{\tfrac{\beta(1+\nu)}{\alpha\nu}}, \qquad \underbrace{\sigma_{\text{alea}}^2 = \tfrac{\beta}{\alpha-1}}_{\text{aleatoric (data) noise}}, \qquad \underbrace{\sigma_{\text{epi}}^2 = \tfrac{\beta}{\nu(\alpha-1)}}_{\text{epistemic (model) uncertainty}}. \tag{23}$$

The Student-t width w_{St} is the identifiable predictive scale used by the loss regulariser and width-cap diagnostics. The active-learning sampler, however, ranks tasks by the mean classic total variance $\sigma_{\text{alea}}^2 + \sigma_{\text{epi}}^2$ (§A.19); the classic split is reported for interpretability.

A.7.4 Aleatoric and epistemic uncertainty

The two variances in Eq. (23) answer fundamentally different questions and must not be conflated (Fig. 9). The *aleatoric* variance $\sigma_{\text{alea}}^2 = \beta/(\alpha - 1)$ is related to the irreducible noise of the data-generating process at a query: instrument and retrieval error, sub-grid variability averaged into a 0.25° cell, and genuine interannual volatility of the target. It is a property of the *world*, not of the model, and it does not shrink with more data: observing the same place and year again does not make the value intrinsically sharper. The *epistemic* variance $\sigma_{\text{epi}}^2 = \beta/(\nu(\alpha - 1))$ is by contrast the model's own ignorance: it is large where the model has seen little relevant evidence (a data-sparse country, a held-out future year, an under-represented climate regime) and *shrinks* toward zero as the pseudo-count ν accumulates evidence. The NIG head exposes both from a single forward pass: β and α set the scale of the inverse-gamma over the observation variance (aleatoric), while ν scales how tightly the latent mean is pinned down (epistemic).

The two axes are orthogonal, and the four corners of the (aleatoric, epistemic) plane all occur in Earth-system data. A daily-scale precipitation climatology that is densely observed is high-aleatoric/low-epistemic (intrinsically noisy but well-known); a smooth field queried at a year beyond the training window is low-aleatoric/high-epistemic (well-behaved but unobserved); a sparsely reported institutional indicator in a small country is high on both. Figure 9 shows the canonical contrast: the aleatoric band is roughly constant across the input, while the epistemic band collapses in the data-rich interior and flares in the data-sparse tails, exactly the behaviour a model should exhibit as it moves away from its training support.

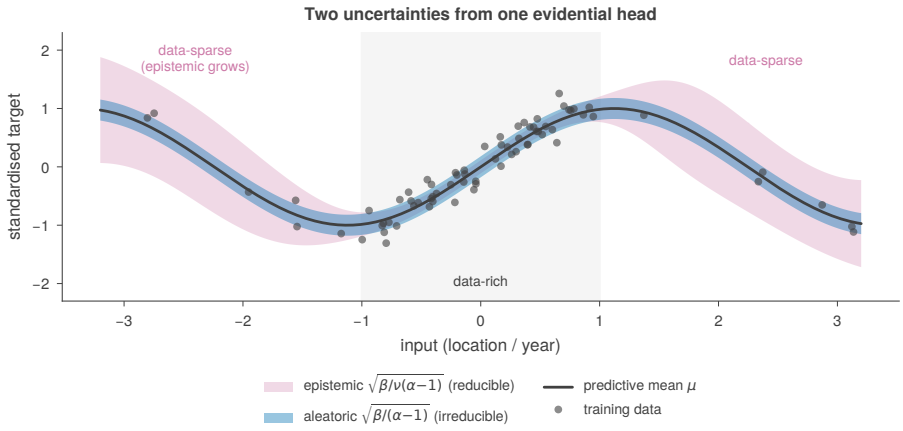

Figure 9: **Two uncertainties from one evidential head.** The aleatoric band $\sqrt{\beta/(\alpha-1)}$ (blue) is the irreducible data noise and is roughly constant; the epistemic band $\sqrt{\beta/(\nu(\alpha-1))}$ (purple) is the model's ignorance and collapses where data is dense, flaring in the data-sparse tails. Both are read off the single NIG output $(\mu, \nu, \alpha, \beta)$ of Eq. (19).

A.7.5 Uncertainty Quantification and Foundation Models

A foundation model is, by construction, applied *beyond* the exact distribution it was trained on: it is queried at new tasks, new years, and new regions, and its outputs feed downstream analyses it was never directly optimised for. Three consequences make a calibrated, decomposed uncertainty a first-class output of TerraNova rather than an afterthought.

(i) **Decision-readiness.** Climate-risk, exposure, and socioeconomic-projection decisions require intervals, not point estimates: a prediction of national vulnerability or local exposure is actionable only with a defensible confidence band. The Student-t predictive (Eq. 22) supplies a per-query interval directly, in physical units after de-standardisation.

(ii) **Extrapolation awareness.** Epistemic uncertainty is the model's internal signal that a query is out-of-distribution: a future year, a data-poor country, an unusual covariate combination. A foundation model whose reported epistemic uncertainty *rises* as it leaves its training support is self-aware about extrapolation; a bare point predictor gives no such warning and fails silently. This is especially important for the Anthropocene setting, where the most policy-relevant queries (future decades, rapidly changing regions) are precisely the least-observed.

(iii) **Compute allocation.** The same uncertainty signal closes a loop with training: the active-resampling sampler (§A.19) directs gradient toward the tasks the model is most uncertain about, so uncertainty is not merely reported but *used* to improve the model.

This is why TerraNova emits a full NIG distribution at every query rather than a mean, and why the reconstruction loss (§A.8) is constructed to keep the two uncertainties identifiable on standardised targets rather than collapsing them.

A.7.6 Embeddings as the joint carrier of signal for prediction and its uncertainty

A theme that runs through the entire architecture is that TerraNova's computation is mediated by *embeddings*: the location, country, time, task, and spatiotemporal vectors of Part B, each rescaled to a fixed radius on a sphere (Eq. N). These embeddings are not an internal convenience to be discarded at the readout: they **are** the object the model actually learns, and they are deliberately made to carry more than the information needed to predict a conditional mean. Three design facts make this explicit.

First, the spatiotemporal embedding $\mathbf{z}_{\text{st}}$ is consumed by *two* different heads with different objectives: the task-conditioned decoder (§A.7), which reads it to predict, and the alignment projections (§A.9.2), which read it to be contrasted against external and country embeddings. A single vector must therefore be simultaneously *predictive* and *retrievable*, which forces its geometry to encode structure that a prediction-only model would never need.

Second, because the decoder is *evidential*, the embedding must carry the information that lets the hypernetwork emit not just μ but a calibrated (ν, α, β). To produce a small ν (high confidence) the network must "know", from the embedding alone, that the query lies in a well-observed part of the input space; to produce a large ν it must recognise an unfamiliar one. The embedding therefore encodes an implicit notion of *how well-determined* the target is at that location and time alongside its predictive content. In this sense the uncertainty is not bolted on at the head; it is a property of where the query lands in embedding space.

Third, the auxiliary objectives act *on the embeddings rather than on the predictions*. External geospatial alignment (§A.9.2) shapes the spatiotemporal embedding so that nearby and visually similar places are close; internal country–location alignment (§A.9.3) makes a country embedding and the mean embedding of its coordinates adjacent; temporal transport (§A.4.5) regularises the geometry of the time embedding. Most of the model's objectives, and a large fraction of its parameters, are spent shaping embedding geometry, because it is the embedding that must transfer across tasks, geometries, and time. The quality of TerraNova as a foundation model is, to first order, the quality of these embeddings: their accuracy, their calibration, their cross-geometry correspondence, and their alignment to external representations are all properties of the same learned vectors.

Part C. Learning objective

The total objective (Eq. 35) is a sum of one reconstruction loss (§A.8) and three auxiliary losses (§A.9), each realised stochastically by the mixed-step schedule and scaled by a fixed category weight. This Part derives the evidential reconstruction loss from the predictive

distribution of §A.7.3, explains the two corrections that make the evidential split well-posed, and then specifies each auxiliary objective and its weight. Homoscedastic (Kendall) task weighting (Kendall et al., 2018) is disabled in the final model.

A.8 Reconstruction: evidential NIG loss

For a z-scored target y and NIG output $(\mu, \nu, \alpha, \beta)$, write $\Omega = 2\beta(1+\nu)$. The NIG negative log-likelihood (Amini et al., 2020) is the exact negative log-density of the Student-t marginal (22),

$$\mathcal{L}_{\mathrm{NLL}} = \tfrac{1}{2}\log\frac{\pi}{\nu} - \alpha\log\Omega + \left(\alpha+\tfrac{1}{2}\right)\log\!\left(\nu(y-\mu)^2+\Omega\right) + \log\Gamma(\alpha) - \log\Gamma\!\left(\alpha+\tfrac{1}{2}\right). \quad (24)$$

Minimising (24) alone is insufficient. The original evidential regulariser $|y-\mu|\,(2\nu+\alpha)$ Amini et al. (2020) pushes the total evidence down where the residual is large, but Meinert et al. (2023) show it drives a $\nu \to 0$ degeneracy and, more fundamentally, that the aleatoric/epistemic split is *not* fully separately identifiable from the marginal likelihood without a prior on one of the two scales: a genuinely noisy region (large aleatoric variance) and an epistemically uncertain one produce the same marginal and are conflated. We adopt two corrections. First, we normalise the residual by the identifiable Student-t width before weighting by the total evidence, so the penalty is on *standardised* error rather than raw noise level:

$$\mathcal{R}_{\mathrm{evi}} = \left|\frac{y-\mu}{w_{\mathrm{St}}}\right|^{p}(2\nu+\alpha), \qquad p=1. \quad (25)$$

Second, we add a soft one-sided cap on the predictive width: because targets are z-scored, a width beyond a few σ is implausible, and this prior supplies the kind of information on one of the two scales that Meinert et al. (2023) argue is required, alongside a smoothness assumption in x, to make the split well-posed, while closing the $w_{\mathrm{St}} \to \infty$ escape hatch that the normalised regulariser would otherwise leave open:

$$\mathcal{R}_{\mathrm{cap}} = \mathrm{relu}\!\left(w_{\mathrm{St}} - W_{\max}\right)^2, \qquad W_{\max} = 3. \quad (26)$$

The per-sample reconstruction loss, mean-reduced over the batch (with optional off-support sample weights w_i; §A.15), is

$$\boxed{\mathcal{L}_{\mathrm{rec}} = \left\langle w_i\left[\mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{ev}}\,\mathcal{R}_{\mathrm{evi}} + \kappa\,\mathcal{R}_{\mathrm{cap}}\right]\right\rangle, \qquad \lambda_{\mathrm{ev}} = 0.3,\ \kappa = 0.1.} \quad (27)$$

All NIG terms are evaluated in `fp32` with degenerate-parameter floors ($\nu \geq 10^{-8}$, $\alpha - 1 \geq 10^{-6}$, $\beta \geq 10^{-8}$) for numerical safety, since the log and $\log\Gamma$ terms are unstable in `bf16`; the reconstruction category weight is $w_{\mathrm{cat}}^{\mathrm{rec}} = 1$.

Design rationale. The training-policy study sweeps the evidence coefficient over $\lambda_{\mathrm{ev}} \in \{0.1, 0.3\}$ jointly with the curriculum and active-learning switches and selects 0.3: relative to 0.1 it modestly lowers reconstruction error and improves calibration, so accuracy and calibration are positively associated rather than in tension at this setting. The width cap and its weight κ are held fixed across the study. Replacing a bare MSE with this evidential objective is what lets TerraNova emit a calibrated, decision-ready uncertainty with every point prediction; the same uncertainty also drives active resampling (§A.19), so loss and the training policy are coupled.

A.9 Auxiliary objectives

A.9.1 Temporal transport

For a fixed (τ, u) observed at source year t and target year $t + \Delta$, the model forms time embeddings $\boldsymbol{\tau}_s, \boldsymbol{\tau}_{s+\Delta}$; the transported time $T(\boldsymbol{\tau}_s, \Delta)$ is fed back through fusion and the decoder to give a transported prediction $\hat{y}^{\mathrm{tr}}$. Four terms, with stop-gradient targets, supervise transport simultaneously as a *predictor* and as a well-behaved *operator*:

$$\mathcal{L}_{\mathrm{pred}} = \mathcal{L}_{\mathrm{rec}}^{\mathrm{NIG}}\big(\hat{y}^{\mathrm{tr}},\, y_{t+\Delta}\big) \qquad \text{(transported prediction is accurate)}, \tag{28}$$

$$\mathcal{L}_{\mathrm{match}} = \big\| T(\boldsymbol{\tau}_s, \Delta) - \mathrm{sg}[\boldsymbol{\tau}_{s+\Delta}] \big\|_2^2 \qquad \text{(transport lands on the true target-year embedding)}, \tag{29}$$

$$\mathcal{L}_{\mathrm{id}} = \big\| T(\boldsymbol{\tau}_s, 0) - \mathrm{sg}[\boldsymbol{\tau}_s] \big\|_2^2 \qquad \text{(zero shift is the identity)}, \tag{30}$$

$$\mathcal{L}_{\mathrm{sg}} = \big\| T\big(T(\boldsymbol{\tau}_s, \Delta_1), \Delta_2\big) - \mathrm{sg}[T(\boldsymbol{\tau}_s, \Delta_1{+}\Delta_2)] \big\|_2^2 \qquad \text{(semigroup)}, \tag{31}$$

with $\Delta \in \{\pm1, \pm2, \pm4, \pm6, \pm8\}$ (offsets beyond $\Delta_{\max} = 4$ saturate the normalised-offset clamp of Eq. 9) and semigroup pairs $(\Delta_1, \Delta_2) \in \{(1,1), (1,2), (2,2), (4,4), (2,6)\}$ drawn per step. Here $\mathcal{L}_{\mathrm{rec}}^{\mathrm{NIG}}$ is the same evidential prediction criterion as Eq. 27 (NIG NLL plus the evidence regulariser and width-cap penalty), evaluated on the transported prediction rather than a separate bare NLL. The family loss is the fixed combination

$$\mathcal{L}_{\mathrm{tr}} = w_{\mathrm{cat}}^{\mathrm{tr}}\big(0.35\,\mathcal{L}_{\mathrm{pred}} + 0.10\,\mathcal{L}_{\mathrm{match}} + 0.02\,\mathcal{L}_{\mathrm{id}} + 0.05\,\mathcal{L}_{\mathrm{sg}}\big), \qquad w_{\mathrm{cat}}^{\mathrm{tr}} = 1. \tag{32}$$

The stop-gradients prevent the consistency terms from collapsing the time encoder itself (they constrain T, not $\boldsymbol{\tau}$); the weighting emphasises predictive accuracy while keeping the operator geometrically well-behaved.

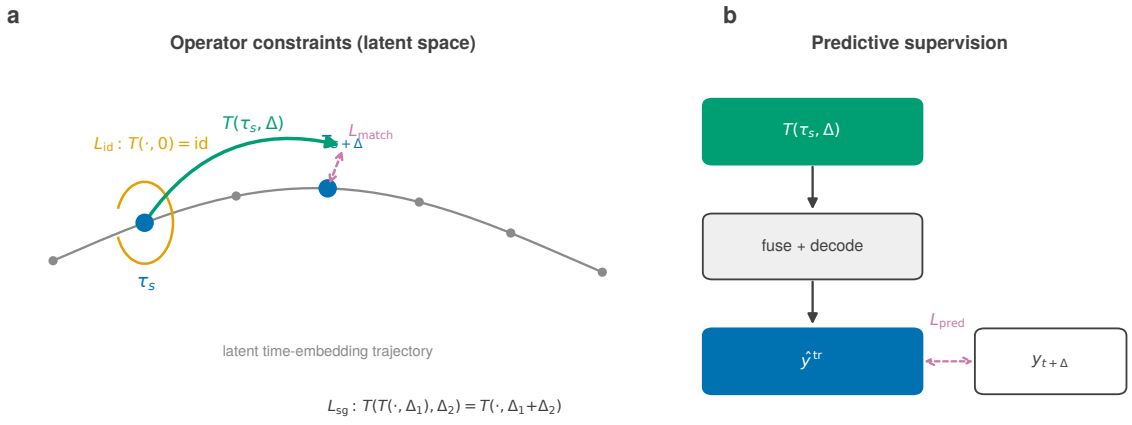

Figure 10: The four temporal-transport constraints (Eq. 32). **(a)** In latent time-embedding space, the operator $T(\boldsymbol{\tau}, \Delta)$ should land on the true target-year embedding ($\mathcal{L}_{\mathrm{match}}$), be the identity at zero shift ($\mathcal{L}_{\mathrm{id}}$), and compose additively ($\mathcal{L}_{\mathrm{sg}}$, semigroup). **(b)** The transported embedding, once fused and decoded, must also predict the true future target ($\mathcal{L}_{\mathrm{pred}}$). Targets are stop-gradient, so the consistency terms shape T without collapsing the time encoder.

A.9.2 External geospatial alignment (spatiotemporal)

This objective aligns TerraNova's *spatiotemporal* embedding $\mathbf{z}_{\text{st}}$ to frozen external geospatial embeddings: SatCLIP (Klemmer et al., 2025), GeoCLIP (Vivanco Cepeda et al., 2023), and Copernicus-Embed (Wang et al., 2025), sampled with equal probability per step (§A.13). The targets are static one-vector-per-coordinate products, but $\mathbf{z}_{\text{st}}$ is time-dependent, so the alignment is anchored in time: each bank is queried at the year(s) in which *its own* training data were generated (the ref. years of Table 6). One such year is drawn uniformly per batch and used as the query time, so $\mathbf{z}_{\text{st}}$ is always read in the bank's native temporal context rather than at an arbitrary date; the contrast is thus spatiotemporal on the TerraNova side and matched in time to each target.

The bridge between the two spaces is a small learned head. The source is projected by a per-bank two-layer MLP $P_{\text{h}} : \mathbb{R}^{256} \to \mathbb{R}^{512}$ (Linear $\to$ LN $\to$ SiLU $\to$ Linear, hidden 512) into the bank's 512-D space. SatCLIP and GeoCLIP are natively 512-D and are taken directly; Copernicus-Embed is 768-D and passes through a single linear projection to 512 first. Both sides are then ℓ_2-normalised. They are contrasted with cosine similarity $s(a, b) = \hat{a}^\top \hat{b}$ and temperature $\tau_{\text{T}} = 0.07$ against $K = 32$ inverse-distance (hard) negatives $\{\mathbf{v}_k^-\}$, in the sampled-negative InfoNCE form (van den Oord et al., 2018; Radford et al., 2021):

$$\mathcal{L}_{\text{geo}} = -\log \frac{\exp\big(s(P_{\text{h}}\mathbf{z}_{\text{st}}, \mathbf{v}^+)/\tau_{\text{T}}\big)}{\exp\big(s(\cdot, \mathbf{v}^+)/\tau_{\text{T}}\big) + \sum_{k=1}^{K} \exp\big(s(\cdot, \mathbf{v}_k^-)/\tau_{\text{T}}\big)}, \qquad w_{\text{cat}}^{\text{geo}} = 1. \tag{33}$$

Negatives are drawn from a softmax over inverse distances, $p(x_k^-) \propto \exp\big((d_{\text{hav}} + \varepsilon)^{-1}/T_{\text{neg}}\big)$ with $T_{\text{neg}} = 0.07$, beyond a 50 km exclusion radius, where d_{hav} is the great-circle (haversine) distance, so the contrast is tilted toward the hardest negatives (nearby but distinct locations), which forces the source embedding to resolve fine spatial distinctions rather than merely global gradients.

Design rationale. Our ablation treats external alignment as an auxiliary *representation* objective rather than a prediction task, and sweeps the loss family (inverse- vs linear-distance weighting), the per-epoch step budget, and the loss weight against a no-alignment baseline. Equal-weighted inverse-distance alignment to all three banks at the largest budget is selected: it raises top-1 retrieval into the external spaces from near-chance (0.008) to 0.875 and, crucially, improves gridded held-out correlation by $\Delta \geq 0.05$ without measurable loss of reconstruction accuracy. Because the external banks are visual (satellite imagery for SatCLIP and Copernicus, ground-level photographs for GeoCLIP), this is evidence that anchoring the spatiotemporal space to image-derived structure regularises it toward better spatial generalisation, consistent with the *platonic representation hypothesis* (Huh et al., 2024): models trained on different modalities of the same world converge toward a shared representation of its underlying state. The intuition is specific and load-bearing here, because it justifies spending compute on visual contrastive losses while the prediction targets are numeric. A satellite tile or a ground-level photograph is a dense, high-bandwidth observation of the *physical* character of a place, its land cover, terrain, built form, vegetation and hydrography, which the frozen image encoder has already distilled into structure; regression onto numeric targets alone never sees that signal and is free to arrange its latent space in whatever way happens to fit the training fields. Pulling $\mathbf{z}_{\text{st}}$ toward the image-derived banks therefore supplies an inductive bias the targets cannot: it forces the latent geometry to encode intrinsic physical features of the Earth's surface, so that places are separated by what they physically *are* rather than only by where their training values happened to land. This is why a visual contrastive loss improves *numeric* prediction at all, and why it helps most where tabular signal is sparse, a data-poor cell or country inherits a physically grounded location representation from imagery even when its own labels are few. Country-level tasks see a small accuracy reduction (these visual losses are not designed for them), which the internal country–location objective (§A.9.3) is introduced to address.

A.9.3 Internal country–location alignment (spatiotemporal)

This is the objective that makes the two native geometries mutually informative, and like the external alignment it operates in *spatiotemporal* space. One year y is drawn per batch from the years for which a gridded population-density raster is available, and that single year does double duty: **(i)** it is the query time passed to the time encoder for *both* the country embedding and every in-country coordinate embedding, and **(ii)** it selects the contemporaneous population raster P_y that weights the coordinate sampler. The two effects are complementary. The first compares the country code and the coordinate codes within a single temporal slice rather than as static maps. The second makes the spatial prior over positives itself time-consistent: because the population field is re-read for each sampled year, the positives track where people *actually were* in year y, so as populations redistribute over decades (urbanisation, regional growth and decline) an early-year batch concentrates its positives differently from a recent-year batch, instead of re-using a fixed present-day mask. Concretely, for a sampled country c with polygon support Ω_c we draw $C = 128$ positive coordinates inside Ω_c with year-specific density $q_y(x) \propto \log(1 + P_y(x))$, where P_y is the population-density raster for year y, and $K = 16$ negatives from other countries. Mean-pool the positive coordinates' spatiotemporal embeddings, $\bar{\mathbf{h}}_c^+ = \frac{1}{C} \sum_{x \in P_c} \mathbf{z}_{\text{st}}(x)$, and

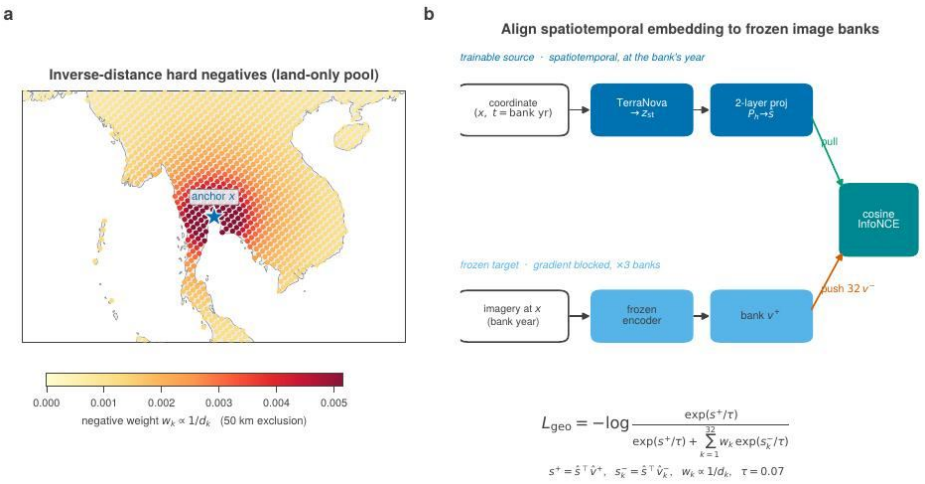

Figure 11: External geospatial alignment (Eq. 33). **(a)** Hard negatives are drawn from a softmax over inverse great-circle distances beyond a 50 km exclusion radius around the anchor, so the contrast is tilted toward nearby-but-distinct locations. **(b)** The trainable spatiotemporal embedding is projected to $\hat{s}$ and pulled toward the positive bank vector $\mathbf{v}^+$ produced by a *frozen* image encoder, while being pushed from 32 hard negatives, by sampled-negative InfoNCE at temperature τ_{T}=0.07. The three banks (SatCLIP, GeoCLIP, Copernicus-Embed) are rotated across steps.

align the country's own spatiotemporal embedding $\mathbf{h}_c = \mathbf{z}_{\mathrm{st}}(c)$ (the country routed through the geometry router, then ST-fused at the same year) to this within-country mean by sampled-negative InfoNCE ($\tau_{\mathrm{T}} = 0.07$):

$$\mathcal{L}_{\mathrm{cl}} = -\log \frac{\exp\big(s(\mathbf{h}_c, \bar{\mathbf{h}}_c^+)/\tau_{\mathrm{T}}\big)}{\exp\big(s(\mathbf{h}_c, \bar{\mathbf{h}}_c^+)/\tau_{\mathrm{T}}\big) + \sum_{x^- \in N_c} \exp\big(s(\mathbf{h}_c, \mathbf{h}_{x^-})/\tau_{\mathrm{T}}\big)}, \qquad w_{\mathrm{cat}}^{\mathrm{cl}} = 0.25. \tag{34}$$

Design rationale. We ablate two different alignment mechanisms: this contrastive alignment and a prediction-consistency regulariser that forces a country's prediction to match the mean of coordinate predictions inside it, and selects population-weighted alignment-only, rejecting consistency. The predictive gain concentrates on the country side of the joint benchmark, which is the informative result: because WorldTensor is not merely a gridded copy of the country panel but carries distinct variables (climate, air quality), the relatively sparse country tasks can borrow the higher-density coordinate signal through this soft coupling. The $\log(1 + P_y)$ density (borrowed from high-dynamic-range tone-mapping) compresses the heavy-tailed population field so dense cities do not monopolise the within-country sampler, while still focusing positives where people (and thus most socioeconomic signal) are. Consistency is discarded because it lowered the predictive selection metric (it is too strong a regulariser) and is therefore set to $w^{\mathrm{cons}} = 0$.

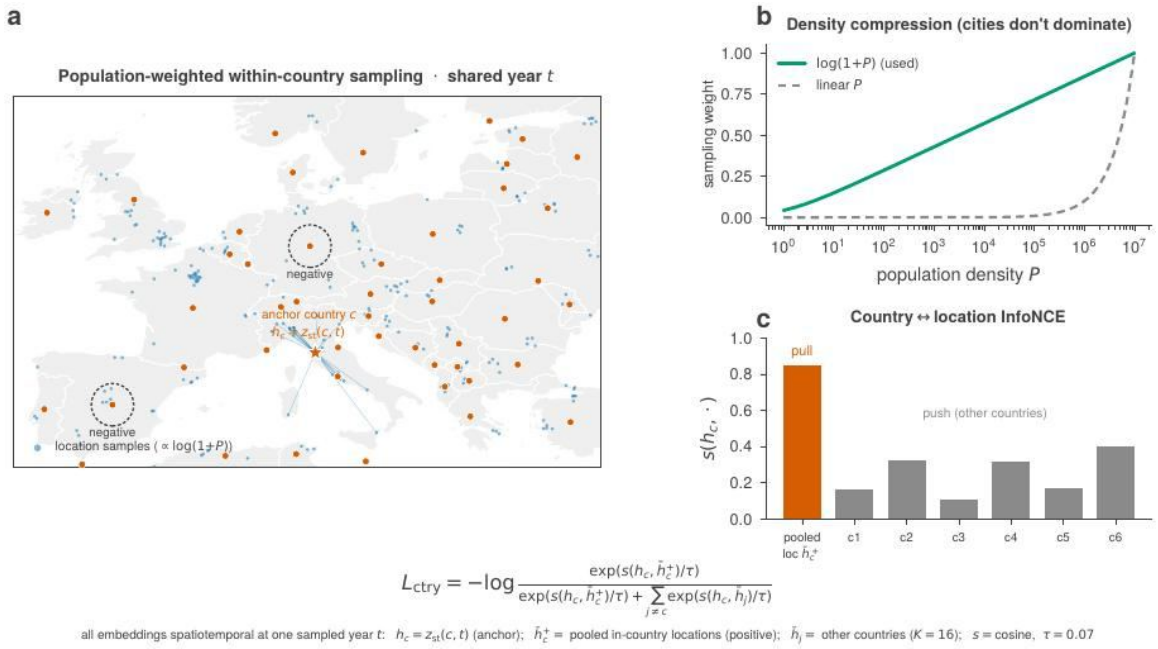

Figure 12: Internal country–location alignment (Eq. 34). **(a)** Positive coordinates are drawn inside the anchor country's polygon, population-weighted using the density raster for the sampled year; **(b)** the $\log(1 + P_y)$ density compresses the heavy-tailed population field so dense cities do not monopolise the sampler. **(c)** The country embedding $\mathbf{h}_c$ is pulled toward the mean-pooled in-country location embedding $\bar{\mathbf{h}}_c^+$ and pushed from other countries' embeddings by sampled-negative InfoNCE, softly coupling the two native geometries so the sparse country tasks can borrow the denser coordinate signal.

A.10 Total objective and category weighting

Over a mixed mini-epoch the optimiser minimises the sum of the active families' losses,

$$\boxed{\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{cl}}} \tag{35}$$

with category weights $(w^{\text{rec}}, w^{\text{tr}}, w^{\text{geo}}, w^{\text{cl}}) = (1, 1, 1, 0.25)$ and no learned homoscedastic weighting. Because each minibatch realises exactly one family (§A.17), (35) is optimised in expectation, with the step counts of Table 7 setting the effective mixing proportions.

Design rationale. Kendall homoscedastic weighting (Kendall et al., 2018) re-weights tasks by a learned per-task observation noise. It is disabled here because the evidential head already learns per-prediction scale through (ν, α, β); layering a second, per-task learned scale on top of 1,024 tasks was found redundant and was switched off in the final recipe to avoid two interacting uncertainty mechanisms competing during optimisation.

Part D. Data

TerraNova is trained on three data products: the gridded environmental fields of WorldTensor (§A.11), the country-level socioeconomic and institutional indicators of CountryTensor (§A.12), and three external geospatial embedding banks used only as alignment targets (§A.13). The first two are the reconstruction targets; all three are released separately as companion resources and are documented in dedicated data papers and code releases, which we summarise here to the extent needed to reproduce the training inputs. The model uses a curated core subset of 512 *fields from each of WorldTensor and CountryTensor.*

A.11 WorldTensor: gridded environmental fields

The gridded component is drawn from WorldTensor (Rodriguez-Pardo and Tavoni, 2026), a harmonised global dataset that aligns hundreds of environmental and human-system variables to a common annual 0.25° latitude–longitude grid (matching the ERA5 reanalysis grid; Hersbach et al., 2020) with standardised coordinates, CF-style metadata, and a NetCDF release. The full release comprises 757 variable families (52,823 NetCDF files, ≈ 46 GB) organised into 13 thematic domains plus a static-context domain, spanning 1900–2025 with domain-dependent coverage. TerraNova trains on the core-512 manifest, a 512-field subset (415 annual fields and 97 static layers) curated for coverage and quality.

Construction. WorldTensor is built around four principles: shared spatial support (the 0.25° grid), shared temporal resolution (annual), multi-domain coverage, and machine-readability. Achieving them requires more than format conversion, because sources differ in native resolution, projection, temporal structure, and geometry type. Gridded reanalysis products are regridded to the canonical grid with longitude-seam and polar handling and bilinear (continuous) or nearest-neighbour/category-specific (discrete) resampling; high-frequency products are aggregated to annual summary statistics (mean, standard deviation, min, max, or sum, as appropriate to the variable); and point, line, and polygon datasets are rasterised into spatially meaningful gridded fields (e.g. counts, cumulative quantities, active

stocks, or distance-to-event surfaces) rather than discarded. Each released field carries a short source code and a statistic suffix (e.g. `t2m_mean`, `tp_sum`, `pm2p5_std`).

Table 4: WorldTensor core-512 gridded fields by domain. Representative variables are summarised from Rodriguez-Pardo and Tavoni (2026); coverage is the released span for that domain. All fields are delivered by WorldTensor itself.

domain	#	representative variables	coverage
climate	138	2 m temperature, precipitation, sea-level pressure (mean/std/min/max)	1940–2025
static	97	elevation, slope, bathymetry, distance-to-coast/river, soil texture, travel time	static
emissions	67	CH_4, N_2O, biogenic & fossil CO_2, shipping NO_x	1970–2024
energy	52	active/added/retired/net generating capacity, plant diversity	1900–2025
air_quality	40	$PM_{2.5}$, atmospheric composition (annual summaries)	2003–2024
hazards_and_conflict	28	earthquake, cyclone, volcanic, disaster, conflict counts/intensities	1900–2025
human_systems	22	population density/count, sectoral GDP, nighttime lights, built surface, human footprint	1972–2024
extremes	14	land & marine heatwave indices, SPI/SPEI drought	1901–2025
land_use	14	primary/secondary forest, pasture, rangeland, urban, crop fractions	1900–2024
agriculture	12	crop production, livestock density, N/P/K fertiliser application	1961–2021
vegetation	10	NDVI, EVI, burned area, vegetation optical depth	2000–2025
hydrology	8	terrestrial water storage, soil moisture, snow water equiv., wetland inundation	2000–2025
cryosphere	6	snow cover, glacier mass/area, permafrost extent/thickness	1976–2025
ocean	4	chlorophyll-a (mean/std/min/max)	2010–2023
			total: 512 fields

Domain inventory. Table 4 lists the core-512 domains, representative variables, and coverage; the per-domain field counts are visualised in Figure 13. Coverage is deliberately heterogeneous across domains: climate reaches back to 1940, land use and energy to 1900, while satellite-derived ocean and vegetation products begin in the 2000s. The annual resolution is a design choice that enables cross-domain alignment at the cost of attenuating sub-annual dynamics; the 0.25° grid balances detail, global completeness, and tractability (Rodriguez-Pardo and Tavoni, 2026).

A.12 CountryTensor: socioeconomic and institutional indicators

The country component is 512 national indicators derived from the Quality-of-Government (QoG) data ecosystem of the University of Gothenburg (Teorell et al., 2021), indexed by ISO3 over an annual panel that follows the QoG time-series coverage (most series 1946–2020, with some historical institutional series reaching earlier), within the model's temporal normalisation range (Eq. 8). The variables are grouped into 65 thematic families (which roll up into 10 higher-level buckets) spanning human development, political economy, political institutions, governance and rights, inequality, demography, conflict, health, and climate vulnerability (all 65 are listed in Table 5; the family-size distribution is shown in Figure 14). Of the 512 variables, 338 are continuous and 174 are one-hot/binary category indicators; *all* are trained as scalar regression targets under the evidential loss (Eq. 27), and the binary metadata carried by the data export is deliberately ignored so that a single objective and head serve every task.

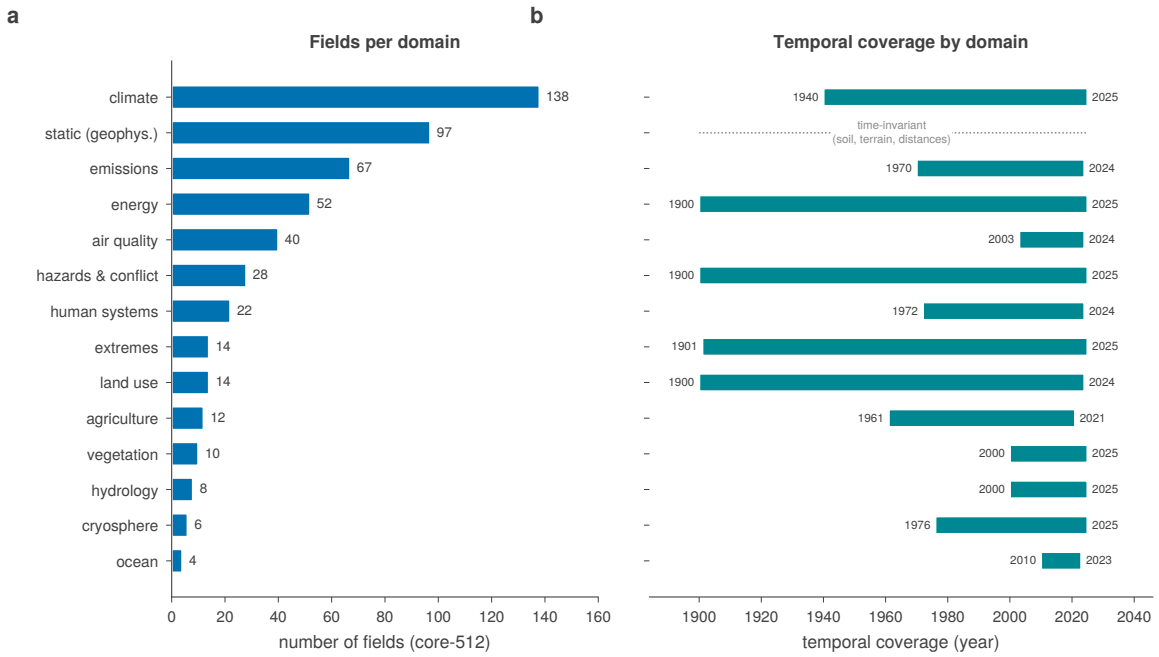

Figure 13: WorldTensor core-512 inventory (real counts and spans from the manifest and Table 4). **(a)** Fields per domain: physical domains dominate by count, but task sampling (§A.17) uses a socioeconomic-first selection order and explicit boosts so that policy-relevant human-system fields are not crowded out. **(b)** Temporal coverage is deliberately heterogeneous: energy, land use and hazards reach back to 1900, climate to 1940, while satellite-derived vegetation, hydrology and ocean products begin only in the 2000s. The "static (geophys.)" domain (97 fields) is time-invariant and therefore has no temporal extent: it collects fixed geophysical descriptors: SoilGrids soil properties and texture (Poggio et al., 2021) (72 fields over standard depth profiles), vegetation/land-cover class, topography, distance-to-coast/river, bathymetry, that provide spatial context shared across all years. All time-varying series are aligned to the common annual grid and to the model's temporal normalisation range (Eq. 8).

Table 5: CountryTensor: all 65 QoG-derived thematic families (by variable count). At a higher level these roll up into 10 buckets: human welfare (116), macro development (111), institutions & governance (107), climate & environment (65), historical/structural (30), conflict/external (22), political process (20), social attitudes (4), regional (1), and a residual *other* (36). Total 512 variables (338 continuous, 174 binary), all regressed.

family	#	family	#	family	#
Human development	43	Aid & development	6	Economic complexity	2
Political economy	39	Historical institutions	6	Freedom House political	2
Political institutions	35	Living-conditions qual.	6	Geography & history	2
Economy & development	34	Representation inequal.	6	Hist. econ. development	2
Governance & rights	31	Climate disaster risk	5	Hist. infrastructure	2
Environment & energy	26	Gender parliament repr.	5	Informal economy	2
Inequality & living std.	23	Regime classification	5	Peace & security	2
Regime/historical typ.	18	State antiquity/history	5	Polity regime	2
Demography & labour	14	Dependency ratios	4	Prosperity & poverty	2
Climate vulnerability	13	Electoral-system design	4	State capacity (latent)	2
Conflict & security	13	Ethnocultural diversity	4	Top-income concentr.	2
Financial integ./trade	12	Gender gap	4	Uncertainty	2
Health	12	Health outcomes/longev.	4	Country risk	1
Agriculture & land use	10	Human-rights violations	4	Electoral attitudes	1
Elections & events	9	Party-system dynamics	4	Environ. footprint (ext.)	1
Hist. state formation	9	Tax-revenue structure	4	Freedom House nations	1
Public-sector spending	9	Culture & values	3	Governance capacity	1
Regime failure/transition	8	Digital/info. access	3	Metabolic health	1
Ecological footprint	7	Remittances & diaspora	3	Regional integration	1
External relations/sec.	7	Resource/fuel depend.	3	Screening & prevention	1
Gender repr. systems	7	Voter turnout/compet.	3	State fragility	1
Macro trade & growth	7	Basic capabilities	2		

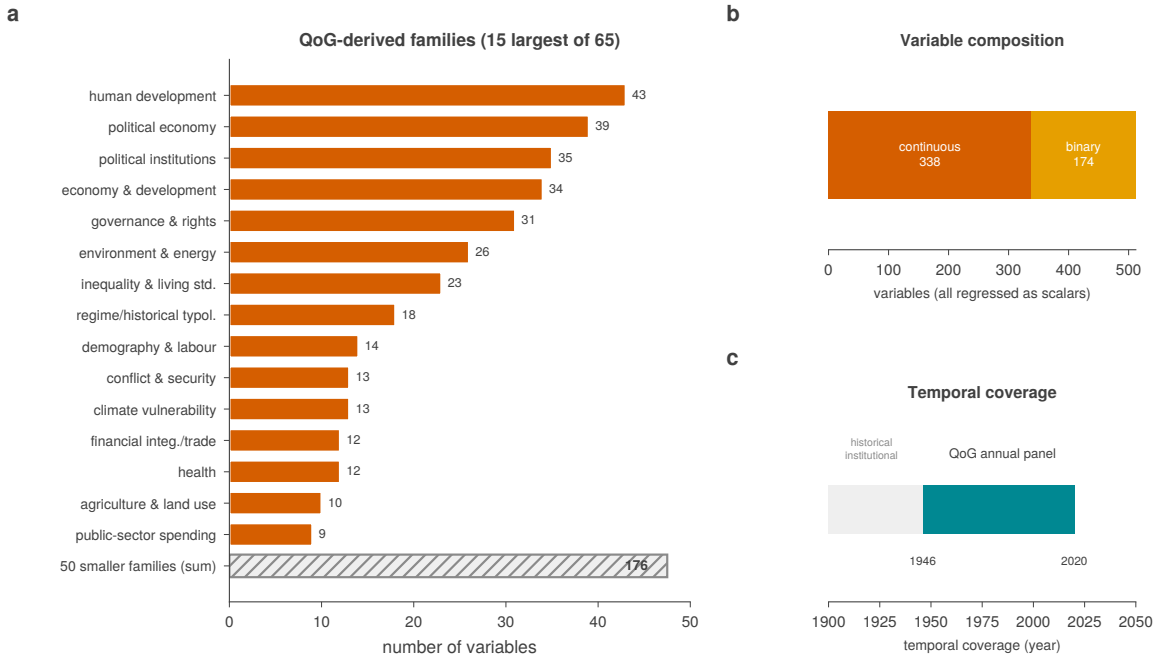

Figure 14: CountryTensor inventory (real counts from `countrytensor_core_512.yaml`). **(a)** Variable counts for the 15 largest QoG-derived families; the long tail of 50 further families is shown as a single hatched aggregate (bar clipped, true total 180 variables) so it does not visually dominate the named families. **(b)** All 512 variables (338 continuous, 174 one-hot/binary) are trained as scalar regression targets under the evidential loss. **(c)** The national panel is annual, following the QoG time-series (1946–2020 for most series, with some historical institutional series extending earlier), indexed by ISO3 within the model's temporal normalisation range (Eq. 8).

A.13 External geospatial embedding banks

The alignment targets of §A.9.2 are precomputed embedding banks over a fixed 500k land-coordinate pool (`land_only_500k`), generated with a companion geospatial-embeddings resource (Rodriguez-Pardo, 2026). The generation pipeline (Algorithm 2) samples candidate coordinates with a Fibonacci-sphere lattice (uniform on the sphere, with the lattice phase randomised per run so repeated runs do not reuse the same locations), filters them to land using Natural Earth polygons (Natural Earth, 2022), and embeds the surviving coordinates with each wrapped encoder. Banks are read-only and never updated during training; the source projections in the alignment loss are the only trainable part of that pathway.

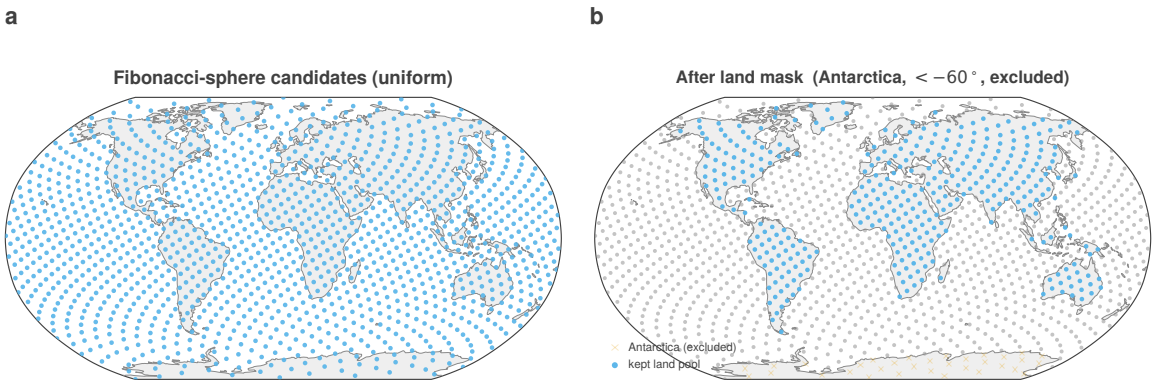

Figure 15: Land-coordinate pool construction (Algorithm 2). **(a)** A Fibonacci-sphere lattice places candidate points uniformly over the globe (equal-area, with no clustering at the poles, unlike a latitude–longitude grid). **(b)** Candidates are filtered to land with Natural Earth polygons (Natural Earth, 2022), leaving the surviving coordinates as the fixed 500k land pool over which all three embedding banks are precomputed.

Table 6: External geospatial embedding banks used as alignment targets (the three static products from the wrapper used by TerraNova). Each stores coordinates and one embedding vector per point over the 500k land pool, at its own encoder's native width.

bank	file	dim	ref. years	modality
SatCLIP	`satclip_land_500k.pt`	512	2021–2023	satellite imagery
GeoCLIP	`geoclip_land_500k.pt`	512	2004–2014	ground-level images
Copernicus	`copernicus_land_500k.pt`	768	2021	Copernicus rasters

Algorithm 2 Geospatial embedding-bank generation (Rodriguez-Pardo, 2026)

Require: point budget N ($= 500{,}000$), encoder set $\mathcal{E} = \{\text{SatCLIP, GeoCLIP, Copernicus}\}$

$\{x_i\}_{i=1}^{M} \leftarrow \textsc{FibonacciSphere}(N', \text{phase} \sim \mathcal{U})$ ▷ $N' \geq N$ candidates, randomised phase

$\{x_i\} \leftarrow \{x_i : x_i \in \textsc{NaturalEarthLand}\}$ ▷ land mask; keep first N

for each encoder $E \in \mathcal{E}$ **do**

  $\mathbf{V}^E \leftarrow [\, E(x_i) \,]_{i=1}^{N} \in \mathbb{R}^{N \times d_E}$ ▷ batched embedding

return bank $= \{\texttt{coordinates\_lonlat}= \{x_i\}, \{E\texttt{\_embeddings}= \mathbf{V}^E\}_{E \in \mathcal{E}}\}$

Design rationale. The Fibonacci-sphere lattice is used because a naïve latitude–longitude grid over-samples the poles, biasing any contrastive loss toward high-latitude geometry; the Fibonacci construction is uniform on the sphere and deterministic up to its phase, so the land pool is geographically balanced. Land filtering restricts the banks to the support of the alignment task (the external encoders are defined on land), and the shared coordinate pool across all three banks means the three alignment objectives see the same locations, so their gradients are comparable. The same Fibonacci-on-sphere construction, intersected with the WorldTensor ocean mask, supplies the open-ocean coordinates for the `OCN` synthetic country (§A.15).

Part E. Preprocessing

Three preprocessing mechanisms turn the raw data products into training-ready targets: per-task standardisation (§A.14), which places all targets in a common z-scored space; the support policy and ocean handling (§A.15), which deal with fields that are undefined over parts of the globe; and the train/validation split (§A.16), which defines a held-out-time generalisation test. Each is described with the exact rule the code applies, because all three interact with the loss (Eq. 27) and the evidential caps (Eq. 20).

A.14 Per-task standardisation

A central registry stores per-task standardisation parameters and applies one of three invertible transforms. The reconstruction objective uses *standardisation* (z-scoring), with per-task mean and standard deviation $(\mu_\kappa, \sigma_\kappa)$ estimated on the training split:

$$\tilde{y} = \frac{y - \mu_\kappa}{\sigma_\kappa}, \qquad y = \tilde{y}\,\sigma_\kappa + \mu_\kappa. \tag{36}$$

Two further strategies are supported for diagnostics and for variables better suited to a bounded range: min–max scaling to $[0, 1]$, $\tilde{y} = (y - y_{\min})/(y_{\max} - y_{\min})$, and a robust transform that maps the 5th/95th percentiles to $[0, 1]$ and clips, $\tilde{y} = \text{clip}\big((y - q_{05})/(q_{95} - q_{05}), 0, 1\big)$, which is less sensitive to heavy tails. Non-finite inputs are sanitised ($\pm\infty \to \text{NaN}$) before normalisation so that masked pixels are excluded from the loss and metrics rather than corrupting the per-task statistics.

Design rationale. Standardisation is the final choice because it is what makes the evidential machinery well-posed. The width cap $W_{\max} = 3$ and the variance cap $\sigma_{\max} = 3$ are meaningful only because every target has unit marginal variance: "three standard deviations" is then a universal, task-independent bound, so a single pair of caps serves all 1,024 tasks (§A.7.2, §A.8). It also equalises the loss scale across tasks of wildly different physical units (temperature in kelvin, GDP in dollars), so no task dominates the gradient by virtue of its units, and the per-task RMSE used for model selection (§A.20) is comparable across tasks.

A.15 Support policy and ocean handling

Many fields are undefined over part of the globe: land-domain quantities (population, built surface, land use) have no value over the ocean, and ocean-domain quantities have no value over land. A *support policy* resolves this per field. Off-support cells are filled with a deterministic value (typically 0, e.g. population over open ocean), so the field is dense and gridded, and each policy carries a per-sample weight $w_i = w_{\text{off}} \leq 1$ applied to those imputed pseudo-labels in Eq. (27); on-support cells carry $w_i = 1$. The gridded (WorldTensor) policies keep $w_{\text{off}} = 1$, supervising the filled cells at full weight so the model learns where each quantity is defined; the country-level land-zero policy down-weights them to $w_{\text{off}} = 0.8$. The active-learning sampler additionally drops imputed cells from its uncertainty estimate (§A.19), so they do not distort task reweighting.

The land/ocean classification used by the support policy and the country/location samplers is a composite mask built once from the static WorldTensor fields,

$$\texttt{is\_land} = (\texttt{bathymetry\_elevation} \geq 0) \ \vee \ \text{isfinite}(\texttt{land\_area}), \tag{37}$$

which keeps inland lakes and Greenland as land (the `land_area` disjunct catches points where the bedrock dips below sea level) while classifying open ocean as ocean and excluding Antarctica whenever its non-negative bathymetry marks it as land. A synthetic country `OCN` is registered for the samplers; its coordinates are drawn from a deterministic Fibonacci-sphere lattice (§A.13) intersected with the ocean mask, uniform on the sphere with no polar clustering, and it receives $\texttt{ocn\_sampling\_fraction} = 0.1$ of the country-sampling weight so the model gets meaningful open-ocean exposure despite only a minority of country tasks having defined ocean values.

A.16 Train/validation splits

All splits use $\texttt{split\_seed} = 42$. Gridded fields use a *temporal-year* holdout: for each field, $\texttt{gridded\_holdout\_years} = 3$ whole years are sampled for validation from the *interior* of the record (Algorithm 3), the two earliest and two latest years are excluded from the holdout pool, so the model is never asked to extrapolate beyond the temporal envelope it was trained on, only to interpolate unseen interior years. Country variables use a per-pool validation fraction $\texttt{val\_fraction} = 0.02$. Validation during training is therefore a held-out-*time* generalisation test: the model must reconstruct unseen years of seen fields, which directly probes the temporal pathway (§A.4.4, §A.4.5) rather than spatial extrapolation.

Algorithm 3 Temporal-year holdout per gridded field

Require: sorted years $y_{0..n-1}$, holdout count H $(= 3)$, seed
1: **if** $n \geq H + 4$ **then** interior $\leftarrow y_{2..n-3}$ ▷ exclude two edge years each side
2: **else** interior $\leftarrow y_{0..n-1}$
3: $H' \leftarrow \min(H, |\text{interior}|, n - 1)$
4: holdout $\leftarrow$ sorted H' years sampled without replacement from interior
5: **return** train = years $\setminus$ holdout, val = holdout

Part F. Training

Training interleaves the five step families on a fixed mixed schedule (§A.17), optimised with AdamW under a single warmup–cosine schedule with weight EMA (§A.18), and is steered by an uncertainty-driven active-resampling callback (§A.19). The full loop is Algorithm 4. This Part gives the schedule arithmetic, the task-sampling mechanics, the optimiser settings, and the active-learning algorithm, together with the study that fixed the three training-time switches.

Algorithm 4 TerraNova training loop

1: restore RNGs, optimiser, EMA, and AL state from `last.ckpt` if present
2: **for** epoch $e = e_0, \ldots, E-1$ **do**
3: $\mathcal{S} \leftarrow$ MixedSchedule($\{n_{\text{fam}}\}$, window 50) ▷ Alg. 5
4: **for** each step type $\zeta \in \mathcal{S}$ **do**
5: draw a batch for family ζ from its provider ▷ task-sampler probs for reconstruction
6: $\mathcal{L}_\zeta \leftarrow$ family loss (Eqs. 27–34)
7: backprop; clip $\|\nabla\|_2 \leq 1$; AdamW step; update EMA
8: step LR schedule (Eq. 39)
9: **if** $e \equiv 0 \pmod 2$ **then** run active resampling ▷ Alg. 6
10: **if** $e \equiv 0 \pmod 3$ **then** validate; checkpoint best-by-`val/recon_rmse_mean`
11: checkpoint (atomic) every 2 epochs; permanent milestone every 10
12: **if** walltime budget exhausted **then** graceful stop + checkpoint + exit-42

A.17 Mixed-step schedule

One *epoch* interleaves the four active families in the per-epoch budget of Table 7. The sequence is built by expanding the exact multiset of step tokens, shuffling it globally, then chunking the shuffled stream into mini-epoch windows of 50 steps and reshuffling each chunk (Algorithm 5). This prevents long homogeneous runs while preserving the exact family counts; balance within any single 50-step window is statistical rather than guaranteed by construction. The three geospatial-alignment heads split the 256 geo steps as evenly as integer counts allow, with the single remainder step rotating across epochs.

Table 7: Per-epoch step budget..

family	steps/epoch	notes
reconstruction	32768	85/15 gridded/country base (AL-adjusted)
transport	512	batch 512
geo_alignment	256	3 heads, rotating remainder, batch 512
country_alignment	256	128 countries, 128 coords, 16 negatives

Modality balance and task sampling. Within the 32768 reconstruction steps, the gridded/country split is *initialised* at 85/15, so the much larger gridded inventory does not drown out the country tasks. This 85/15 is a nominal base, not a hard constraint: it fixes the per-task sampling probabilities at start-up, but the active-learning resampler (§A.19) then multiplies each task's base probability by a bounded uncertainty factor in $[0.5, 2.0]$ every two epochs, *independently of modality*. The *realised* gridded/country share therefore drifts away from 85/15 over training, toward whichever modality carries the larger predictive uncertainty; the drift is bounded because every multiplier is clamped to $[0.5, 2.0]$ and re-anchored to the fixed base each cycle (it does not compound). Within the gridded bucket, per-task selection probabilities are shaped by multiplicative boosts and then renormalised so the modality split is preserved. Writing π^0_κ for a task's base (importance-prior) probability and b_κ for its boost factor, the gridded sampling probability is

$$\pi_\kappa = \frac{\mathrm{clip}\big(b_\kappa\, \pi^0_\kappa,\ r_{\min}/N_{\mathrm{g}},\ r_{\max}/N_{\mathrm{g}}\big)}{\sum_{\kappa' \in \mathrm{grid}} \mathrm{clip}(\cdot)}, \qquad r_{\min} = 0.04,\ r_{\max} = 25, \tag{38}$$

where N_{g} is the number of gridded tasks, the clip bounds are relative to the *uniform* probability $1/N_{\mathrm{g}}$, not to the task's own prior, with a per-domain floor guaranteeing each domain at least a fraction 0.01 of the total sampling weight (across both modalities). The headline boosts (`task_boosts`) up-weight the urban/population and basic-climate fields the model is expected to find hardest and most policy-relevant: population density, nighttime lights, built surface, impervious surface, urban vegetation class, and human footprint at 10×; mean temperature and total precipitation at 5×; their standard deviations at 2×, on top of a base table that further up-weights anthropogenic-activity proxies and down-weights smooth distance/accessibility fields.

Algorithm 5 Mixed mini-epoch step sequence

Require: family counts $\{n_{\mathrm{rec}}, n_{\mathrm{tr}}, n_{\mathrm{geo}}, n_{\mathrm{cl}}\}$, window w=50

$S \leftarrow$ multiset with n_{fam} tokens of each family (geo tokens cycle the 3 heads)
shuffle S globally
partition the shuffled S into consecutive windows of size w
for each window **do** shuffle its tokens uniformly
return concatenation of windows ▷ family counts preserved exactly

Design rationale. The global shuffle followed by mini-epoch reshuffling avoids the grouped ordering in which the optimiser would see one objective for long contiguous stretches. It gives a statistically mixed stream while the exact family counts are preserved. The 85/15 modality split and the task boosts together counteract the raw inventory imbalance (the gridded set is dominated by climate and static fields, Figure 13) so that the human-system and climate fields central to the model's purpose receive adequate gradient.

A.18 Optimisation

Optimiser and precision. AdamW (Loshchilov and Hutter, 2019) with peak learning rate $\eta_{\max} = 1.2 \times 10^{-4}$, weight decay 0.01, $\beta = (0.9, 0.999)$; gradients clipped to $\|\nabla\|_2 \leq 1.0$; `bf16-mixed` precision; effective batch 8192. Orthogonal initialisation (Saxe et al., 2014) is applied to all non-positional-encoding parameters; the location and time encoders are excluded so they keep their WIRE/hash/Fourier-specific initialisations (§A.4.1, §A.4.4).

Learning-rate schedule. A single training scale (the spatial-scale curriculum is off; §A.19) of $E = 128$ epochs: exponential warmup for $w = 5$ epochs from $\eta_0 = 10^{-6}$ to $\eta_{\max}$, then cosine decay to $\eta_{\min} = 10^{-6}$ (Loshchilov and Hutter, 2017),

$$\eta(e) = \begin{cases} \eta_0 \, (\eta_{\max}/\eta_0)^{e/w}, & 0 \leq e < w, \\ \eta_{\min} + (\eta_{\max} - \eta_{\min}) \cdot \frac{1}{2}\left(1 + \cos \pi \frac{e-w}{E-w-1}\right), & w \leq e \leq E-1. \end{cases} \tag{39}$$

The scheduler advances then applies (no one-epoch lag), so the final trained epoch lands exactly on $\eta_{\min}$. The implementation is a per-scale scheduler that resets a full warmup→cosine cycle at each curriculum scale; with the curriculum disabled it runs as a single cosine over all 128 epochs (Fig. 16).

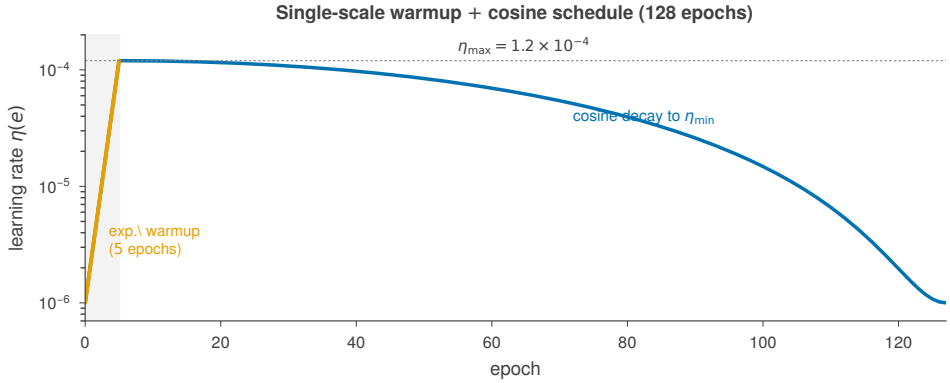

Figure 16: The learning-rate schedule (Eq. 39): a 5-epoch exponential warmup from $\eta_0 = 10^{-6}$ to the peak $\eta_{\max} = 1.2 \times 10^{-4}$ (shaded), followed by a single cosine decay back to $\eta_{\min} = 10^{-6}$ over the remaining 123 epochs. Advance-then-apply stepping makes the final epoch land exactly on $\eta_{\min}$.

Weight EMA. An exponential moving average of the weights with decay 0.9999, ramped linearly over the first 2000 steps, is maintained on-GPU. Evaluation, mid-training plots, and checkpoints are produced for both the raw and the EMA weights.

Design rationale. Advance-then-apply stepping matters because it guarantees the finest training reaches $\eta_{\min}$ exactly at the last epoch. The EMA provides a lower-variance weight estimate for evaluation and is checkpointed separately as the sibling of each raw checkpoint, so raw and EMA weights from the same selected training state can be compared post-hoc without retraining.

A.19 Active-learning resampling

Every $K_{\mathrm{AL}} = 2$ epochs the per-task evidential uncertainty is recomputed on held-out points and used to reweight the reconstruction task sampler, so that tasks the model is most uncertain about receive more gradient in the next interval (Lewis and Gale, 1994; Settles, 2009). Coverage is per-task by construction: the callback iterates over the live sampler's task set and draws $n = 64$ validation points for *every* task, so none is left at its default weight merely because it did not appear in a random validation slice. For each task t it forms the mean total uncertainty $u_t = \langle \sigma^2_{\mathrm{alea}} + \sigma^2_{\mathrm{epi}} \rangle$ (Eq. 23), dropping imputed off-support cells, smooths it with an EMA $\bar{u}_t \leftarrow d\,\bar{u}_t + (1{-}d)u_t$ ($d = 0.5$) to avoid thrashing, and converts the smoothed signal into a multiplicative adjustment of the task's base probability:

$$m_t = \mathrm{clip}\Big(\big(\bar{u}_t / \langle \bar{u} \rangle\big)^s,\, m_{\min}, m_{\max} \Big), \qquad p_t \propto p_t^0\, m_t, \qquad s = 2.0,\ [m_{\min}, m_{\max}] = [0.5, 2.0]. \tag{40}$$

The exponent s tunes aggressiveness (larger s sharpens the reweighting) and the clamps bound how far any task may drift from its prior. The callback logs diagnostic statistics each pass (mean/percentile uncertainty, the sampling-distribution entropy and its KL to the base, the effective sample size, and the clip rates) for monitoring.

Algorithm 6 Evidential active resampling (every K_{AL} epochs)

Require: base probabilities p^0, decay d, strength s, clamps $m_{\min}, m_{\max}$, budget n

for each task t in the live sampler **do**
  draw n held-out points; evaluate $(\mu, \nu, \alpha, \beta)$; drop imputed cells
  $u_t \leftarrow \big\langle \beta/(\alpha{-}1) + \beta/(\nu(\alpha{-}1)) \big\rangle$ ▷ mean total uncertainty
  $\bar{u}_t \leftarrow d\,\bar{u}_t + (1{-}d)\,u_t$ ▷ EMA; init $\bar{u}_t = u_t$
$\langle \bar{u} \rangle \leftarrow \mathrm{mean}_t\, \bar{u}_t$
for each task t **do**
  $m_t \leftarrow \mathrm{clip}\big((\bar{u}_t/\langle \bar{u} \rangle)^s, m_{\min}, m_{\max}\big)$; $w_t \leftarrow p_t^0\, m_t$
push $p_t \leftarrow w_t / \sum_{t'} w_{t'}$ into the sampler; append audit log

Design rationale. The $2 \times 2 \times 2$ ablation study crosses the spatial-scale curriculum, active learning, and the evidence coefficient and finds active resampling to be the single most valuable training-time lever: it lowers reconstruction error by $\widehat{\Delta}_{\mathrm{RMSE}} = -0.052$ (a large effect at $d = -0.89$) and roughly halves the calibration NLL, and every cell on the accuracy–calibration Pareto front is an active-learning cell. The same study finds the spatial-scale curriculum significantly *hurts* reconstruction ($\widehat{\Delta}_{\mathrm{RMSE}} = +0.027$): a plausible explanation is that the multi-resolution hash branch of the location encoder (§A.4.1) is designed to ingest high-resolution detail from the outset, so spending early budget at coarse resolutions wastes capacity the model would otherwise use for fine structure. The curriculum is therefore switched off, and the winning cell is the final recipe: SC-OFF · AL-ON · λ_{ev}=0.3, which is rank-one across all three seeds.

Part G. Evaluation

Validation runs every 3 *epochs over up to* 200 *batches and reports three families of metrics: reconstruction accuracy(§A.20), uncertainty calibration and sharpness (§A.21), and representation diagnostics (§A.22). Metrics are computed in standardised space.*

A.20 Reconstruction accuracy

For each task t with pooled held-out predictions $\{(\hat{y}_i, y_i)\}_{i \in \mathcal{V}_t}$ (point estimate $\hat{y} = \mu$), we compute

$$\mathrm{RMSE}_t = \sqrt{\tfrac{1}{|\mathcal{V}_t|} \textstyle\sum_i (\hat{y}_i - y_i)^2}, \qquad \mathrm{MAE}_t = \tfrac{1}{|\mathcal{V}_t|} \textstyle\sum_i |\hat{y}_i - y_i|, \tag{41}$$

$$R^2_t = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}, \qquad \rho_t = \frac{\sum_i (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}, \tag{42}$$

(ρ_t is Pearson correlation). For per-field comparisons we also report PSNR, $\mathrm{PSNR}_t = 10 \log_{10}\big(\mathrm{range}(y_t)^2 / \mathrm{MSE}_t\big)$, computed in standardised space (the range^2/MSE ratio is invariant under the affine de-standardisation, so the value coincides with its physical-unit counterpart); it is the standard image-reconstruction quality measure and makes heterogeneous fields comparable on a decibel scale. Aggregates are reported overall (the selection metric is the task-mean RMSE) and split by modality ($\overline{\mathrm{RMSE}}_{\mathrm{grid}}$, $\overline{\mathrm{RMSE}}_{\mathrm{country}}$, and the corresponding mean correlations), since the two geometries have different difficulty profiles.

A.21 Calibration and sharpness

Calibration is assessed through the probability integral transform (PIT) of the held-out targets under a Gaussian moment-match of the predictive, $\mathcal{N}\big(\mu,\ \sigma^2{=}\sigma^2_{\mathrm{alea}} + \sigma^2_{\mathrm{epi}}\big)$, where $\sigma^2 = \beta/(\alpha{-}1) + \beta/(\nu(\alpha{-}1))$. With standardised residual $z_i = (y_i - \mu_i)/\sigma_i$ and $\mathrm{PIT}_i = \Phi(z_i)$,

a well-calibrated predictive yields PIT values uniform on $[0, 1]$. We report:

$$\text{NLL} = \left\langle \tfrac{1}{2}\log(2\pi\sigma_i^2) + \tfrac{(y_i-\mu_i)^2}{2\sigma_i^2} \right\rangle, \qquad \text{sharpness} = \langle \sigma_i \rangle, \tag{43}$$

$$\text{ECE (Kuleshov et al., 2018)} = \tfrac{1}{B}\sum_{b=1}^{B} \left|\widehat{F}(\text{PIT} \le e_b) - \bar{e}_b\right|, \qquad \text{KS-PIT} = \sup_u \left|\widehat{F}_{\text{PIT}}(u) - u\right|, \tag{44}$$

$$\text{Cov}_\ell = \left\langle \mathbf{1}[\,|y_i - \mu_i| \le z_\ell\, \sigma_i] \right\rangle, \qquad \ell \in \{0.50, 0.80, 0.95\}, \tag{45}$$

where $B = 10$ equal-width PIT bins have right edges e_b and centres $\bar{e}_b$, $\widehat{F}$ is the empirical CDF, $z_\ell = \Phi^{-1}(\frac{1+\ell}{2})$ is the two-sided normal quantile, and $\mathbf{1}[\cdot]$ the indicator. NLL summarises overall calibration; sharpness is the mean predictive width (lower is sharper, to be read jointly with calibration); ECE, in the binned form of Guo et al. (2017) applied to the probability integral transform rather than to classifier confidence (Kuleshov et al., 2018), and KS-PIT measure deviation of the PIT from uniformity (the former binned, the latter a supremum statistic); and coverage compares the empirical frequency of targets inside the nominal ℓ interval against ℓ.

Design rationale. The marginal predictive is a Student-t (Eq. 22), but its aleatoric/epistemic split is not exactly separately identifiable from the likelihood (Meinert et al., 2023); calibration therefore uses their summed predictive variance for PIT/ECE/coverage, rather than attempting to validate the two components separately. Calibration is reported as a co-equal diagnostic and deliberately kept out of model selection, which uses RMSE only: this avoids selecting a model that is well-calibrated but inaccurate. The width cap (26) supplies the scale prior that makes the moment-matched Gaussian a faithful summary on z-scored targets.

A.22 Representation diagnostics

Alignment retrieval. For each external bank, the projected source and target embeddings give a cosine-similarity matrix whose diagonal is the positive pair. We measure top-1 accuracy, recall@5/@10, mean reciprocal rank, median rank, the median great-circle error of the top-1 retrieval (km), the fraction retrieved within $\{25, 100, 500\}$ km, and the mean positive cosine, a battery that captures both how often the correct location is retrieved and how geographically close the mistakes are.

Intrinsic dimension. The latent geometry of each encoder is summarised by the intrinsic dimension of its embeddings. Two batteries are used. Validation during training logs TwoNN (Facco et al., 2017), which fits the distribution of the ratio of second- to first-nearest-neighbour distances, together with a local-PCA estimate. The post-hoc latent-geometry analysis uses the Levina–Bickel local maximum-likelihood estimator (Levina and Bickel, 2004), whose pointwise estimate from k neighbours is $\hat{m}_i = \left[\frac{1}{k-1}\sum_{j=1}^{k-1} \log \frac{r_k(i)}{r_j(i)}\right]^{-1}$ ($r_j(i)$ the distance to the j-th neighbour), together with method-of-moments, TLE (Amsaleg et al., 2019), and FisherS (Albergante et al., 2019) estimates. Reporting global and pooled-local estimates across estimators guards against the failure modes of any single method, and the intrinsic dimension is used qualitatively to compare how the continuous, lookup, and hybrid encoders organise their latent space.

A.23 Adaptation to unseen tasks

The training objective (Part C) fits the $T = 1{,}024$ reconstruction tasks of the taxonomy (§A.3); a foundation model earns its name only if a task *absent* from that set can be attached cheaply afterwards. This section specifies how a *frozen* TerraNova is adapted to an unseen regression target: the parameter surface on which adaptation acts, the family of adaptation operators we compare around the native MiSS operator, how a new task embedding is initialised, and the held-out protocol under which adaptation skill, calibration, and cost are scored. The protocol is specified here; the quantitative study is reported in the main text.

A.23.1 The adaptation surface

The task encoder is a pure lookup $E_\tau \in \mathbb{R}^{N_\tau \times 256}$ whose final rows are *reserved for dynamic registration* (§A.4.6). Adapting to an unseen task $\tau^\star$ first instantiates one such row, $\mathbf{e}_{\tau^\star} = \mathrm{N}_{256}\big(E_\tau[\tau^\star]\big)$. The native operator then learns that row jointly with Matrix Shard Sharing (MiSS) (Kang and Yin, 2026) residuals on the linear maps $\ell \in \mathcal{M}$ of the task·ST fusion. For the row-vector convention used here,

$$\widetilde{f}_\ell(\mathbf{h}) = \mathbf{h}W_\ell^\top + (\mathbf{h}A_\ell)B_{:,:d_\ell^{\mathrm{out}}}, \quad A_\ell \in \mathbb{R}^{d_\ell^{\mathrm{in}} \times r}, \qquad B \in \mathbb{R}^{r \times d_{\max}^{\mathrm{out}}}. \tag{46}$$

Each map has its own trainable input factor A_ℓ, whereas one trainable output factor B is shared across maps and sliced to d_ℓ^{out}; the frozen base map W_ℓ is never modified. This differs from the published operator in where the sharing happens: MiSS ties one block within a layer across shards of its output dimension, whereas we tie the output factor across layers. Adaptation uses $r = 4$ over all eligible maps in the 16-block task·ST fusion. This gives

$$256 + 4\underbrace{\sum_{\ell \in \mathcal{M}} d_\ell^{\mathrm{in}}}_{99{,}840} + 4\underbrace{d_{\max}^{\mathrm{out}}}_{4{,}096} = 416{,}000 \tag{47}$$

trainable scalars, including the fresh task row. The one exception is the gridded label-budget sweeps, which fix $r = 8$ and therefore train 831,744 scalars; every other adaptation result uses the rank-4 operator. The location and time encoders, spatiotemporal fusion, hypernetwork, evidential head, base task·ST weights, and all existing task rows remain frozen.

The two learned components play complementary roles. Moving $\mathbf{e}_{\tau^\star}$ locates the variable in the pretrained task geometry; the shared MiSS shard and layer-specific factors allow a low-rank correction to how that new task interacts with the spatiotemporal state before decoder generation. The resulting adapter is stored and activated with $\tau^\star$. With it inactive, all pretrained-task computations remain bit-identical to the base model.

Design rationale. This is the operational form of the premise stated in §A.2 and §A.7.5: the pretrained embedding geometry is the transfer substrate, while MiSS supplies limited interaction capacity when a new target cannot be represented by moving one task vector alone. Sharing B across the fusion maps makes that correction substantially smaller than an independent low-rank adapter at every layer. Because neither base weights nor old task rows change, adapters can be registered per task without cumulative fine-tuning drift.

A.23.2 A ladder of adaptation operators

Given a labelled adaptation set $\mathcal{D}_{\tau^\star} = \{(u_i, t_i, y_i)\}_{i=1}^{n}$ at *label budget* n, every operator minimises the same standardised mean-squared error on the predicted mean,

$$\mathcal{L}_{\text{adapt}} = \big\langle \big(\mu_\theta(\tau^\star, u_i, t_i) - \tilde{y}_i\big)^2 \big\rangle, \qquad \tilde{y}_i = \text{per-task } z\text{-score of } y_i \ (\S A.14), \tag{48}$$

with Adam under a per-operator step budget (the single-vector and PEFT operators run 2× the base step count; the heavy operators are capped at 150 steps). The cost measurements reported separately do not use this schedule: there every gradient operator runs the same 800 steps, so that wall-clock compares operators at equal work rather than at equal budget. In both cases the operators differ only in *which* parameters are unfrozen (Table 8). Let $\mathcal{M}$ index the linear maps of the task·ST fusion (§A.6.2), the PEFT target, with in/out widths $d_\ell^{\text{in}}, d_\ell^{\text{out}}$. MiSS at $r = 4$ is the native operator used for the main adaptation results; the remaining rows form the controlled capacity–cost ladder.

Table 8: Adaptation operators, ordered by trainable-parameter count. All fit μ by Eq. (48); r is the PEFT rank. The linear probe touches no model weights (a ridge read-out of the frozen spatiotemporal embedding); every gradient operator masks the embedding gradient to the single new row of E_τ, so the 1,024 trained task vectors are preserved exactly.

operator	trained quantity	trainable parameters
linear probe	ridge weights on frozen $\mathbf{z}_{\text{st}}$ (no model grad.)	256+1
task embedding	$\mathbf{e}_{\tau^\star}$ only (lower-capacity baseline)	256
VeRA (r)	$\mathbf{e}_{\tau^\star}$ + per-layer scales $(\mathbf{d}_\ell, \mathbf{b}_\ell)$, shared frozen A, B	$256 + \sum_{\ell\in\mathcal{M}}(r + d_\ell^{\text{out}})$
MiSS ($r = 4$)	$\mathbf{e}_{\tau^\star}$ + $\{A_\ell\}$, single shared B	416,000
LoRA (r)	$\mathbf{e}_{\tau^\star}$ + adapters $A_\ell B_\ell$	$256 + \sum_{\ell\in\mathcal{M}} r\,(d_\ell^{\text{in}} + d_\ell^{\text{out}})$
finetune-last	$\mathbf{e}_{\tau^\star}$ + last two task·ST blocks + output proj.	block-dependent
(decoder / full)	*hypernetwork+head / all θ*	*upper bounds*

The probe (Alain and Bengio, 2016) and the task-embedding baseline touch no backbone weights at all; the PEFT operators (LoRA (Hu et al., 2022), MiSS (Kang and Yin, 2026), VeRA (Kopiczko et al., 2024)) add low-rank residuals to the task·ST fusion; finetune-last unfreezes its top blocks; decoder/full fine-tuning are included only as upper bounds. Each weight-space operator carries its own learning rate: a 256-vector tolerates a large step whereas weight-matrix adapters diverge unless the step is much smaller, and sharing one learning rate across operators is the single largest source of an unfair parameter-efficiency comparison.

A.23.3 Initialising a new task embedding

The reserved row is initialised on the same sphere of radius $\sqrt{256}$ (Eq. N) as the trained tasks, then optimised. We consider three initialisers: the *arithmetic mean* of the trained rows (used as-is); the *spherical mean* (the ℓ_2-renormalised mean of the unit directions, rescaled to the mean trained-row norm); and *random-sphere* (a seeded random unit direction, rescaled likewise). A grid over (initialiser, learning rate) selects random-sphere as the default: it is best-or-tied everywhere and, unlike the two mean initialisers, never collapses a fit into a near-constant field.

Design rationale. A centroid warm start can sit in a flat, low-loss basin for the task-row component, where the gradient on $\mathbf{e}_{\tau^\star}$ is weak and the optimisation can stall at a degenerate (near-constant) decoder; a random point on the sphere breaks that symmetry while the fixed-radius convention keeps the new row at the same scale as trained tasks in every attention logit it feeds.

A.23.4 Held-out evaluation protocol

Splits. Adaptation is scored under a split that stresses generalisation of the new task *across space*, complementary to the held-out-*time* validation used in training (§A.16). Country tasks use a *group* split: a fixed fraction of whole countries is held out, so the task is fit on some nations and tested on entirely unseen ones. Gridded tasks use a *spatial-block* split: contiguous longitude/latitude blocks are assigned wholesale to train or test, so spatial autocorrelation cannot leak a memorised neighbour across the boundary. The label budget n subsamples the train side only; the held-out side $\mathcal{V}$ is fixed across budgets and operators.

Skill. Point skill on the standardised target uses the Pearson correlation ρ and the RMSE/MAE of §A.20. For time-varying tasks we additionally report the *anomaly correlation*, the Pearson correlation after removing each location's temporal mean,

$$\rho_{\text{anom}} = \operatorname{corr}\big(\hat{y}_i - \bar{\hat{y}}_{\ell(i)},\ \ y_i - \bar{y}_{\ell(i)}\big), \qquad \ell(i) \text{ the location of sample } i, \tag{49}$$

over locations sampled at least twice. It isolates inter-annual skill and is ≈ 0 for a predictor that only fits the spatial climatology, so it is the decisive metric on country×year tasks, where there is no spatial structure to exploit.

Calibration. The operators of Table 8 fit only μ, so the predictive width (ν, α, β) at the new task is whatever the *frozen* hypernetwork emits for $\mathbf{e}_{\tau^\star}$. Because the mean is re-fit while the width is not, the inherited width can be globally mis-scaled on the new task; before scoring we therefore apply one recalibration constant per task, fitted on data disjoint from the side that is scored, and log the uncalibrated (raw frozen-head) coverage alongside. Two forms are used: σ-scaling, $\sigma \leftarrow s\,\sigma$ with $s^2 = \langle (y-\mu)^2/\sigma^2 \rangle$ estimated on the adaptation *train* split, and a split-conformal multiplier fitted directly at the nominal level, which is the interval reported and drawn in the main text. Under the same Gaussian moment-match as §A.21, $\sigma^2 = \sigma^2_{\text{alea}} + \sigma^2_{\text{epi}}$ (Eq. 23), we report the 90% interval coverage (Eq. 45) and the closed-form Gaussian CRPS (Gneiting and Raftery, 2007),

$$\operatorname{CRPS}\big(\mathcal{N}(\mu,\sigma^2), y\big) = \sigma\Big[z\big(2\Phi(z)-1\big) + 2\varphi(z) - \tfrac{1}{\sqrt{\pi}}\Big], \qquad z = \tfrac{y-\mu}{\sigma}, \tag{50}$$

with φ the standard-normal density. Because the *shape* of the width is inherited rather than re-fit: recalibration adjusts one global scalar: these diagnostics measure whether the model's *pretrained* uncertainty *transfers* to an unseen task up to a single scale factor.

Cost. Each adaptation is logged with its trainable-parameter count (Table 8) and its wall-clock fit time on a single GPU, so skill is always reported jointly with the label and compute budget that produced it.

Design rationale. The group and spatial-block splits make the test a genuine *new-region* generalisation rather than an interpolation among autocorrelated neighbours, which a random per-sample split would silently allow. Reading calibration off the frozen head, rather than re-fitting a variance model (only a single global scale is recalibrated, on the train split), is deliberate: the question a foundation model must answer is whether its *own* uncertainty remains trustworthy on a task it was never trained on.

Algorithm 7 MiSS adaptation of a frozen TerraNova

Require: frozen θ, task encoder E_τ, labelled set $\mathcal{D}_{\tau^\star}$, budget n, steps S, rank $r = 4$, lrs (η_e, η_m), init $\in \{\text{mean}, \text{sph}, \text{rand}\}$

register a new id $\tau^\star$ in a reserved row of E_τ; initialise $E_\tau[\tau^\star]$ on the sphere of radius $\sqrt{256}$ (init)
freeze all base parameters and existing task rows
for every eligible fusion map W_ℓ, attach A_ℓ; create shared B as in Eq. (46); initialise $A_\ell \sim \mathcal{N}(0, 0.01^2)$ and $B = 0$
mark only $E_\tau[\tau^\star]$, $\{A_\ell\}$, and B trainable
split $\mathcal{D}_{\tau^\star}$ (group / spatial-block); subsample train to n; z-score by train statistics (§A.14)
for $s = 1, \ldots, S$ **do**
    sample a minibatch; $\hat{\mu} \leftarrow \mu_\theta(\tau^\star, \cdot)$ via Eqs. (16)–(19)
    $\mathcal{L} \leftarrow \langle(\hat{\mu} - \tilde{y})^2\rangle$; backprop; **mask** the embedding gradient to row $\tau^\star$; Adam step with (η_e, η_m)
return held-out ρ, ρ_{anom}, RMSE, $\text{Cov}_{0.9}$, CRPS, trainable-parameter count, walltime

Part H. Reference

This Part collects the consolidated hyperparameters (§A.24) and the reproducibility and compute setup; the mapping from each component to the study that selected it is provided in the ablation appendix. All values are the resolved `evidential_full_512_drop015_tr2048` *settings.*

A.24 Hyperparameters

Table 9 consolidates every resolved hyperparameter: encoders, fusion, decoder, objective, optimisation, and data; values differing from the ablation common width or the main-text draft are flagged in the ablation appendix.

Table 9: Production hyperparameters (consolidated).

group	parameter	value
location enc.	WIRE U-Net widths	$[512, 1024, 2048, 4096, 2048, 1024, 512]$
	WIRE $(\omega_0, s_0, \omega_0^{\text{first}})$	$(10, 10, 30)$
	hash $(L, F, \log_2\lvert T\rvert, N_{\min}, N_{\max})$	$(16, 4, 16, 16, 1024)$
	gate init / clamp	1.386 / $[0.15, 0.85]$
country enc.	table dim / MLP (depth,width)	64 / $(4, 256)$, SiLU, LN
time enc.	arch / dim / width / depth / K	residual / 32 / 256 / 4 / 32
	year normalisation $[t_{\min}, t_{\max}]$	$[1900, 2035]$
transport	mode / δ-enc / depth / width	residual-add / Fourier+RBF / 4 / 256
	$\Delta_{\max}$ / δ-Fourier / δ-RBF	4 / 16 freq. (32-D) / 16 (ς_0=0.15)
task enc.	arch / dim	embedding-only / 256
ST fusion	type / width / heads / depth / out	Tr-A / 1024 / 16 / 8 / 256
	residual gate init (α)	-1.1 ($g \approx 0.275$)
task·ST fusion	type / width / depth / out	Tr-A / 1024 / 16 / 512
decoder	hypernet hidden / depth	512 / 1
	generated hidden H / weight-norm	32 / row-ℓ_2
NIG head	mode / $\sigma_{\max}$	single / 3
objective	λ_{ev} / κ / $W_{\max}$ / p	0.3 / 0.1 / 3 / 1
	category weights (rec,tr,geo,cl)	$(1, 1, 1, 0.25)$
	contrastive temperature τ_{T} / geo negs / excl.	0.07 / 32 / 50 km
training	epochs / batch	128 / 8192
	optimiser / lr / wd / clip	AdamW / 1.2×10^{-4} / 0.01 / 1.0
	warmup / $\eta_{\min}$ / precision	5 ep / 10^{-6} / bf16-mixed
	embedding / general dropout	0.15 / 0.15
	EMA decay / warmup	0.9999 / 2000 steps
	per-epoch (rec,tr,geo,cl)	$(32768, 512, 256, 256)$
	AL: K / n / s / d / clamp	2 / 64 / 2.0 / 0.5 / $[0.5, 2.0]$
data	gridded / country tasks	512 / 512
	split / holdout yrs / val frac / seed	temporal-year / 3 / 0.02 / 42
	country fraction / OCN fraction	0.15 / 0.10

Appendix B. Full Ablation Programme

B.1 Ablation Studies

B.1.1 Overview

As we have described, TerraNova is composed of a set of modular representation, fusion, transport, and objective components. A fully combinatorial ablation over all such design choices, compounded with optimization hyperparameter choices, would be both computationally prohibitive and difficult to interpret, since multiple interacting changes could improve or degrade performance simultaneously. Inspired by the engineering process of modern neural network architectures Liu et al. (2022), we adopt a sequential ablation strategy in which a single component family is varied at a time while all remaining components are either fixed or disabled. After each study, the strongest configuration is promoted and used as the default choice in subsequent stages.

Unless otherwise stated, all ablation studies follow the same two-stage protocol. When computationally feasible, we first perform a screening phase in which each candidate configuration is trained with a single seed over a shared learning-rate sweep, to find suitable optimization setups for each configuration. We then perform a confirmation phase in which the best learning rate for each configuration is retrained with three seeds, a longer optimization budget, and cosine learning-rate decay. This procedure allows us to explore broad configuration spaces efficiently while still basing final decisions on more stable multi-seed comparisons.

Each component is evaluated on a benchmark matched to its intended role in the full model. Spatial encoders are assessed on multitask geospatial field reconstruction, country encoders on country-level panel reconstruction, temporal encoders on global temporal trend encoding, and compositional modules such as transport or fusion on downstream benchmarks that require interaction between multiple pretrained components. This design avoids evaluating every module on the same proxy task, and instead measures each one under the setting in which it is expected to contribute most directly. For every experiment, unless stated otherwise, we use the AdamW Loshchilov and Hutter (2019) optimizer. The remainder of this section follows the same structure for each component family: we first motivate the corresponding design question, then describe the configurations tested, report the main empirical results, and finally state the configuration that is carried forward into the next stage of the ablation setup.

B.1.2 Location encoder

Motivation

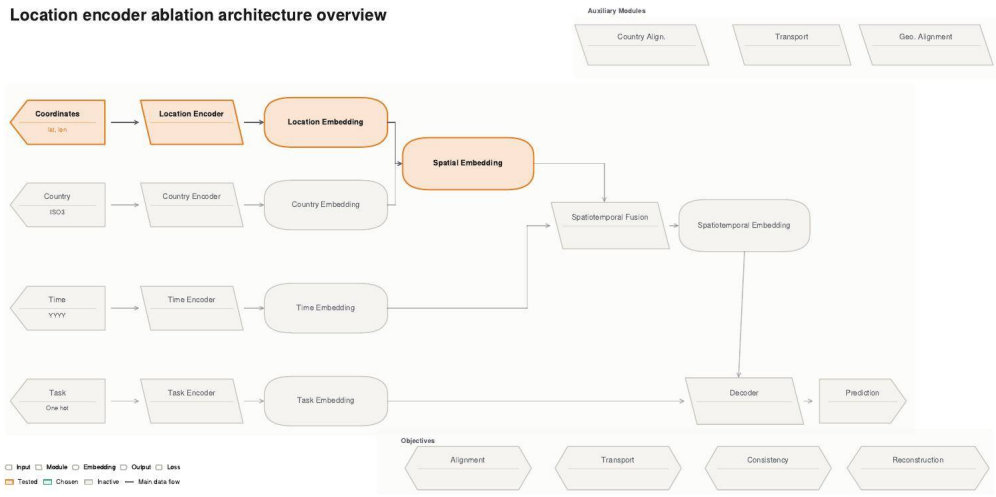

Figure 17: **Ablation scope.** The location input, location encoder, and location embedding are varied across a 44-variant search space (12 standalones + 32 cross-family hybrids), of which 38 variants advance to the confirmation phase. Every other TerraNova block is held fixed and participates only as the shared task head. The benchmark is WorldTensor restricted to 51 gridded fields available in the 2015 reference year.

TerraNova's spatial pathway must reconstruct fields whose spatial irregularity spans several orders of magnitude, from zonally smooth climatological means (sea level pressure or mean temperature) to sparse, high-frequency geographic rasters (population density, wildfire area). The literature in computer graphics and implicit neural representations has presented several approaches for encoding 3-dimensional signals, optimizing for precision, speed, smoothness, or generalization. Overall, we categorize existing approaches into two families of functions that can be used for encoding location embeddings from raw coordinates (ϕ, θ):

1. **Continuous implicit neural representations (INRs)** like Sinusoidal Representation Networks (SIREN) (Sitzmann et al., 2020), Wavelet Implicit Neural Representations (WIRE) (Saragadam et al., 2023), Fourier Feature Networks or Neural Radiance Field (NeRF) positional encodings (Mildenhall et al., 2022; Tancik et al., 2020), or spherical harmonics (SH) (Rußwurm et al., 2024; Rao et al., 2026) applied to geographical data introduce a local smoothness prior that excels on smoother fields but is bandwidth-bounded: they struggle to express the high-frequency, locally discontinuous detail demanded by sparse rasters, which are common in gridded socioeconomic datasets. This has been the most common approach in the geospatial embedding literature.
2. **Lookup-table (LUT) encodings** : multi-resolution hash grids (Müller et al., 2022), Gaussian splatting (Kerbl et al., 2023), or tri-planes (Shue et al., 2023) introduce a

local, high-capacity prior that excels on discontinuous fields but tends to leak noise on smooth targets because their representation is globally unregularised.

This ablation isolates the *location encoder* block of the TerraNova architecture (Fig. 17) and asks whether a single coordinate backbone can dominate both regimes, or whether a principled combination of different approaches is required. We consider 12 standalone backbones, then form cross-family gated hybrids between the top-4 continuous and top-4 LUT standalones, yielding 32 hybrid variants and 44 active configurations in total on 51 geospatial fields from the WorldTensor benchmark. Our goal in this experiment is to design an architecture that can handle both the continuous and smooth patterns common in atmospheric data, and the high-frequency and sparse characteristics of socioeconomic, infrastructure, and settlement data. The design of this location encoder can play a very significant role in the overall performance of the full model, and we therefore use a significant budget of the total ablation study into this key component.

B.1.3 Encoder zoo

Let $c = (\lambda, \phi) \in [-180, 180) \times [-90, 90]$ denote a (longitude, latitude) pair in degrees, and let $f_\theta : \mathbb{R}^2 \to \mathbb{R}^{128}$ be the function we wish to fit. All backbones share the same output dimension $d = 128$. Below, we list our neural architecture designs, and their sizes in terms of layers or neurons per layer. Note that, unless stated otherwise, these were chosen so the amount of trainable parameters is comparable across model families.

Simple SIREN. A fully-connected network with sine activations $\sigma_\omega(u) = \sin(\omega_0 u)$, $\omega_0 = 30$, initialised following Sitzmann et al. (2020):

$$f_\theta(c) = W_L\, \sigma_\omega\Big(W_{L-1}\, \sigma_\omega(\cdots \sigma_\omega(W_1\, g(c) + b_1) \cdots) + b_{L-1}\Big) + b_L, \tag{51}$$

with hidden widths $[704, 704, 1408, 2816, 5632, 2816, 1408, 704, 704]$ and WrapNorm input $g(c)$ (Sec. B.1.4).

Residual SIREN. A SIREN of widths $[512, 512, 1024, 2048, 4096, 4096, 2048, 1024, 512, 512]$ with pre-activation residuals He et al. (2016a) between layers of matching width (scale-matched skip projection where the widths differ).

SIREN U-Net. U-shaped SIREN of widths $[512, 1024, 2048, 4096, 2048, 1024, 512]$ with symmetric skip connections between encoder- and decoder-side layers of equal width. This kind of architecture has shown success in both generative models and location encoding of high-frequency spherical and hemispherical data (Rußwurm et al., 2024).

WIRE U-Net. SIREN U-Net with the sine activation replaced by the complex Gabor wavelet of Saragadam et al. (2023):

$$\sigma_{\text{WIRE}}(u) \;=\; \exp(j\, \omega_0 u)\, \exp(-|s_0 u|^2), \tag{52}$$

with $\omega_0 = 10$, $s_0 = 10$ in the recurrent layers and $\omega_{0,\text{init}} = 30$ in the first layer. These models have shown to encode high frequency patterns better than sinusoidal INRs.

SH + SIREN U-Net. Following Rußwurm et al. (2024), a SIREN U-Net preceded by a real spherical-harmonic basis with `legendre_polys`=8, yielding a 64 dimensional SH feature that is passed through the U-Net. A similar approach was presented in Klemmer et al. (2025).

NeRF PE + MLP. Deterministic multi-band Positional Encoding following the seminal NeRF definition (Mildenhall et al., 2022): $\gamma_L(x) = [\sin(2^k \pi x), \cos(2^k \pi x)]_{k=0}^{L-1}$ with $L = 8$, applied elementwise to Cartesian-3D coordinates, followed by a residual MLP. A similar approach was presented in Vivanco Cepeda et al. (2023).

Spherical Harmonics + MLP. The same 64-dimensional real SH basis used above, followed by a residual MLP.

Hash Grid (S/M/L). Multi-resolution hash encoding (Müller et al., 2022): L levels of a hash table with T rows and feature dim F, with per-level resolutions $N_\ell = N_{\min} \cdot b^\ell$ and $b = (N_{\max}/N_{\min})^{1/(L-1)}$. We keep $F = 4$, $N_{\min} = 16$, and sweep the capacity axis with

$$\begin{aligned} \texttt{S} &: L = 8, T{=}2^{14}, N_{\max} = 256; \\ \texttt{M} &: L = 12, T{=}2^{15}, N_{\max} = 512; \\ \texttt{L} &: L = 16, T{=}2^{16}, N_{\max} = 1024. \end{aligned}$$

Coordinates are passed as longitude/latitude in degrees and normalised internally to a wrapped $[0, 1) \times [0, 1)$ grid (Sec. B.1.4). This sort of model acts as a multi-resolution LUT and has shown significant success in computer graphics for quickly encoding high-frequency data.

RBF + MLP. $K = 256$ learnable Gaussian centres $\mu_k \in \mathbb{S}^2$ with learned log-bandwidths $\log \sigma_k$; features $\phi_k(c) = \exp(-\|x_{\text{cart}}(c) - \mu_k\|^2 / 2\sigma_k^2)$ are concatenated and passed to a residual MLP.

Gaussian Splatting. $K = 4096$ learnable isotropic Gaussian splats Kerbl et al. (2023) on the sphere with centres $\mu_k \in \mathbb{S}^2$, log-widths $\log \sigma_k$, opacity logits α_k, and per-splat features $f_k \in \mathbb{R}^{256}$. Given Cartesian input $x(c)$, the encoder computes

$$g_k(c) = \exp\big(-\|x(c) - \mu_k\|^2 / 2\sigma_k^2\big), \tag{53}$$

$$w_k(c) = \operatorname{softmax}_k(\alpha_k \, g_k(c)), \tag{54}$$

$$\phi(c) = \sum_k w_k(c) \, f_k, \tag{55}$$

followed by a residual MLP decoder to $\mathbb{R}^{128}$.

B.1.4 Coordinate preprocessing

Each backbone receives one of four fixed coordinate transforms, denoted $g(\cdot)$ throughout the zoo:

WrapNorm Used by all SIREN/WIRE variants. Maps (λ, ϕ) to a wrapped 5-dimensional spherical feature with an additional $\pi/2$ longitude phase shift, allowing for a simple yet effective encoding with no discontinuities at the meridian:

$$g_{\text{WN}}(\lambda, \phi) = \big(\cos\phi\cos\lambda,\ \cos\phi\sin\lambda,\ \sin\phi,\ \cos\phi\cos(\lambda + \tfrac{\pi}{2}),\ \cos\phi\sin(\lambda + \tfrac{\pi}{2})\big) \in \mathbb{R}^5.$$

Cartesian 3D Used by NeRF PE, RBF, and Gaussian splatting. Maps a geodesic coordinate pair to the unit sphere:

$$g_{C3}(\lambda, \phi) = \big(\cos\phi\cos\lambda, \cos\phi\sin\lambda, \sin\phi\big).$$

Spherical Used by Spherical Harmonics. The implementation shifts to $\varphi = \text{deg2rad}(\lambda+180)$ and $\theta = \text{deg2rad}(\phi + 90)$ before evaluating the SH basis functions.

Raw degrees Used only as the public interface to the hash-grid variants. Internally the encoding wraps longitude, clamps latitude, and maps (λ, ϕ) to $[0, 1) \times [0, 1)$ before querying the multi-resolution tables.

B.1.5 Gated hybrid construction

Given two standalone backbones $f_A, f_B : \mathbb{R}^2 \to \mathbb{R}^{128}$, the gated hybrid is

$$\tilde{a} = \text{LN}_A(f_A(c)), \qquad \tilde{b} = \text{LN}_B(f_B(c)), \tag{56}$$

$$g = \sigma(\alpha) \in (0, 1)^{128}, \qquad \alpha \in \mathbb{R}^{128}, \tag{57}$$

$$z = g \odot \tilde{a} + (1 - g) \odot \tilde{b}, \tag{58}$$

$$f_{\text{hybrid}}(c) = \mathcal{C}(z), \tag{59}$$

where LN_A, LN_B are affine LayerNorms, α is initialised to zero so that $g = 0.5$ at step 0 (symmetric blend), and $\mathcal{C}$ is a post-combine head with two variants:

- `linear`: $\mathcal{C}(z) = Wz + b$, with $W = I$, $b = 0$ at initialisation (identity-initialised, so the hybrid reduces to the blend at step 0).
- `mlp`: $\mathcal{C}(z) = z + \text{MLP}(z)$, where MLP is a 4×256 GELU residual MLP with zero-initialised final projection. This guarantees $\mathcal{C}(z) = z$ at step 0 and the MLP correction is learned purely as a residual on top of the linear blend.

Our hybrid family fixes f_A to one of {Simple SIREN, Residual SIREN, SH+SIREN U-Net, WIRE U-Net} and f_B to one of {Hash-S, Hash-M, Hash-L, Gaussian Splat}, yielding $4 \times 4 = 16$ pairs. Each pair is evaluated with both combiners, for 32 hybrid variants in total. The per-dimension gate g serves also as a diagnostic: after training, the mean of g over the 128 feature dimensions measures how much of the final embedding is contributed by the continuous branch, and the std measures the spread of per-dimension specialisation.

B.1.6 Task heads and loss

The backbone output is LayerNormed again and fed to 51 independent linear task heads, one per field:

$$\hat{y}_t(c) = w_t^\top \text{LN}(f_\theta(c)) + b_t, \qquad t \in \{1, \ldots, 51\}. \tag{60}$$

Each field is z-score normalised at load time using the training split, so the task-loss is the mean squared error in normalised space:

$$\mathcal{L}(\theta) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{|\mathcal{B}_t|} \sum_{(c,y)\in\mathcal{B}_t} \big(\hat{y}_t(c) - y\big)^2, \tag{61}$$

with $T = 51$ and $|\mathcal{B}_t| = 512$. A single backbone forward pass serves all tasks in a step: coordinates from all tasks are concatenated, embedded once, and routed to their respective linear heads.

B.1.7 Training protocol

All runs use AdamW Loshchilov and Hutter (2019) (weight decay 0.05), gradient clipping at ℓ_2 norm 1.0, batch size $B = 512$ per task, and automatic mixed precision (`bf16`). The embedding dimension is fixed to 128 for every variant. We hold out 5 000 validation points per field (fixed seed for the data split); the remaining points form the training pool.

Phase 1: screening. 12 standalone backbones trained at seed 42 under a 7-point log-uniform LR ladder $\{10^{-6}, 3{\cdot}10^{-6}, 10^{-5}, 3{\cdot}10^{-5}, 10^{-4}, 3{\cdot}10^{-4}, 10^{-3}\}$, flat LR, 5 000 steps, eval every 500 steps.

Phase 2: hybrid screening. 32 hybrids trained across a 3-point local LR mini-sweep chosen from the global ladder and centred on the median of the two parent best LRs, one seed, 5 000 steps, to lock the per-hybrid LR before confirmation. The modal best LR is 10^{-4} (29/32 hybrids); only Simple SIREN×Hash-M, SH+UNet×Hash-M, and SH+UNet×Hash-S (MLP) peak at $3.16{\cdot}10^{-4}$.

Phase 3: confirmation. 38 variants (the 6 best standalones plus all 32 hybrids) are re-trained at their best LR with a cosine-with-warmup schedule (warmup fraction 5%), 15 000 steps, and three seeds $\{42, 137, 2024\}$.

Phase 4: extended runs (diagnostic). The top-3 confirmed variants are re-trained for 60 000 steps under the same cosine schedule at seed 42, to verify the short-budget ranking holds under extended training. Results used only for sanity-checking convergence.

B.1.8 Evaluation

The primary metric is validation PSNR in *physical* (denormalised) units, averaged over the 51 fields:

$$\mathrm{PSNR}_t = 10 \log_{10} \left(\frac{\mathrm{range}(y_t)^2}{\frac{1}{|\mathcal{V}_t|} \sum_{(c,y)\in\mathcal{V}_t} \big(\hat{y}_t(c) - y\big)^2} \right), \qquad \overline{\mathrm{PSNR}} = \frac{1}{T} \sum_t \mathrm{PSNR}_t. \tag{62}$$

Unless stated otherwise, reported numbers are the mean ± one standard deviation over the three confirmation seeds. SSIM, MAE, MSE, R^2, Pearson ρ, and RMSE are also logged per field for completeness (Table 11).

B.1.9 Results

The best performing model is the gated hybrid with linear combiner: **WIRE U-Net × Hash Grid (L)**, achieving $\overline{\mathrm{PSNR}} = 43.05 \pm 0.08$ dB at 50.1 M parameters, and winning 30/51 fields against the best standalone on each field (Table 11). The `mlp`-combiner variant is second at 42.98 ± 0.08 dB. The two combiners are indistinguishable in aggregate mean PSNR and in per-field rankings, so the extra post-mix MLP is not warranted by its parameter overhead. This is generally true for every hybrid: adding an MLP combiner provides no improvements in terms of accuracy but they add unneeded computational overhead.

Pareto frontier. Figure 18 places every confirmation variant in the (parameters, $\overline{\text{PSNR}}$) plane. The Pareto frontier has five points: the three Hash Grid standalones (`Hash-S` 2.2 M / 40.14 dB, `Hash-M` 3.3 M / 40.83 dB, `Hash-L` 5.9 M / 41.38 dB) in the low-parameter regime, and the two best hybrids (Simple SIREN×Hash-M at 46.0 M / 42.55 dB and the winner at 50.1 M / 43.05 dB) in the high-parameter regime. No standalone continuous backbone is Pareto-optimal: WIRE U-Net (44.1 M / 40.54 ± 0.05 dB) and Simple SIREN (42.7 M / 40.33 ± 0.02 dB) are both strictly dominated by Hash-L, which uses roughly $7\times$ fewer parameters and attains higher PSNR. The winner dominates both standalone parents in PSNR while trading parameters for accuracy in a principled way.

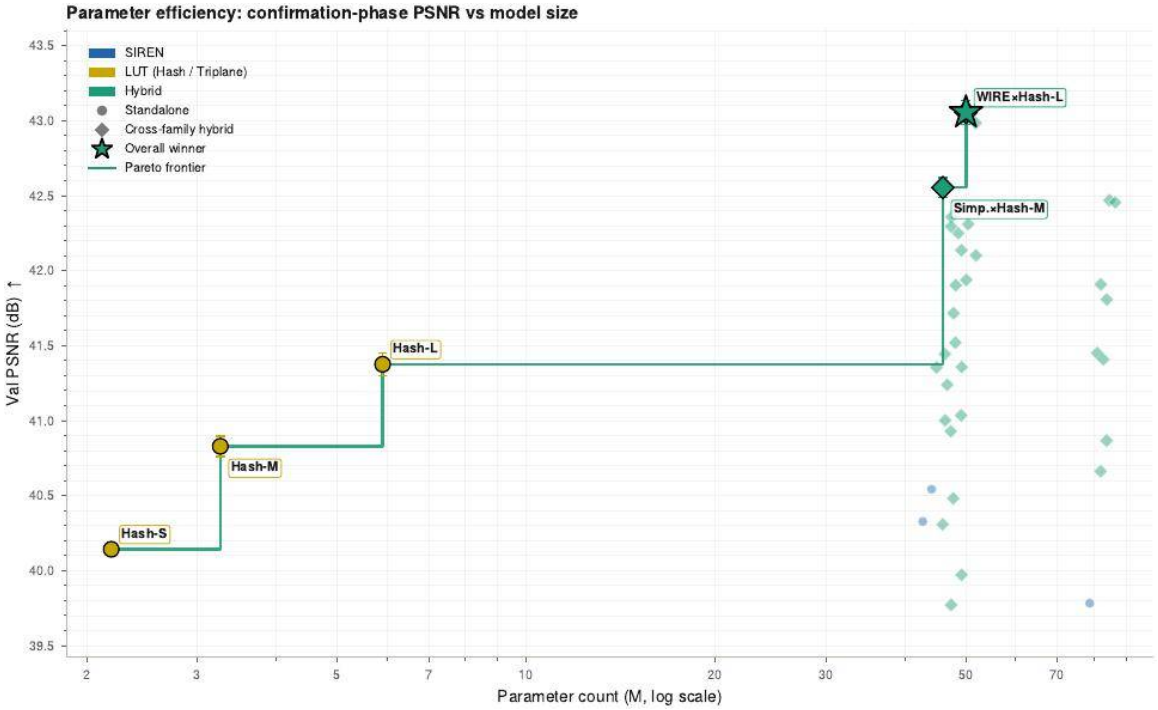

Figure 18: **Efficiency Pareto.** Parameter count vs. validation PSNR for every confirmation variant. Points on the Pareto frontier are bolded; the star marks the winner. Log-scaled x-axis.

Latent geometry. Figure 19 visualises a 3-component ICA projection of each encoder's learned 128-d embedding, and mapped to RGB. Continuous backbones produce smooth, globally organised latent maps with strong zonal gradients; hash-grid LUTs produce high-frequency, locally textured maps without a coherent global structure; the hybrids' final embedding (right panel of Fig. 20) blends both smooth zonal gradients carried by the continuous branch plus high-frequency detail obtained from the LUT branch.

Per-field behaviour. Figure 21 renders winner, best continuous, and best LUT predictions on 8 representative fields spanning the PSNR spectrum (easy at top, hard at bottom). The smooth-vs-sparse regime boundary is visible within the figure:

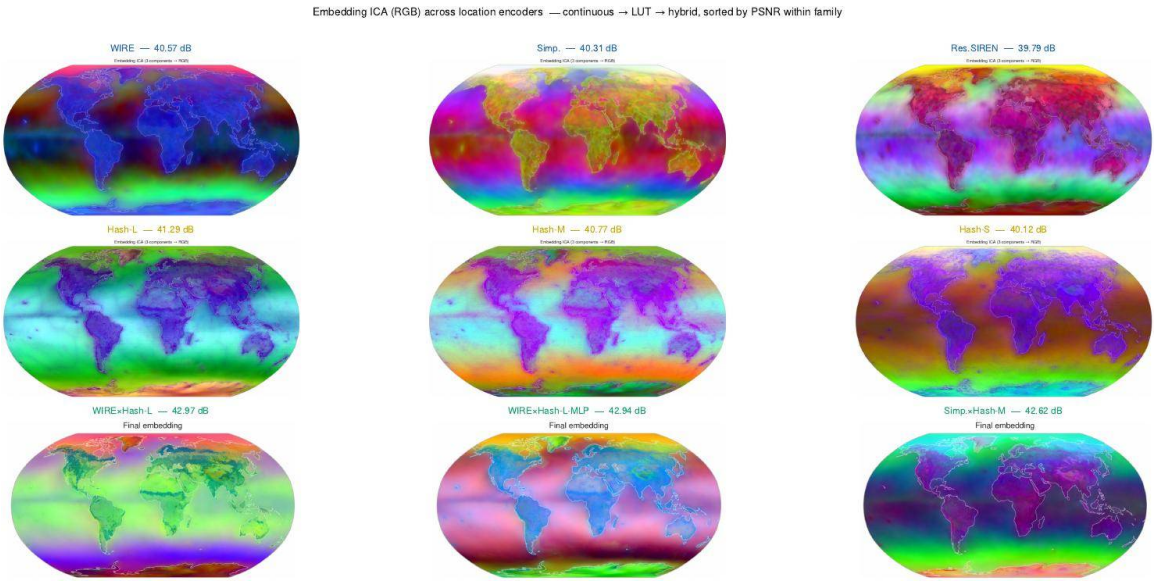

Figure 19: **Latent embedding geometry across encoders.** FastICA over the 128-d location embedding, three components mapped to RGB. Rows correspond to family (continuous, LUT, hybrid); subtitles report the mean confirmation PSNR. Hybrid panels show the final embedding only (see Fig. 20 for the per-branch decomposition).

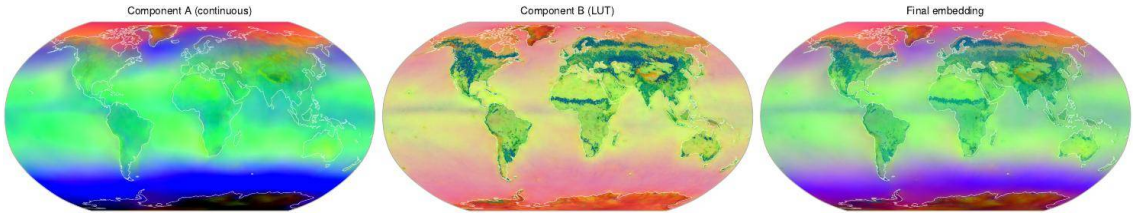

Figure 20: **Winner's per-branch ICA decomposition.** Left: continuous branch (WIRE U-Net) after LayerNorm, showing smooth zonal structure with low-frequency global gradients. Middle: LUT branch (Hash Grid (L)) after LayerNorm, showing localised high-frequency detail without global coherence. Right: final embedding after the per-dimension gate and linear combiner. The final embedding inherits the global gradients from the continuous branch and the sparse, discontinuous detail from the LUT branch.

- Row 1 (`so2`) saturates the PSNR cap for all three models and serves mostly as a sanity-check that trivially localised structure is easy for every backbone.
- On smooth and broad climatological fields (`msl`, `t2m`, `blh`), the hybrid is clearly best and improves over both WIRE and Hash-L simultaneously.
- On moderately irregular ecological and land-management fields (`evi`, `fertilizer`), the hybrid remains best or ties Hash-L while staying visibly less noisy than the pure LUT baseline.
- On the hardest sparse/extreme fields (`spi_03`, `population_density`), WIRE outperforms both LUT-heavier alternatives; the hybrid still stays above Hash-L on `population_density` but gives back much of the WIRE advantage.

Quantitative regime boundary. Figure 22 makes the regime boundary explicit across all 51 fields. For each field t we compute the paired per-seed delta $\Delta_t^{(s)} = \mathrm{PSNR}_{s,t}^{\mathrm{winner}} - \mathrm{PSNR}_{s,t}^{\mathrm{baseline}}$ and plot the mean over seeds $s \in \{42, 137, 2024\}$ with ± 1 std error bars. Fields with $|\bar{\Delta}_t| < 0.25\,\mathrm{dB}$ against both baselines are suppressed (6/51 fields), leaving 45 fields where at least one comparison is meaningfully non-zero. We can observe that:

- The hybrid exceeds Hash-L on 39/45 shown fields and falls behind it on 6/45, so the LUT branch is usually helpful but not uniformly dominated.
- The hybrid exceeds WIRE on 34/45 shown fields and falls behind it on 11/45. The clearest WIRE advantages occur on `burned_area`, `livestock_cattle`, `livestock_chickens`, `population_density`, and several drought / land-cover fields (`spi_03`, `spi_12`, `spei_12`, `vegclass_*`).
- The largest gain over Hash-L is $+5.5\,\mathrm{dB}$ on `msl_mean`, and the largest gain over WIRE is $+7.7\,\mathrm{dB}$ on the same field; the largest deficit to either parent is $-3.9\,\mathrm{dB}$ to WIRE on `burned_area`.

The resulting picture is asymmetric rather than *no-lose*: the hybrid substantially improves over Hash-L across most smooth and moderately irregular fields, but the continuous WIRE parent retains a real edge on a noticeable sparse/extreme subset.

Tables. Table 10 lists Phase 1 results (best-LR PSNR per standalone backbone, seed 42). Table 11 lists the Phase 3 confirmation leaderboard: mean ± std over three seeds, together with parameter count and family tag.

Table 10: Location encoder screening results — 12 standalone backbones. Best learning rate selected by maximum mean Val PSNR (single seed, seed 42). Winner highlighted. Values computed on held-out validation points (skip_plots=True). Single seed (seed 42). LR selected by highest mean Val PSNR across 51 WorldTensor fields.

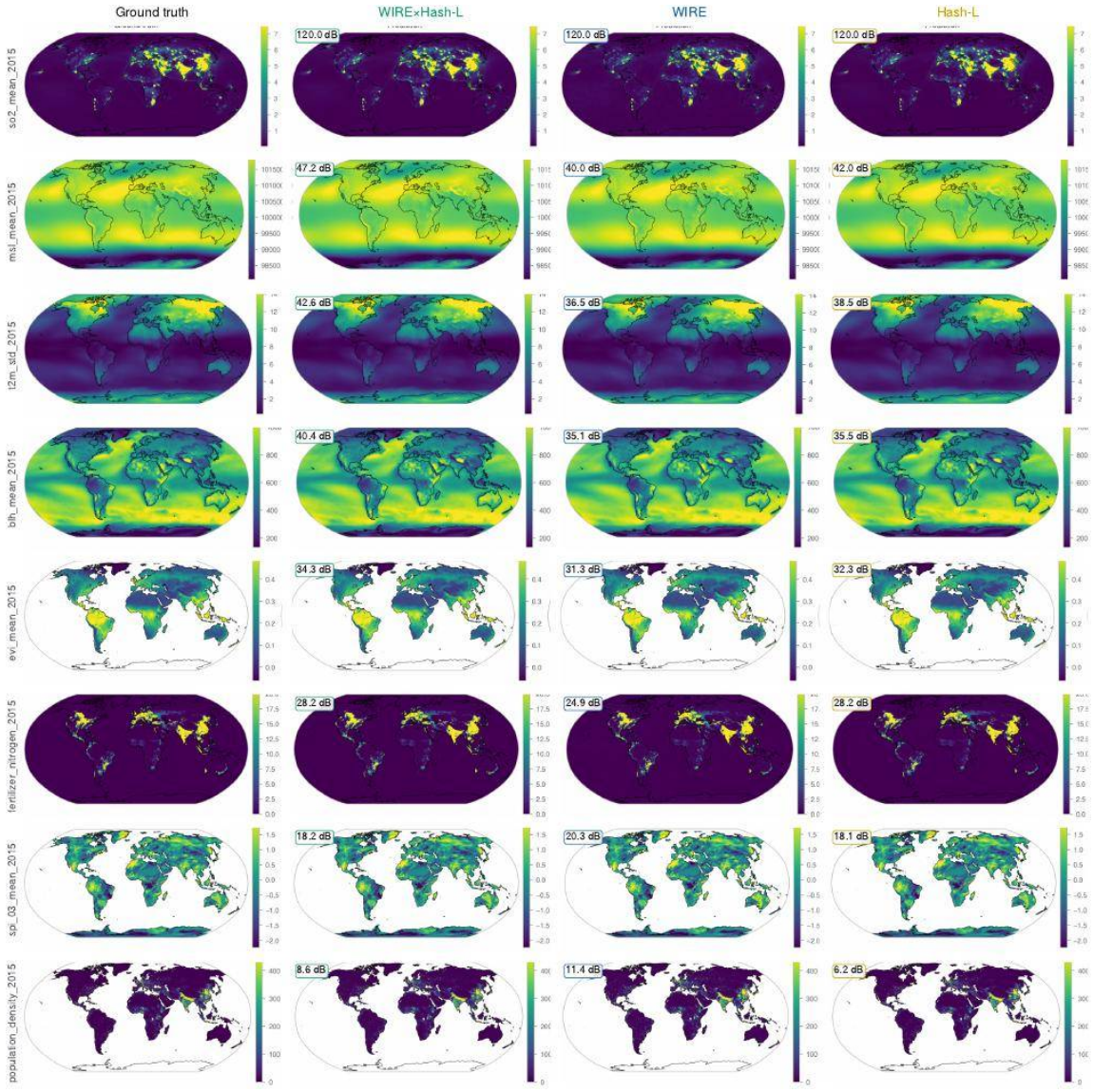

Figure 21: **Predictions across** 8 **representative fields.** Easy fields on top, hard fields at the bottom. Columns from left to right: ground truth, winner (WIRE U-Net×Hash Grid (L)), best continuous standalone (WIRE U-Net), best LUT standalone (Hash Grid (L)). Per-panel badges show denormalised PSNR (dB).

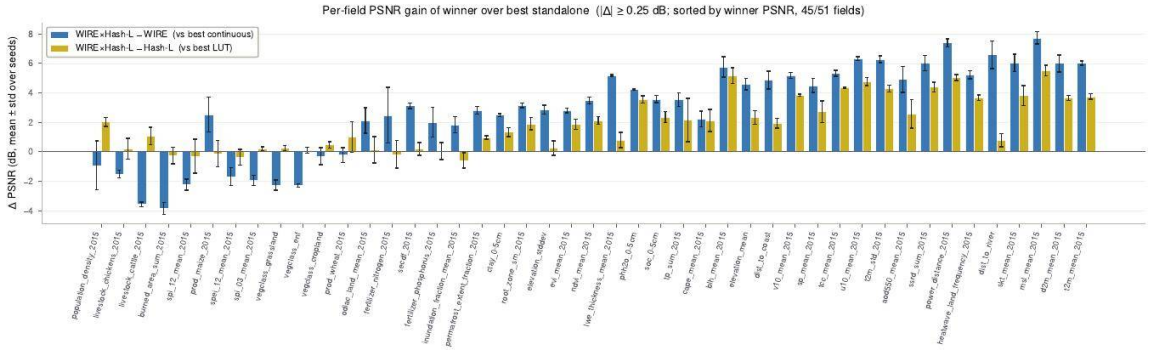

Figure 22: **Per-field PSNR gain of the hybrid winner over the best continuous and best LUT standalones.** $\overline{\Delta}_t$ is the per-seed mean gain in dB; error bars show ±1 std across three seeds. Fields sorted left-to-right by winner PSNR (hardest to easiest); fields with both bars within ±0.25 dB are suppressed (6/51).

Backbone	Family	Params	Best LR	Val PSNR (dB)↑	Val MSE↓	Pearson r↑
Hash Grid (L)	LUT	5M	3.16×10^{-4}	**40.02**	**0.1152**	0.93
Hash Grid (M)	LUT	3M	10^{-3}	39.21	0.1116	0.93
Hash Grid (S)	LUT	2M	10^{-3}	38.50	0.0961	0.94
WIRE U-Net	SIREN	44M	3.16×10^{-5}	38.19	0.1028	0.93
Simple SIREN	SIREN	42M	10^{-4}	38.16	0.1044	0.93
Residual SIREN	SIREN	78M	3.16×10^{-5}	38.02	0.1024	0.93
SH + SIREN U-Net	SIREN	44M	10^{-4}	37.96	0.1073	0.93
SIREN U-Net	SIREN	44M	3.16×10^{-4}	37.95	0.1078	0.93
NeRF PE + MLP	Fourier / PE	1M	10^{-3}	37.12	0.1001	0.93
Sph. Harmonics + MLP	Fourier / PE	1M	10^{-3}	36.62	0.1313	0.91
RBF + MLP	RBF / Kernel	2M	10^{-3}	36.39	0.1422	0.91
Gaussian Splatting	LUT	3M	10^{-3}	33.33	0.2230	0.85

Table 11: Location encoder confirmation results. Val PSNR (dB, ↑), Val MSE (↓), and Pearson r (↑) averaged over 51 WorldTensor fields and 3 seeds. Best entry **bold** and highlighted. Confirmation contains the 6 best standalones plus all 32 hybrids formed from the top-4 continuous × top-4 LUT screening survivors. Values are mean ± s.d. across 3 seeds. Cosine LR schedule, best LR selected during screening. Hybrids use a per-dimension learnable gate.

Backbone	Family	Params	Val PSNR (dB)↑	Val MSE↓	Pearson r↑
WIRE U-Net × Hash Grid (L)	Hybrid	50M	**43.05 ± 0.08**	**0.1050 ± 0.0056**	0.94 ± 0.00
WIRE U-Net × Hash Grid (L) (MLP)	Hybrid	51M	42.98 ± 0.08	0.1062 ± 0.0017	0.94 ± 0.00
Simple SIREN × Hash Grid (M)	Hybrid	46M	42.55 ± 0.07	0.0914 ± 0.0024	0.94 ± 0.00
Residual SIREN × Hash Grid (L)	Hybrid	84M	42.47 ± 0.11	0.1184 ± 0.0064	0.93 ± 0.00
Residual SIREN × Hash Grid (L) (MLP)	Hybrid	86M	42.45 ± 0.02	0.1097 ± 0.0017	0.93 ± 0.00
SH + SIREN U-Net × Hash Grid (M)	Hybrid	47M	42.36 ± 0.11	0.0934 ± 0.0043	0.94 ± 0.00
Simple SIREN × Hash Grid (L) (MLP)	Hybrid	50M	42.31 ± 0.05	0.1042 ± 0.0052	0.94 ± 0.00
WIRE U-Net × Hash Grid (M)	Hybrid	47M	42.29 ± 0.04	0.0896 ± 0.0026	0.95 ± 0.00
Simple SIREN × Hash Grid (L)	Hybrid	48M	42.25 ± 0.08	0.1050 ± 0.0033	0.94 ± 0.00
WIRE U-Net × Hash Grid (M) (MLP)	Hybrid	49M	42.14 ± 0.07	0.0949 ± 0.0033	0.94 ± 0.00
SH + SIREN U-Net × Hash Grid (L) (MLP)	Hybrid	51M	42.10 ± 0.02	0.1058 ± 0.0035	0.94 ± 0.00
SH + SIREN U-Net × Hash Grid (L)	Hybrid	50M	41.94 ± 0.03	0.1117 ± 0.0020	0.93 ± 0.00
Residual SIREN × Hash Grid (M)	Hybrid	82M	41.91 ± 0.02	0.0998 ± 0.0031	0.94 ± 0.00
SH + SIREN U-Net × Hash Grid (S) (MLP)	Hybrid	48M	41.90 ± 0.09	0.0775 ± 0.0014	0.95 ± 0.00
Residual SIREN × Hash Grid (M) (MLP)	Hybrid	83M	41.81 ± 0.03	0.0956 ± 0.0015	0.94 ± 0.00
Simple SIREN × Hash Grid (M) (MLP)	Hybrid	47M	41.72 ± 0.01	0.0930 ± 0.0012	0.94 ± 0.00
WIRE U-Net × Hash Grid (S) (MLP)	Hybrid	48M	41.52 ± 0.04	0.0760 ± 0.0013	0.95 ± 0.00
Residual SIREN × Hash Grid (S)	Hybrid	80M	41.45 ± 0.07	0.0776 ± 0.0014	0.95 ± 0.00
WIRE U-Net × Hash Grid (S)	Hybrid	46M	41.44 ± 0.03	0.0832 ± 0.0017	0.95 ± 0.00
Residual SIREN × Hash Grid (S) (MLP)	Hybrid	82M	41.41 ± 0.01	0.0767 ± 0.0012	0.95 ± 0.00
Hash Grid (L)	LUT	5M	41.38 ± 0.08	0.1171 ± 0.0104	0.93 ± 0.00
SH + SIREN U-Net × Hash Grid (M) (MLP)	Hybrid	49M	41.36 ± 0.05	0.0913 ± 0.0038	0.94 ± 0.00
Simple SIREN × Hash Grid (S)	Hybrid	44M	41.35 ± 0.07	0.0811 ± 0.0002	0.95 ± 0.00
Simple SIREN × Hash Grid (S) (MLP)	Hybrid	46M	41.24 ± 0.01	0.0771 ± 0.0014	0.95 ± 0.00
WIRE U-Net × Gaussian Splatting (MLP)	Hybrid	49M	41.03 ± 0.04	0.0740 ± 0.0008	0.95 ± 0.00
SH + SIREN U-Net × Hash Grid (S)	Hybrid	46M	41.00 ± 0.04	0.0822 ± 0.0020	0.95 ± 0.00
WIRE U-Net × Gaussian Splatting	Hybrid	47M	40.93 ± 0.05	0.0741 ± 0.0007	0.95 ± 0.00
Residual SIREN × Gaussian Splatting (MLP)	Hybrid	83M	40.87 ± 0.04	0.0748 ± 0.0020	0.95 ± 0.00
Hash Grid (M)	LUT	3M	40.83 ± 0.07	0.0937 ± 0.0032	0.94 ± 0.00
Residual SIREN × Gaussian Splatting	Hybrid	81M	40.66 ± 0.02	0.0744 ± 0.0008	0.95 ± 0.00
WIRE U-Net	SIREN	44M	40.54 ± 0.05	0.0747 ± 0.0006	0.95 ± 0.00
Simple SIREN × Gaussian Splatting (MLP)	Hybrid	47M	40.48 ± 0.05	0.0754 ± 0.0002	0.95 ± 0.00
Simple SIREN	SIREN	42M	40.33 ± 0.02	0.0782 ± 0.0004	0.95 ± 0.00
Simple SIREN × Gaussian Splatting	Hybrid	45M	40.31 ± 0.04	0.0775 ± 0.0007	0.95 ± 0.00
Hash Grid (S)	LUT	2M	40.14 ± 0.03	0.0784 ± 0.0002	0.95 ± 0.00
SH + SIREN U-Net × Gaussian Splatting (MLP)	Hybrid	49M	39.97 ± 0.01	0.0786 ± 0.0008	0.95 ± 0.00
Residual SIREN	SIREN	78M	39.78 ± 0.02	0.0806 ± 0.0003	0.95 ± 0.00
SH + SIREN U-Net × Gaussian Splatting	Hybrid	47M	39.77 ± 0.03	0.0814 ± 0.0002	0.95 ± 0.00

B.1.10 Discussion

The ablation shows that the trade-off that separates implicit neural representations from lookup-table encodings is *architecturally mixable*. A symmetric, identity-initialised gate over LayerNormed sub-embeddings learns to route each latent dimension to the branch whose inductive bias fits the per-dimension signal, and at inference the resulting hybrid is the best overall confirmation variant while strictly outperforming both named parents on 31/51 WorldTensor fields and tying or beating the better parent on 40/51 (within a 0.25 dB tolerance).

Three design choices matter:

1. **LayerNorm before the gate.** Without it the two branches produce embeddings at disparate magnitudes and the gate becomes a scale fixer rather than a routing mechanism. LayerNormed branches isolate g to its intended role (per-dimension mixture), which is what makes the gate values diagnostically meaningful.
2. **Identity-initialised combiner.** Both `linear` and `mlp` combiners are initialised so that $\mathcal{C}(z) = z$ at step 0, guaranteeing that the hybrid reduces to a simple 50/50 blend of its parents at initialisation. This yields stable initial training even when the two branches are optimised at different effective learning rates.
3. **Combiner complexity not warranted.** The `mlp` combiner is indistinguishable from the `linear` combiner in aggregate (Fig. 18, top two ranks), confirming that the gain is from *mixing* two complementary priors, not from additional post-mix nonlinearity.

The remaining gaps concentrate on very sparse, extreme, and partly categorical fields, suggesting a future direction toward per-task or spatially adaptive gating, or uncertainty-aware branch selection that learns when to trust the continuous prior more heavily. We leave this to the uncertainty extension of the model.

The chosen configuration is ***WIRE U-Net × Hash Grid (L) hybrid with a 128-dimensional embedding.*** The ablation is run at 128 dimensions to keep the sweep affordable; the production model scales the selected architecture to 256, as it does for the other encoders.

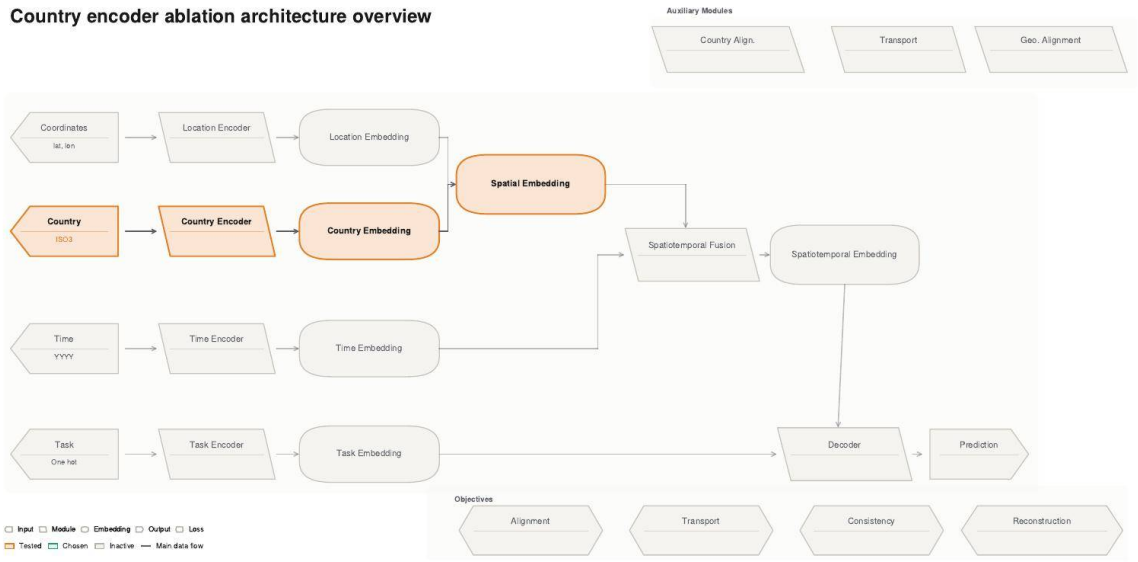

Figure 23: **Country encoder ablation setup.** Components marked in blue were fixed or carried forward, components in orange were tested in this study, and inactive components were not used. The experiment isolates country representation design while holding the rest of the model scaffold fixed.

B.1.11 Country encoder

Motivation

The country encoder is intended to provide a compact representation of country identity that can support a large number of country-level prediction tasks within a shared model. Unlike the location encoder, its input is discrete and low-dimensional, so the main question is not geometric expressivity but representational sufficiency: how much embedding capacity is required to encode heterogeneous national attributes, and whether a learned nonlinear refinement is useful beyond a plain embedding table. We therefore study the country encoder as a representation-capacity benchmark rather than as a transfer or generalization benchmark.

Configurations tested

We evaluated two country-encoder families: a pure learned embedding table (obtained by transforming ISO3 code into a One-hot encoded representation) and a learned embedding followed by a residual MLP. For both families, we swept the embedding dimensionality over $\{8, 16, 32, 64, 128\}$. The residual MLP used four hidden layers of width 256 with learned residual scaling, matching the general design adopted in the other encoder studies. Each configuration was trained in a shared multitask setting to reconstruct 150 country-level variables at the target year 2015 over all available countries. Targets were standardized for optimization and the model was trained with mean squared error. Following the common ablation protocol, we first performed a single-seed learning-rate screening and then a three-seed confirmation run at the best learning rate for each configuration.

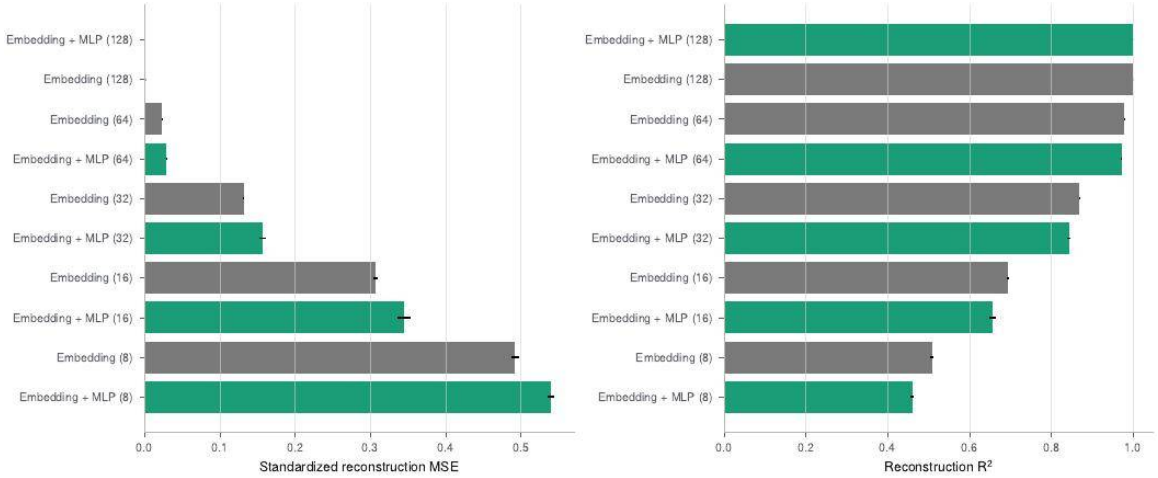

Figure 24: **Country encoder ablation.** Comparison of embedding-only and embedding-plus-residual-MLP country encoders across embedding dimensionalities. Performance is reported as mean standardized reconstruction error over the 150-task benchmark at year 2015, with error bars showing variability across seeds. The best overall configuration is obtained with a 128-dimensional embedding followed by a residual MLP.

Let $z_c = E_c(c)$ denote the latent representation of country c, and let

$$\hat{y}_{c,v} = D_v(z_c) \tag{63}$$

be the prediction for variable v. The training objective is

$$\mathcal{L}_{\text{country}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \text{MSE}(\hat{y}_{c,v}, y_{c,v}) \,, \tag{64}$$

where $|\mathcal{V}| = 150$.

Results

The main result is shown in Fig. 24. Performance improves sharply with embedding dimensionality, indicating that the dominant bottleneck in this benchmark is representational capacity rather than encoder depth. Small embeddings underfit the 150-task reconstruction problem; accuracy improves substantially with dimensionality, with the largest gain between 64 and 128 dimensions (so 64 is not a saturation point but a capacity choice).

At the highest capacity, a plain embedding table is already highly competitive, which indicates that much of the country-level signal can be captured by a sufficiently expressive discrete latent representation. However, the best overall configuration is obtained by adding a residual MLP on top of the 128-dimensional embedding. Table 12 shows that this configuration achieves the lowest mean standardized reconstruction error, outperforming the plain 128-dimensional embedding while remaining architecturally simple.

The family-wise analysis in Fig. 25 shows that this pattern is consistent across broad groups of country-level variables rather than being driven by a small subset of tasks. In other words, the gain from larger embedding size generalizes across heterogeneous socioeconomic,

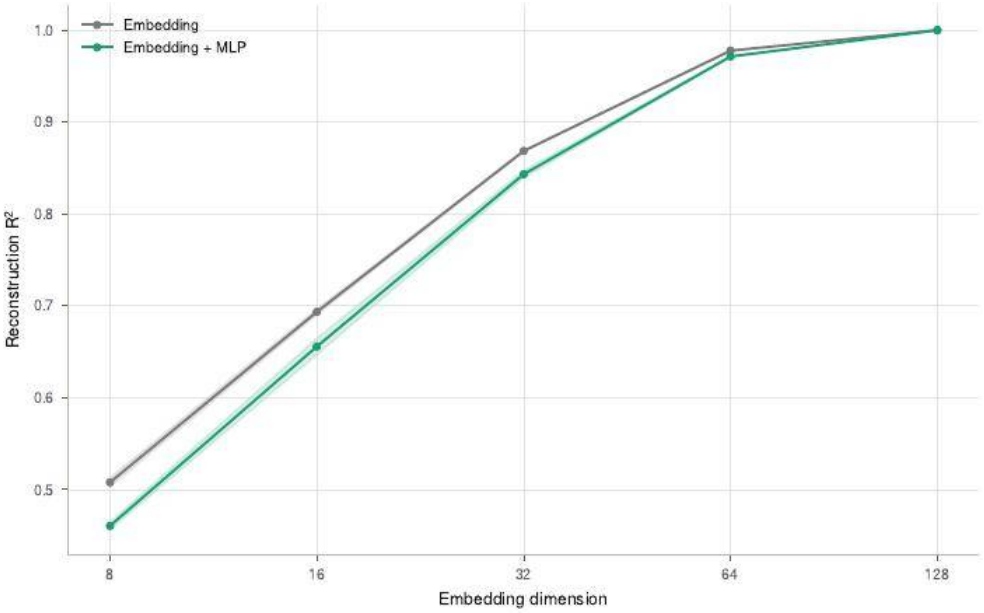

Figure 25: **Country encoder performance by variable family.** Performance trends across major variable families as embedding dimensionality increases. The dominant effect is improved representational capacity, while the residual MLP provides an additional gain once sufficient embedding size has been reached.

institutional, and demographic targets. The residual MLP provides a smaller but still consistent improvement once sufficient latent capacity has been reached.

Figure 23 summarizes the scope of this study. Only the country-encoding pathway is varied, while the rest of the architecture remains inactive or fixed. This makes the resulting comparison easy to interpret: the observed improvements can be attributed directly to country representation design rather than to changes elsewhere in the model.

Conclusion

We select the country encoder consisting of a learned embedding followed by a residual MLP as the default country representation for subsequent experiments. This choice provides the best overall fit while remaining architecturally simple. The ablation also suggests that, for country identity, embedding capacity is more important than architectural depth, and that nonlinear refinement is beneficial only after sufficient latent dimensionality has been reached. The production model deliberately operates this encoder at 64 dimensions — a capacity trade-off that sacrifices some country-reconstruction accuracy so the low-dimensional country pathway does not dominate the much higher-capacity spatial bus in the shared space; the residual MLP is retained for its role in the full model, where the country encoder interacts with the alignment objectives, even though in this isolated panel-reconstruction benchmark it helps mainly at higher dimensionality.

The chosen configuration is a ***Residual MLP with a 128-dimensional embedding****.*

			Standardized fit		
Variant	Best LR	Trainable params	MSE ↓	MAE ↓	R^2 ↑
Embedding + MLP (128)	10^{-3}	2,255,126	**0.0002 ± 0.0000**	**0.0085 ± 0.0001**	**0.9998 ± 0.0000**
Embedding (128)	10^{-3}	84,886	0.0004 ± 0.0000	0.0125 ± 0.0002	0.9996 ± 0.0000
Embedding (64)	10^{-3}	42,518	0.0226 ± 0.0001	0.1074 ± 0.0005	0.9774 ± 0.0001
Embedding + MLP (64)	10^{-3}	2,179,926	0.0290 ± 0.0006	0.1220 ± 0.0012	0.9710 ± 0.0006
Embedding (32)	10^{-3}	21,334	0.1317 ± 0.0005	0.2516 ± 0.0008	0.8683 ± 0.0005
Embedding + MLP (32)	10^{-3}	2,142,326	0.1570 ± 0.0040	0.2705 ± 0.0032	0.8430 ± 0.0040
Embedding (16)	10^{-3}	10,742	0.3068 ± 0.0031	0.3643 ± 0.0021	0.6932 ± 0.0031
Embedding + MLP (16)	10^{-3}	2,123,526	0.3445 ± 0.0088	0.3879 ± 0.0058	0.6555 ± 0.0088
Embedding (8)	10^{-3}	5,446	0.4923 ± 0.0049	0.4583 ± 0.0013	0.5077 ± 0.0049
Embedding + MLP (8)	10^{-3}	2,114,126	0.5396 ± 0.0040	0.4825 ± 0.0023	0.4604 ± 0.0040

Values are mean ± s.d. across three confirmation seeds. Best values are shown in bold. The winning model, Embedding + MLP (128), was selected by standardized reconstruction MSE.

Table 12: **Country encoder confirmation results.** Mean ± standard deviation across confirmation seeds for the tested country-encoder configurations.

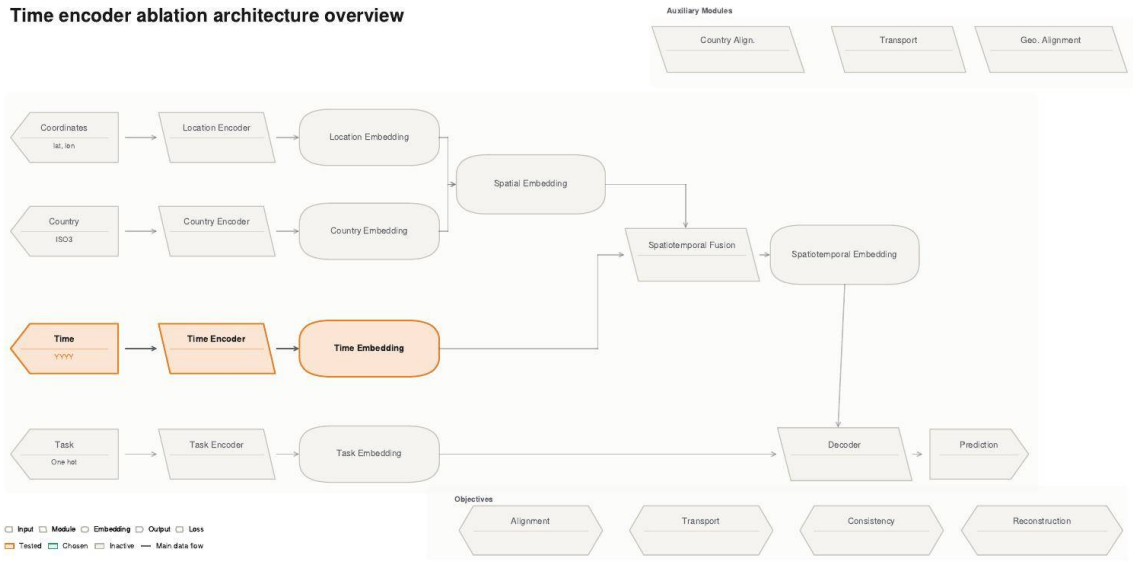

Figure 26: **Time encoder ablation setup.** Components marked in blue were fixed or carried forward, components in orange were tested in this study, and inactive components were not used.

B.1.12 Time encoder

Motivation

The time encoder is responsible for compressing calendar year into a compact latent representation that can support a diverse collection of temporal prediction tasks. Unlike the country encoder, its input is continuous and ordered, so the main design question is which temporal inductive bias is most appropriate: direct scalar projection, periodic feature mappings, localized basis functions, or hybrid combinations thereof. We therefore study the time encoder as a compact temporal representation problem, asking which architecture best captures heterogeneous long-term trends under a strong dimensional bottleneck.

Configurations tested

We evaluated eight time-encoder configurations with a fixed output dimensionality of 8. Let $E_t(\cdot)$ denote the time encoder and let $z_t = E_t(t)$ be the latent representation of year t. The tested families differed in how the scalar year input was lifted into a feature basis before projection into the latent space.

For the **raw scalar** family, the normalized year t is mapped directly to the latent space, either through a single linear projection or through a residual MLP refinement. For the **Fourier** family, the year is first mapped to a sinusoidal basis,

$$\phi_{\text{Fourier}}(t) = [\sin(2\pi\omega_1 t), \cos(2\pi\omega_1 t), \ldots, \sin(2\pi\omega_K t), \cos(2\pi\omega_K t)]\,, \tag{65}$$

where $\{\omega_k\}_{k=1}^{K}$ denotes the set of Fourier frequencies. This basis is then projected to the latent space, optionally followed by residual MLP refinement. We set $K = 32$ in this experiment.

For the **RBF** family, the year is mapped to a radial basis expansion,

$$\phi_{\mathrm{RBF}}(t) = \left[\exp\left(-\frac{(t-\mu_1)^2}{2\sigma^2}\right), \ldots, \exp\left(-\frac{(t-\mu_M)^2}{2\sigma^2}\right)\right], \tag{66}$$

where $\{\mu_m\}_{m=1}^{M}$ are learned or fixed temporal centers and σ controls the basis width. This representation is likewise projected to the latent space, optionally followed by residual MLP refinement. We set $M = 32$ in this experiment.

We also evaluated two hybrid families. In **Hybrid concat + MLP**, the Fourier and RBF expansions are concatenated before residual MLP refinement,

$$\phi_{\mathrm{hyb}}(t) = [\phi_{\mathrm{Fourier}}(t), \phi_{\mathrm{RBF}}(t)]. \tag{67}$$

In **Hybrid gated + MLP**, the two basis families are first projected into a shared hidden space and then combined through a learned gate before residual MLP refinement. In all MLP-based variants, the refinement network consisted of 4 hidden layers of width 256 with learned residual scaling.

The benchmark used 15 temporally varying WorldTensor variables spanning a range of interannual smoothness and volatility. For each variable v, the target $y_{t,v}$ was the area-weighted global annual mean at year t, and predictions were produced through variable-specific decoder heads

$$\hat{y}_{t,v} = D_v(z_t). \tag{68}$$

Targets were standardized for optimization, and all models were trained with mean squared error,

$$\mathcal{L}_{\mathrm{time}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathrm{MSE}(\hat{y}_{t,v}, y_{t,v}), \tag{69}$$

where $|\mathcal{V}| = 15$. Following the common ablation protocol, we first performed a single-seed learning-rate screening and then a three-seed confirmation run at the best learning rate for each configuration.

Results

The main result is shown in Fig. 27. Under the 8-dimensional bottleneck, **Fourier + MLP** clearly outperforms all alternative temporal parameterizations. This result is not marginal: it is the strongest configuration in the confirmation runs and provides the most favorable trade-off between overall fit quality and temporal-dynamics accuracy. In our benchmark, the best learning rate for this model was 10^{-3}.

This pattern is especially informative because the benchmark includes both smooth and irregular temporal signals. A pure raw-scalar encoder is too weak, while localized RBF-style encoders tend to behave more like high-capacity lookup mechanisms and do not organize time as cleanly across heterogeneous temporal regimes. Adding a residual MLP provides a decisive improvement, showing that basis selection alone is not sufficient; nonlinear refinement is needed to map temporal features into a useful shared latent space. However, once this refinement is present, the simpler Fourier basis remains superior to the more flexible hybrid alternatives.

The representative fits in Fig. 28 make this difference visible at the level of individual variables. **Fourier + MLP** tracks smooth long-term climate signals accurately while also remaining

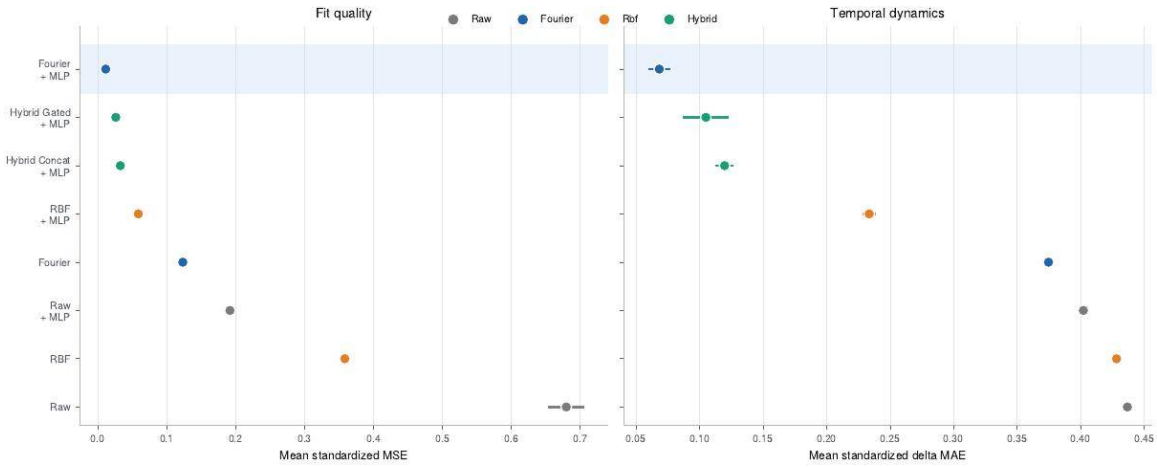

Figure 27: **Time encoder ablation.** Comparison of time-encoder variants on the 15-task temporal benchmark. Performance is reported as mean standardized reconstruction error, with error bars showing variability across seeds. **Fourier + MLP** is the best overall configuration.

			Fit quality			Dynamics
Variant	Best LR	Params	Std. MSE ↓	Corr. ↑	R^2 ↑	$\Delta\text{MAE}_{\text{std}}$ ↓
Fourier + MLP	10^{-3}	547,220	**0.011 ± 0.002**	**0.994 ± 0.001**	**0.856 ± 0.002**	**0.068 ± 0.009**
H. Gated + MLP	3.16×10^{-4}	556,437	0.026 ± 0.006	0.987 ± 0.003	0.841 ± 0.006	0.105 ± 0.018
H. Concat + MLP	3.16×10^{-4}	555,413	0.032 ± 0.001	0.984 ± 0.001	0.835 ± 0.001	0.120 ± 0.007
RBF + MLP	3.16×10^{-4}	539,028	0.058 ± 0.005	0.970 ± 0.003	0.810 ± 0.005	0.233 ± 0.005
Fourier	10^{-3}	656	0.123 ± 0.001	0.935 ± 0.001	0.776 ± 0.001	0.375 ± 0.003
Raw + MLP	10^{-3}	531,091	0.191 ± 0.004	0.894 ± 0.002	0.746 ± 0.007	0.402 ± 0.002
RBF	10^{-3}	400	0.358 ± 0.008	0.772 ± 0.008	0.625 ± 0.002	0.428 ± 0.000
Raw	10^{-3}	151	0.680 ± 0.026	0.637 ± 0.006	0.319 ± 0.026	0.437 ± 0.000

Values are mean ± s.d. across three confirmation seeds. Trainable parameters are exact counts. Best values are shown in bold. The winning model, Fourier + MLP, also achieved the best score on all 15 temporal variables.

Table 13: **Time encoder confirmation results.** Mean ± standard deviation across confirmation seeds for the tested time-encoder configurations.

competitive on noisier and more episodic series such as burned area and $PM_{2.5}$. In contrast, weaker baselines either underfit the global trend or introduce less coherent local deviations. The family-wise and field-wise summaries show that this is a global pattern rather than an isolated advantage on a small subset of variables.

Table 13 quantifies the ranking. **Fourier + MLP** achieves the best mean standardized reconstruction error and the best temporal-dynamics score, and in our confirmation study it is the strongest model on all 15 benchmark variables. Taken together, these results indicate that the dominant temporal structure in this benchmark is smooth and low-dimensional, and that the best encoder is one that captures this structure explicitly rather than over-parameterizing local temporal deviations.

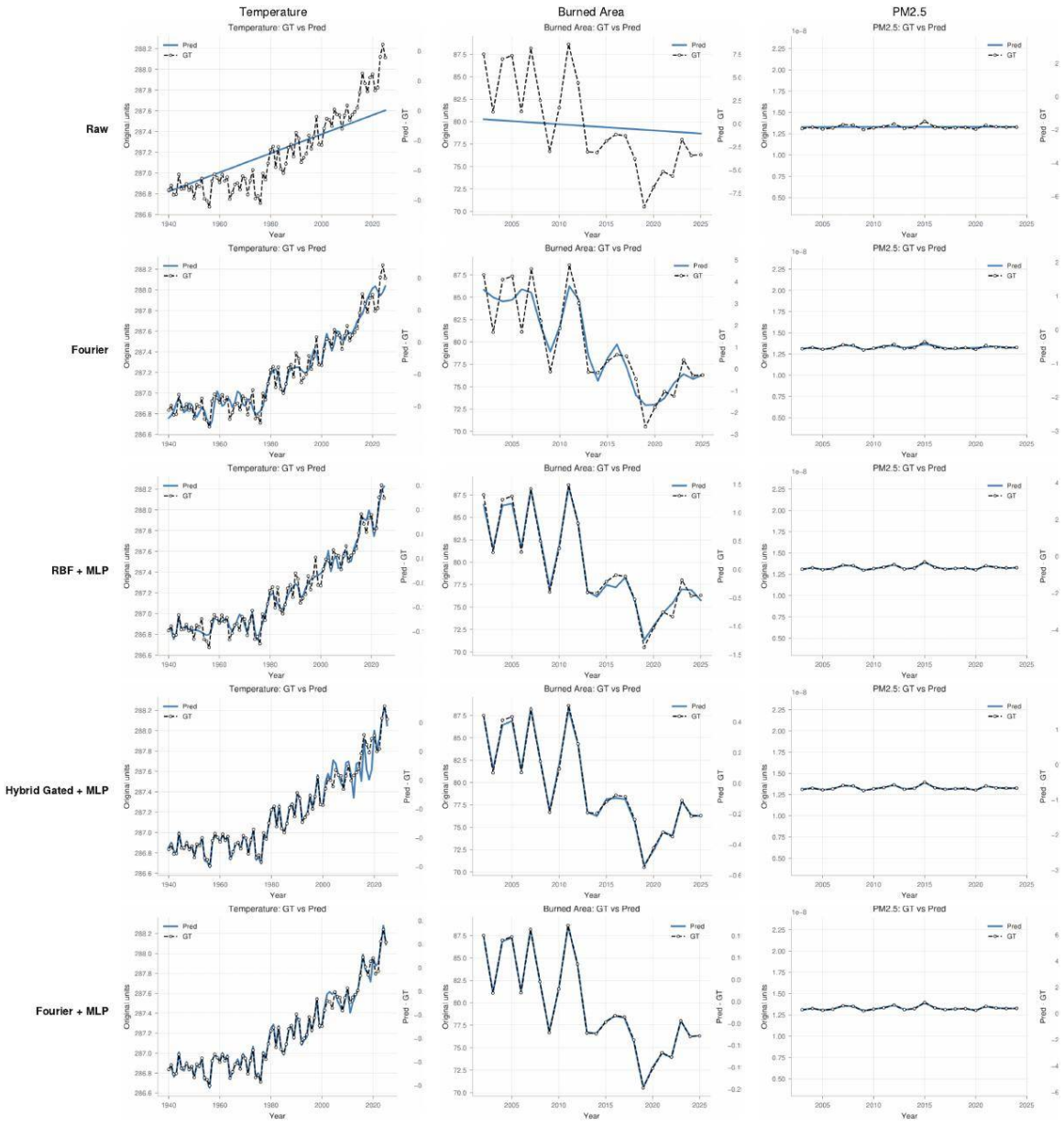

Figure 28: **Representative temporal fits.** Comparison of five time-encoder configurations on three representative variables. Rows correspond to encoder variants and columns to representative temporal regimes. Each panel shows only the main original-units fit view, allowing direct comparison of how well each encoder tracks the observed temporal trend.

Conclusion

We select the time encoder consisting of a Fourier-feature parameterization followed by a residual MLP, with an 8-dimensional output embedding, as the default temporal representation for subsequent experiments. This configuration provides the best trade-off between compactness, stability, and predictive performance. The ablation indicates that temporal structure in

TerraNova is best modeled through an explicitly smooth basis combined with a small learned nonlinear refinement. In contrast to the location encoder experiment, hybridising different sub-architectures did not provide significant improvements with respect to individual models. *The chosen configuration is* ***Fourier + MLP with an 8-dimensional embedding****.*

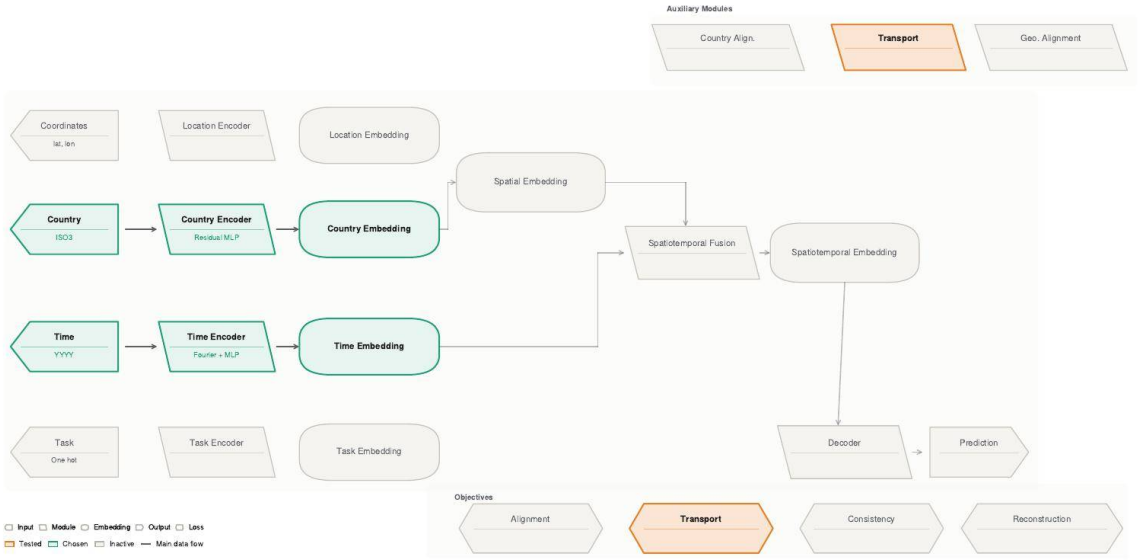

Figure 29: **Transport-module ablation setup.** The country encoder and time encoder were fixed to the strongest configurations identified in the previous studies, while the transport module and its associated objectives were varied. Inactive components were not used.

B.1.13 Transport module

Motivation

The purpose of the transport module is to model explicit temporal evolution in latent space. Once strong location and time encoders are available, the central question is whether an additional temporal operator is still useful, or whether direct prediction from location and time embeddings is already sufficient. We therefore study the transport module not as an isolated predictor, but as a learned operator acting on the time embedding. We want to understand if this module can help regularize the time embedding geometry and improve generalization on held-out years.

Configurations tested

In this study we fixed the country encoder to the strongest country-encoder configuration identified previously, namely a 128-dimensional embedding followed by a residual MLP, and fixed the time encoder to the winning **Fourier + MLP** configuration with an 8-dimensional embedding. Let $z_c = E_c(c)$ denote the country embedding for country c, and let $z_t = E_t(t)$ denote the time embedding for year t. Predictions were produced by a residual decoder $D(\cdot)$ operating on the concatenation of country and time embeddings.
The direct predictive pathway is therefore

$$\hat{y}^{\mathrm{dir}}_{c,t} = D([z_c, z_t])\,, \tag{70}$$

while the transported pathway introduces a transport operator $T(\cdot, \Delta)$ acting only on the time embedding:

$$\tilde{z}_{t+\Delta} = T(z_t, \Delta), \qquad \hat{y}^{\mathrm{tr}}_{c,t+\Delta} = D([z_c, \tilde{z}_{t+\Delta}])\,. \tag{71}$$

We compared five transport formulations:

- **Direct only**: no transport module, using only the direct pathway.
- **Transport prediction only**: transport is trained only through downstream predictive supervision.
- **Transport + match**: predictive supervision plus latent matching between transported and direct target-year embeddings.
- **Transport + match + identity**: latent matching plus an identity constraint at zero time shift.
- **Transport + full consistency**: latent matching, identity, and semigroup consistency.

All variants optimize the direct reconstruction term

$$\mathcal{L}_{\mathrm{dir}} = \mathrm{MSE}\Big(\hat{y}^{\mathrm{dir}}_{c,t}, y_{c,t}\Big)\,, \tag{72}$$

computed on standardized targets. When transport is enabled, we additionally use the transport-prediction term

$$\mathcal{L}_{\mathrm{pred}} = \mathrm{MSE}\big(\hat{y}^{\mathrm{tr}}_{c,t+\Delta}, y_{c,t+\Delta}\big)\,, \tag{73}$$

the latent-matching term

$$\mathcal{L}_{\mathrm{match}} = \|T(z_t, \Delta) - z_{t+\Delta}\|_2^2\,, \tag{74}$$

the identity term

$$\mathcal{L}_{\mathrm{id}} = \|T(z_t, 0) - z_t\|_2^2\,, \tag{75}$$

and the semigroup term

$$\mathcal{L}_{\mathrm{sg}} = \|T(T(z_t, \Delta_1), \Delta_2) - T(z_t, \Delta_1 + \Delta_2)\|_2^2\,. \tag{76}$$

The total objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{dir}} + \lambda_{\mathrm{pred}}\mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{match}}\mathcal{L}_{\mathrm{match}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{sg}}\mathcal{L}_{\mathrm{sg}}. \tag{77}$$

The weights were set as follows:

- **Transport prediction only**: $(\lambda_{\mathrm{pred}}, \lambda_{\mathrm{match}}, \lambda_{\mathrm{id}}, \lambda_{\mathrm{sg}}) = (1, 0, 0, 0)$
- **Transport + match**: $(1, 0.1, 0, 0)$
- **Transport + match + identity**: $(1, 0.1, 0.02, 0)$
- **Transport + full consistency**: $(1, 0.1, 0.02, 0.05)$

The benchmark used the same 150 variables as in the country-encoder study, but over their full temporal panel rather than at a single year. Evaluation was performed on blocked temporal windows for a subset of variables across all countries, so that the model had to reconstruct missing temporal spans rather than isolated random years.

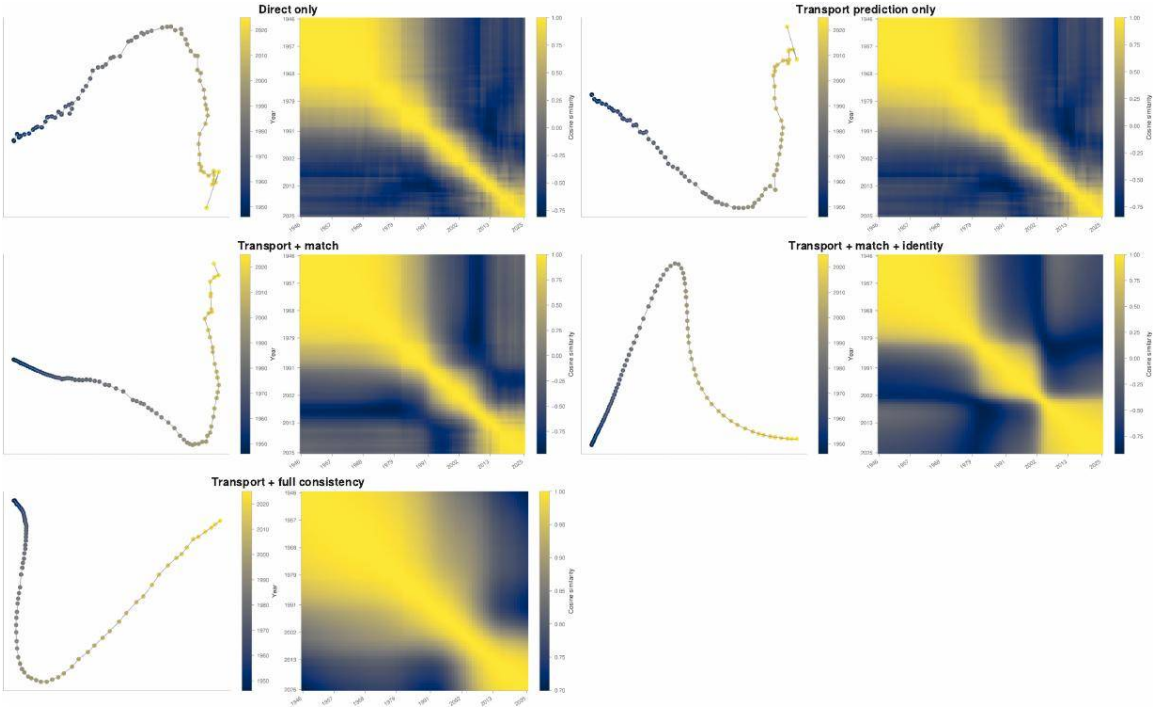

Figure 30: **Time-embedding geometry across transport configurations.** Comparison of the learned temporal geometry for the main transport formulations, using the confirmation-run time-embedding geometry plots for each configuration.

Results

The transport ablation yields a clean but nontrivial outcome. The best overall predictive model and the best explicit transport operator are not the same. Table 14 shows that **Transport + full consistency** achieves the lowest overall held-out error, indicating that transport improves the full model when it is constrained not only by predictive supervision but also by latent geometric structure. By contrast, **Transport + match + identity** yields the best transported-rollout performance, showing that matching and identity constraints are sufficient to produce the strongest explicit transported path. The full-consistency formulation yields the strongest overall temporal regularization, whereas the match-plus-identity formulation yields the cleanest explicit temporal operator. In other words, semigroup consistency appears to improve the structure of the model globally even when it does not produce the best direct rollout path. Table 15 supports this interpretation by showing that **Transport + full consistency** also gives the strongest aggregate geometry metrics, while Fig. 30 makes the qualitative differences in time-embedding structure visible across transport formulations. The no-transport configuration behaves like a look-up table, while the full transport configuration acts as a geometric regulariser and shows smoother embeddings, where dynamics are more predictable.

The key negative result is also clear from this study: predictive supervision alone is not enough. *Transport prediction only* does not outperform the direct baseline, which means that transport becomes useful only when accompanied by additional latent-consistency constraints.

			Held-out fit		Geometry	Transport alignment	
Variant	Best LR	Params	Direct MSE ↓	Transport MSE ↓	Score ↓	Match MSE ↓	Id. MSE ↓
T. + full consistency	3.16×10^{-4}	5,573,433	**0.2252 ± 0.0051**	0.2430 ± 0.0278	**0.058**	0.004 ± 0.000	0.000 ± 0.000
T. + match + id.	10^{-3}	5,573,433	0.2395 ± 0.0119	**0.2341 ± 0.0041**	0.255	0.001 ± 0.000	0.000 ± 0.000
Direct only	3.16×10^{-4}	3,450,663	0.2426 ± 0.0038	—	—	—	—
Transport pred. only	3.16×10^{-4}	5,573,433	0.2486 ± 0.0033	0.2489 ± 0.0028	0.869	0.013 ± 0.002	0.272 ± 0.031
Transport + match	3.16×10^{-4}	5,573,433	0.2519 ± 0.0027	0.2427 ± 0.0086	0.335	0.002 ± 0.000	0.150 ± 0.009

Table 14: **Transport ablation confirmation results.** Mean ± standard deviation across confirmation seeds for the direct-fit and transport-rollout metrics of the tested transport configurations. Values are mean ± s.d. across three confirmation seeds. Direct-fit winner (Transport + full consistency), transport-rollout winner (Transport + match + identity), and geometry winner (Transport + full consistency) are highlighted. Geometry score is the mean of min-max normalized match MSE, $(1 - \cos)$, identity MSE, and semigroup MSE.

Variant	Geometry score ↓	Match MSE ↓	$1 - \cos$ ↓	Identity MSE ↓	Semigroup MSE ↓
T. + match + id.	0.255	0.0012 ± 0.0002	0.00007 ± 0.00003	0.00003 ± 0.00001	0.52974 ± 0.14454
Transport + match	0.335	0.0023 ± 0.0001	0.00037 ± 0.00005	0.14961 ± 0.00916	0.19391 ± 0.02546
T. + full consistency	**0.058**	**0.0038 ± 0.0004**	**0.00005 ± 0.00002**	**0.00032 ± 0.00017**	**0.00952 ± 0.00351**
Transport pred. only	0.869	0.0128 ± 0.0023	0.00099 ± 0.00015	0.27173 ± 0.03082	0.25672 ± 0.04538

Geometry winner (Transport + full consistency) is shown in bold.

Table 15: **Transport geometry diagnostics.** Mean ± standard deviation across confirmation seeds for embedding-match, identity, semigroup, and composite geometry metrics.

Figure 29 summarizes the scope of the study: the encoders are fixed, and only the transport module and its associated objectives are varied.

Conclusion

We carry forward **Transport + full consistency** as the default transport formulation for subsequent experiments. This configuration gives the best overall predictive model and the strongest temporal geometry, even though **Transport + match + identity** yields the best explicit rollout path. The transport ablation therefore shows that transport is useful in TerraNova, but only when constrained by additional latent-consistency objectives rather than by predictive supervision alone.

*The chosen configuration is **Transport + full consistency**.*

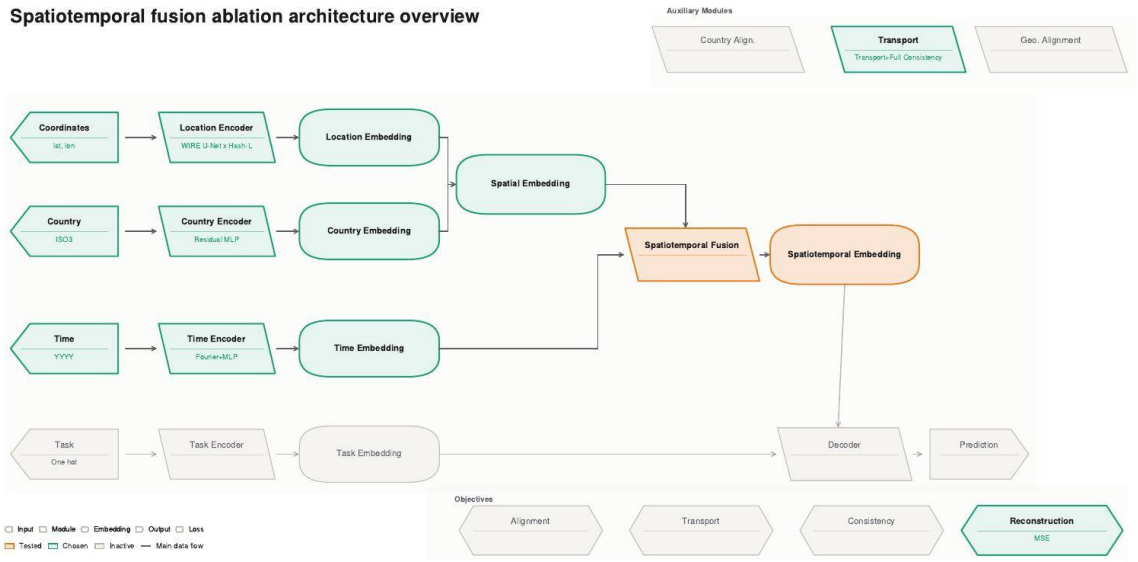

Figure 31: **Spatiotemporal fusion module ablation setup.** Coordinate, country, and time inputs, their encoders, and their embeddings are fixed to the selected upstream configurations. The transport module and reconstruction objective remain active. The tested component is the spatiotemporal fusion pathway, mapping the fixed spatial and temporal embeddings into the 128-dimensional spatiotemporal embedding consumed by the decoder.

B.1.14 Spatiotemporal fusion

Motivation

The location, country, time, and transport ablations identify strong individual representation modules, but TerraNova still requires a mechanism that combines *where* a sample is observed with *when* it is observed. This fusion block is the first point in the architecture where the spatial representation

$$s \in \mathbb{R}^{128}$$

and the temporal representation

$$\tau \in \mathbb{R}^{8}$$

interact before the shared decoder. Our design question in this experiment is therefore how temporal information should condition spatial structure.
This ablation isolates the spatiotemporal fusion block

$$F_\theta : \mathbb{R}^{128} \times \mathbb{R}^{8} \to \mathbb{R}^{128}, \qquad z = F_\theta(s, \tau),$$

while keeping the upstream encoders and downstream decoder fixed (Fig. 31). The spatial input is either the winner of the location-encoder ablation for gridded samples, or the winner of the country-encoder ablation projected onto the same 128-dimensional spatial bus for country samples. The temporal input is the Fourier+MLP time encoder, and the transport module is kept at the selected transport-ablation configuration. Only the fusion map F_θ is varied.

The benchmark deliberately combines both native geometries of TerraNova. The gridded side uses the same 51 WorldTensor fields as the location-encoder study. Time-varying gridded variables are evaluated on held-out years using year-specific targets; static gridded variables are queried at random held-out years with invariant targets. The country side uses 150 temporal country-level variables, with five randomly held-out years per variable. Evaluation is thus held-out-time generalization for both modalities, not a spatial extrapolation benchmark.

B.1.15 Fusion architecture zoo

All variants share the same output dimension 128 and are evaluated at hidden widths $w \in \{256, 512\}$ and depths $d \in \{4, 8\}$. Let $M_{w,d}(\cdot)$ denote a residual MLP with hidden width w and d scaled residual blocks. Unless stated otherwise, the final output of each fusion module is $z \in \mathbb{R}^{128}$.

Spatial-only baseline. The lower-bound model ignores time completely:

$$z = M_{w,d}(s). \tag{78}$$

This baseline tests whether the fixed encoders and decoder can already explain the held-out data without explicit temporal conditioning. Any useful fusion module should improve over this reference, particularly on country variables whose targets change substantially across years.

Concat + MLP. The simplest time-aware baseline concatenates the two embeddings:

$$z = M_{w,d}\big([s; \tau]\big). \tag{79}$$

This option gives the MLP full access to both modalities but imposes no explicit inductive bias about how time should act on space. It is the generic nonlinear-fusion baseline.

FiLM family. Feature-wise linear modulation Perez et al. (2018) treats time as a conditioner that rescales and shifts the spatial representation. With

$$\gamma(\tau) = A_\gamma \tau, \qquad \beta(\tau) = A_\beta \tau,$$

the standard FiLM variant is

$$z = M_{w,d}\big(s \odot (1 + \gamma(\tau)) + \beta(\tau)\big). \tag{80}$$

This tests the hypothesis that year mainly changes the amplitude and offset of spatial factors. We also test two restricted variants:

$$z_{\text{scale}} = M_{w,d}\big(s \odot (1 + \gamma(\tau))\big), \tag{81}$$
$$z_{\text{resid}} = M_{w,d}\big(s + s \odot \gamma(\tau) + \beta(\tau)\big). \tag{82}$$

The scale-only form removes additive temporal shifts, while the residual form keeps the original spatial state explicit and learns a time-dependent delta.

Gated FiLM. The gated FiLM variant learns a per-feature interpolation between the identity spatial representation and the FiLM-modulated representation:

$$\tilde{s} = s \odot (1 + \gamma(\tau)) + \beta(\tau), \tag{83}$$
$$g(\tau) = \sigma(A_g \tau), \tag{84}$$
$$z = M_{w,d}\big(g(\tau) \odot \tilde{s} + (1 - g(\tau)) \odot s\big). \tag{85}$$

The motivation is to avoid forcing every spatial feature to be temporal: some coordinates should remain almost static, while others should respond strongly to year.

Residual gate. Instead of modulating the spatial vector multiplicatively, residual gating adds a learned temporal displacement:

$$g(\tau) = \sigma(A_g \tau), \qquad u(\tau) = A_u \tau, \tag{86}$$
$$z = M_{w,d}\big(s + g(\tau) \odot u(\tau)\big). \tag{87}$$

This is a compact temporal-update model: the spatial embedding remains the anchor, and time contributes only through a gated residual vector.

Cross-attention: spatial queries time. The first attention family uses cross-attention Hou et al. (2019) to let the spatial token query the temporal token:

$$q = W_q s, \qquad k = W_k \tau, \qquad v = W_v \tau, \tag{88}$$
$$h = \mathrm{LN}\big(q + \mathrm{Attn}(q, k, v)\big), \tag{89}$$
$$z = W_o\, B_{d-1}(h), \tag{90}$$

where B_{d-1} is a stack of residual blocks. This direction, denoted S→T in the figures, asks: which temporal information is relevant for a given spatial or country state?

Cross-attention: time queries space. The reverse direction lets the time token query the spatial token and then recombines the attended context with the original spatial embedding:

$$q = W_q \tau, \qquad k = W_k s, \qquad v = W_v s, \tag{91}$$
$$c = \mathrm{LN}\big(\mathrm{Attn}(q, k, v)\big), \tag{92}$$
$$z = W_o\, B_{d-1}\big(W_c[c; s]\big). \tag{93}$$

This tests the opposite asymmetry: temporal state first extracts a spatial context, and the model then reconstructs the fused representation from that context plus the original spatial vector.

Joint self-attention. The symmetric attention baseline treats spatial and temporal embeddings as a two-token set:

$$x_1 = W_s s, \qquad x_2 = W_t \tau, \tag{94}$$
$$H = \mathrm{Attn}\big([x_1, x_2], [x_1, x_2], [x_1, x_2]\big), \tag{95}$$
$$h = \mathrm{LN}\left(\frac{1}{2}\sum_{i=1}^{2}(x_i + H_i)\right), \tag{96}$$
$$z = W_o\, B_{d-1}(h). \tag{97}$$

Unlike directional cross-attention, this option gives neither modality a privileged query role.

Transformer Tr-A: stacked cross-modal decoder. The first transformer Vaswani et al. (2017) family repeats the spatial-query cross-attention pattern with a feed-forward network after each attention layer:

$$q_0 = W_s s, \qquad m = W_t \tau, \tag{98}$$
$$\hat{q}_\ell = \mathrm{LN}\big(q_\ell + \mathrm{Attn}(q_\ell, m, m)\big), \tag{99}$$
$$q_{\ell+1} = \mathrm{LN}\big(\hat{q}_\ell + \mathrm{FFN}(\hat{q}_\ell)\big), \qquad \ell = 0, \ldots, d-1, \tag{100}$$
$$z = W_o q_d. \tag{101}$$

This is the highest-capacity version of the S→T inductive bias: spatial state remains the query throughout, while time is repeatedly injected through cross-attention and nonlinear refinement.

Transformer Tr-B: encoder-decoder fusion. The second transformer family first processes the time token as memory and then uses a spatial query decoder:

$$m_0 = W_t \tau, \qquad q_0 = W_s s, \tag{102}$$
$$m_L = \mathrm{TransformerEncoder}(m_0), \tag{103}$$
$$q_{\ell+1} = \mathrm{DecoderBlock}(q_\ell, m_L), \tag{104}$$
$$z = W_o q_D. \tag{105}$$

This option separates temporal processing from spatial decoding. It is more structured than Tr-A, but also less direct: time is first transformed into a memory token before it conditions the spatial query.

B.1.16 Training and evaluation protocol

The active sweep contains 12 architectures × 2 widths × 2 depths = 48 configurations. Phase 1 screens every configuration at seed 42 over the log-spaced learning-rate ladder

$$\{10^{-3}, 10^{-3.5}, 10^{-4}, 10^{-4.5}, 10^{-5}, 10^{-5.5}, 10^{-6}\},$$

using a flat learning rate for 15,000 steps. Phase 2 confirms every screened configuration, not only the top few, using that configuration's best Phase-1 learning rate, three seeds $\{42, 137, 2024\}$, a warmup-cosine schedule, and 45,000 steps.

Training alternates stochastic minibatches from the two modalities. Gridded batches sample flat tuples $(\text{field}, \text{pixel}, \text{year})$; country batches sample flat tuples $(\text{country}, \text{year})$ with a per-variable observation mask. Both modalities are z-score normalised using training data only, so all reported losses are in standardised units. The training loss is the sum of the modality-specific reconstruction loss and the fixed transport-consistency regularisation inherited from the transport ablation. For country batches, the selected predictive transport term is also active.

The primary metric is held-out-year RMSE averaged over all evaluated variables:

$$\overline{\mathrm{RMSE}} = \frac{1}{|\mathcal{G}| + |\mathcal{C}|} \left[\sum_{g \in \mathcal{G}} \mathrm{RMSE}_g + \sum_{c \in \mathcal{C}} \mathrm{RMSE}_c \right], \tag{106}$$

where $\mathcal{G}$ is the set of gridded fields with held-out-year evaluation and $\mathcal{C}$ is the set of country variables with held-out observations. We also report modality-specific RMSE and Pearson correlation:

$$\overline{\mathrm{RMSE}}_{\mathrm{grid}}, \quad \overline{\mathrm{RMSE}}_{\mathrm{country}}, \quad \bar{r}, \quad \bar{r}_{\mathrm{grid}}, \quad \bar{r}_{\mathrm{country}}.$$

B.1.17 Results

Headline. The best fusion module is the $w = 512$, $d = 8$ cross-modal transformer (Tr-A), with

$$\overline{\mathrm{RMSE}} = 0.1655 \pm 0.0007,$$

an absolute reduction of 0.1818 RMSE units relative to the spatial-only $w = 256, d = 4$ baseline (0.3473 ± 0.0001), or a 52.3% relative reduction. The winner also has strong modality-specific performance:

$$\overline{\mathrm{RMSE}}_{\mathrm{grid}} = 0.2538, \qquad \overline{\mathrm{RMSE}}_{\mathrm{country}} = 0.1326,$$

with correlations

$$\bar{r} = 0.954, \qquad \bar{r}_{\mathrm{grid}} = 0.877, \qquad \bar{r}_{\mathrm{country}} = 0.983.$$

The runner-up is the $w = 512, d = 8$ encoder-decoder transformer (Tr-B), at 0.1730 ± 0.0007 RMSE. The top five configurations are all transformers, showing that the decisive factor is not merely adding time but allowing repeated cross-modal interaction at sufficient width and depth.

Efficiency Pareto. Figure 32 places all confirmation variants in the parameter–error plane. The frontier begins with compact non-transformer and attention variants around 56–61M parameters, but the low-error end is entirely transformer-dominated. The selected Tr-A $w512, d8$ model is the rightmost useful point before further parameter increases stop improving accuracy: Tr-B $w512, d8$ uses more parameters (85.2M) but has higher RMSE (0.1730). Thus the winning module is not only the lowest-error model but also the efficient high-capacity endpoint of the sweep.

Family-level behaviour. Figure 33 summarizes the distribution of confirmation RMSE within each architecture family. The transformer family is separated from the rest of the sweep: its best variants occupy the lowest-error regime, and even its weaker configurations remain competitive with the best attention models. Directional and joint attention form the next tier, but with a much wider spread because the T→S direction collapses to the spatial-only regime. MLP, FiLM, and gated residual variants improve substantially over the spatial-only baseline, but they cluster near 0.21–0.22 RMSE and do not approach the transformer frontier.

Capacity effects. Figure 34 exposes a strong interaction between architecture and capacity. For the transformer families, increasing width and depth is consistently useful: Tr-A improves from 0.1988 at $w256, d4$ to 0.1655 at $w512, d8$, and Tr-B improves from 0.1968 at $w256, d4$ to 0.1730 at $w512, d8$. Attention variants also benefit from larger width and depth in the S→T direction, but saturate above the transformer error floor. By contrast, FiLM and MLP variants show little gain from extra capacity, indicating that their limitation is the fusion mechanism rather than the number of parameters.

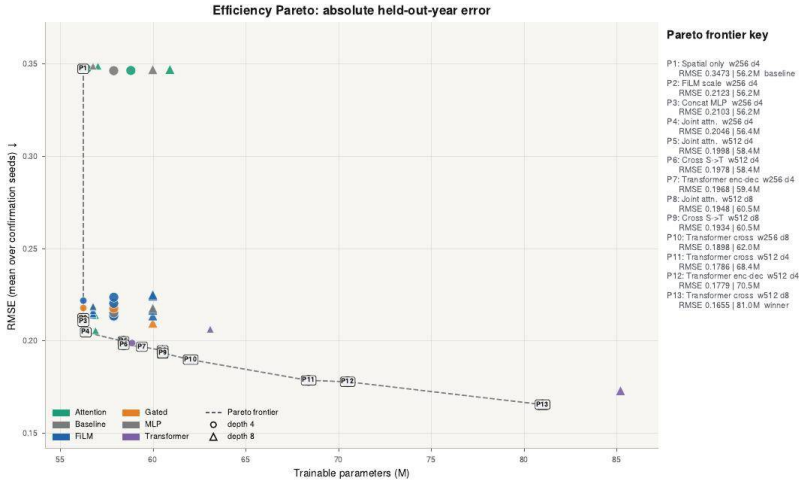

Figure 32: **Efficiency Pareto for spatiotemporal fusion.** Confirmation RMSE versus trainable parameter count for all 48 variants. Pareto-frontier points are labelled in the side key. The selected model is Transformer Tr-A with width 512 and depth 8.

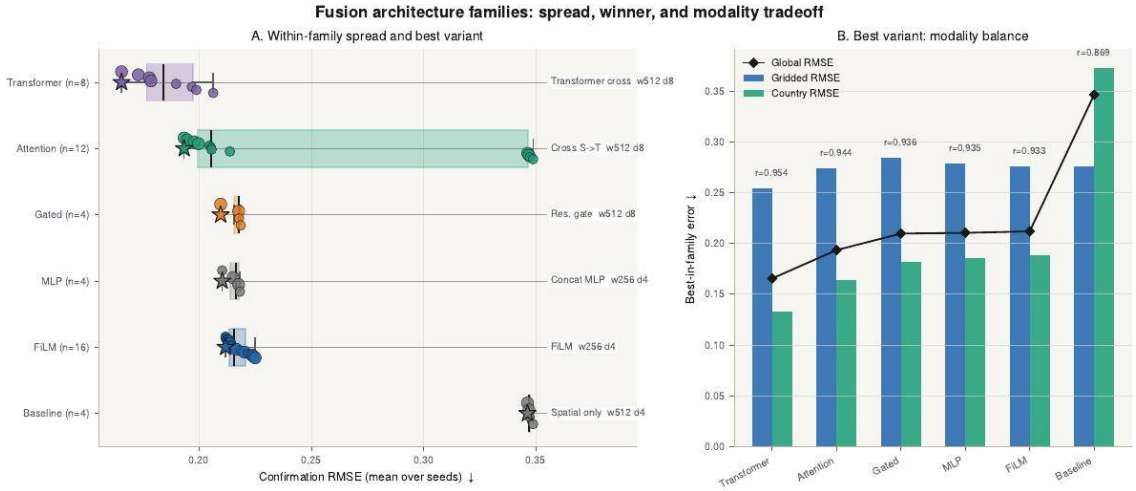

Figure 33: **Fusion family summary.** Left: RMSE distribution within each family, with the best-in-family configuration marked. Right: modality-specific RMSE of the best member of each family. Transformer fusion wins both globally and on the country side, while preserving competitive gridded performance.

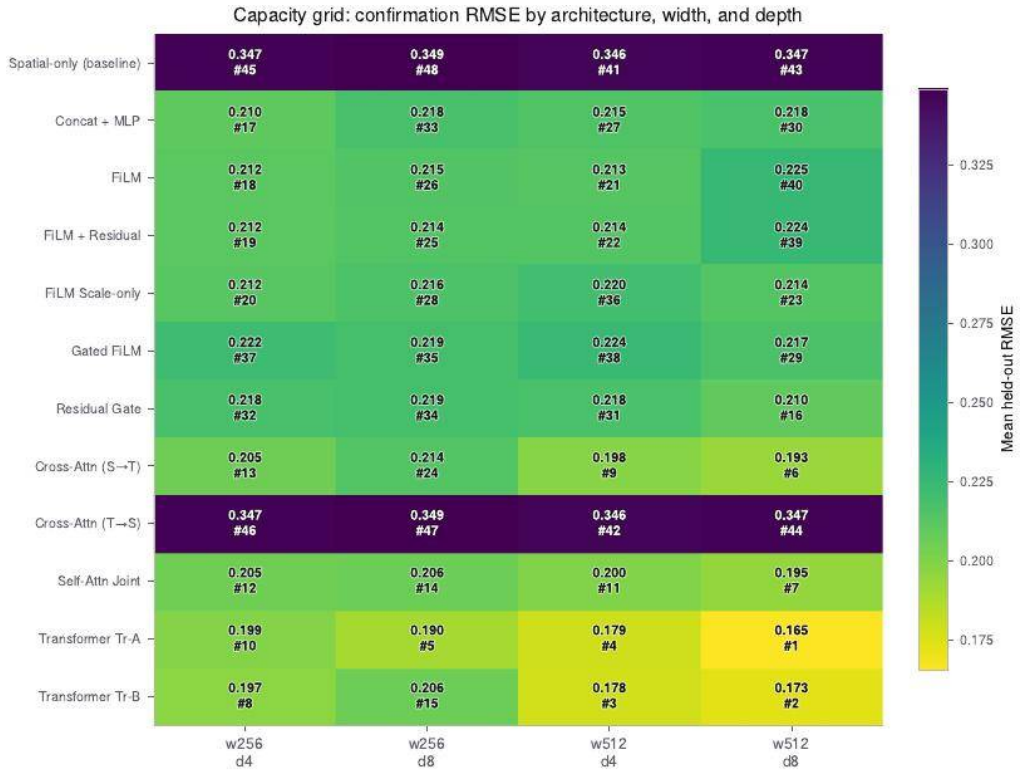

Figure 34: **Capacity grid.** Confirmation RMSE by architecture, hidden width, and depth. Each cell shows mean RMSE and leaderboard rank. Transformer capacity scales cleanly; FiLM, MLP, and gated variants do not close the gap by increasing width or depth.

Modality trade-off. Figure 35 compares gridded and country RMSE. The spatial-only and T→S attention variants have acceptable gridded RMSE but very poor country RMSE, showing that they preserve spatial reconstruction while failing to learn temporal country dynamics. The selected Tr-A model moves sharply down the country-error axis while retaining the best overall balance. This is the most important qualitative outcome of the ablation: the winning module does not win by overfitting one modality, but by improving the harder temporal-country side without destroying gridded performance.

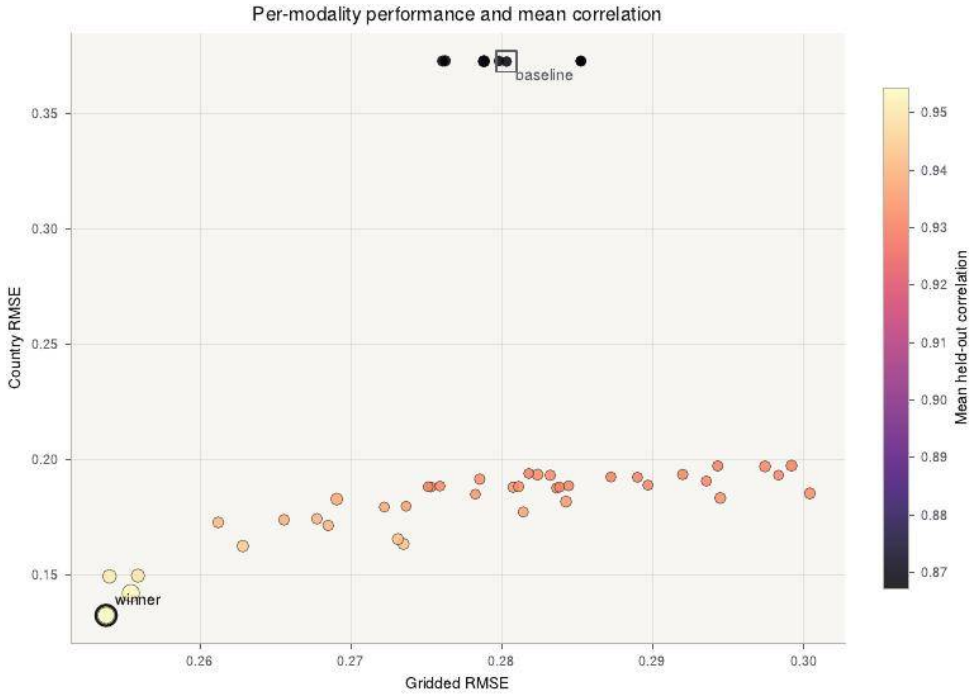

Figure 35: **Per-modality trade-off.** Gridded RMSE versus country RMSE for all confirmation variants. Point colour encodes mean held-out correlation. The winning transformer occupies the best global trade-off region, with low country error and competitive gridded error.

Confirmation table. Table 16 reports the full confirmation leaderboard.

B.1.18 Discussion

The ablation supports three conclusions. First, explicit spatiotemporal fusion is necessary: the spatial-only baseline reconstructs static and weakly temporal structure but fails on held-out country years, yielding a global RMSE around 0.347. Second, the direction of interaction matters. S→T attention is consistently useful, while T→S attention is nearly indistinguishable from the spatial-only baseline. The successful pattern is to keep the spatial or country state as the query and inject temporal information as conditioning memory, not to let the time token query a spatial memory and then reconstruct the state afterward.
Third, the best fusion mechanism is a repeated cross-modal transformer rather than a single modulation operation. FiLM, gated FiLM, residual gates, and concat-MLPs all provide

Table 16: Fusion confirmation leaderboard. Held-out-year RMSE is reported as mean ± s.d. across 3 confirmation seeds. Δ is the percentage change relative to `spatial_only_w256_d4` (negative is better). Grid and Ctry are modality-specific held-out-year RMSE values; r is the mean Pearson correlation across gridded and country variables. Winner **bold** and highlighted.

Variant	Family	Param.	RMSE↓	Δ%	Grid↓	Ctry↓	r ↑
Transformer Tr-A (w512, d8)	Transformer	81.0M	**0.1655 ± 0.0007**	-52.3	0.2538	0.1326	0.954
Transformer Tr-B (w512, d8)	Transformer	85.2M	0.1730 ± 0.0007	-50.2	0.2554	0.1423	0.953
Transformer Tr-B (w512, d4)	Transformer	70.5M	0.1779 ± 0.0006	-48.8	0.2540	0.1495	0.951
Transformer Tr-A (w512, d4)	Transformer	68.4M	0.1786 ± 0.0016	-48.6	0.2559	0.1498	0.949
Transformer Tr-A (w256, d8)	Transformer	62.0M	0.1898 ± 0.0004	-45.3	0.2629	0.1626	0.944
Cross-Attn (S→T) (w512, d8)	Attention	60.5M	0.1934 ± 0.0033	-44.3	0.2735	0.1636	0.944
Self-Attn Joint (w512, d8)	Attention	60.5M	0.1948 ± 0.0013	-43.9	0.2731	0.1656	0.943
Transformer Tr-B (w256, d4)	Transformer	59.4M	0.1968 ± 0.0014	-43.3	0.2612	0.1729	0.941
Cross-Attn (S→T) (w512, d4)	Attention	58.4M	0.1978 ± 0.0006	-43.0	0.2685	0.1715	0.941
Transformer Tr-A (w256, d4)	Transformer	58.9M	0.1988 ± 0.0008	-42.7	0.2656	0.1740	0.940
Self-Attn Joint (w512, d4)	Attention	58.4M	0.1998 ± 0.0014	-42.5	0.2678	0.1744	0.940
Self-Attn Joint (w256, d4)	Attention	56.4M	0.2046 ± 0.0011	-41.1	0.2722	0.1795	0.938
Cross-Attn (S→T) (w256, d4)	Attention	56.4M	0.2053 ± 0.0008	-40.9	0.2737	0.1799	0.937
Self-Attn Joint (w256, d8)	Attention	56.9M	0.2056 ± 0.0010	-40.8	0.2814	0.1774	0.937
Transformer Tr-B (w256, d8)	Transformer	63.1M	0.2063 ± 0.0008	-40.6	0.2691	0.1829	0.938
Residual Gate (w512, d8)	Gated	60.0M	0.2096 ± 0.0004	-39.6	0.2843	0.1819	0.936
Concat + MLP (w256, d4)	MLP	56.2M	0.2103 ± 0.0007	-39.4	0.2783	0.1851	0.935
FiLM (w256, d4)	FiLM	56.2M	0.2118 ± 0.0007	-39.0	0.2753	0.1882	0.933
FiLM + Residual (w256, d4)	FiLM	56.2M	0.2118 ± 0.0001	-39.0	0.2751	0.1883	0.933
FiLM Scale-only (w256, d4)	FiLM	56.2M	0.2123 ± 0.0008	-38.9	0.2759	0.1886	0.932
FiLM (w512, d4)	FiLM	57.9M	0.2132 ± 0.0018	-38.6	0.2808	0.1881	0.934
FiLM + Residual (w512, d4)	FiLM	57.9M	0.2135 ± 0.0016	-38.5	0.2811	0.1884	0.934
FiLM Scale-only (w512, d8)	FiLM	60.0M	0.2135 ± 0.0026	-38.5	0.2945	0.1834	0.934
Cross-Attn (S→T) (w256, d8)	Attention	56.9M	0.2137 ± 0.0007	-38.5	0.2836	0.1877	0.933
FiLM + Residual (w256, d8)	FiLM	56.8M	0.2141 ± 0.0008	-38.4	0.2838	0.1881	0.932
FiLM (w256, d8)	FiLM	56.8M	0.2147 ± 0.0009	-38.2	0.2844	0.1887	0.932
Concat + MLP (w512, d4)	MLP	57.9M	0.2152 ± 0.0013	-38.0	0.2785	0.1916	0.933
FiLM Scale-only (w256, d8)	FiLM	56.8M	0.2163 ± 0.0009	-37.7	0.2897	0.1890	0.932
Gated FiLM (w512, d8)	FiLM	60.0M	0.2166 ± 0.0015	-37.6	0.3004	0.1854	0.933
Concat + MLP (w512, d8)	MLP	60.0M	0.2176 ± 0.0004	-37.3	0.2824	0.1935	0.934
Residual Gate (w512, d4)	Gated	57.9M	0.2177 ± 0.0014	-37.3	0.2832	0.1933	0.933
Residual Gate (w256, d4)	Gated	56.2M	0.2178 ± 0.0008	-37.3	0.2818	0.1940	0.931
Concat + MLP (w256, d8)	MLP	56.8M	0.2182 ± 0.0017	-37.2	0.2872	0.1925	0.931
Residual Gate (w256, d8)	Gated	56.8M	0.2186 ± 0.0008	-37.1	0.2890	0.1924	0.931
Gated FiLM (w256, d8)	FiLM	56.8M	0.2186 ± 0.0015	-37.0	0.2936	0.1907	0.932
FiLM Scale-only (w512, d4)	FiLM	57.9M	0.2203 ± 0.0038	-36.6	0.2920	0.1936	0.932
Gated FiLM (w256, d4)	FiLM	56.2M	0.2218 ± 0.0012	-36.1	0.2983	0.1933	0.932
Gated FiLM (w512, d4)	FiLM	57.9M	0.2236 ± 0.0015	-35.6	0.2943	0.1973	0.931
FiLM + Residual (w512, d8)	FiLM	60.0M	0.2243 ± 0.0013	-35.4	0.2974	0.1971	0.932
FiLM (w512, d8)	FiLM	60.0M	0.2250 ± 0.0013	-35.2	0.2992	0.1974	0.931
Spatial-only (baseline) (w512, d4)	Baseline	57.9M	0.3463 ± 0.0001	-0.3	0.2761	0.3724	0.869
Cross-Attn (T→S) (w512, d4)	Attention	58.8M	0.3464 ± 0.0002	-0.3	0.2763	0.3725	0.869
Spatial-only (baseline) (w512, d8)	Baseline	60.0M	0.3469 ± 0.0002	-0.1	0.2788	0.3723	0.868
Cross-Attn (T→S) (w512, d8)	Attention	60.9M	0.3470 ± 0.0001	-0.1	0.2788	0.3724	0.868
Spatial-only (baseline) (w256, d4)	Baseline	56.2M	0.3473 ± 0.0001	0.0	0.2803	0.3722	0.868
Cross-Attn (T→S) (w256, d4)	Attention	56.5M	0.3473 ± 0.0001	0.0	0.2798	0.3724	0.868
Cross-Attn (T→S) (w256, d8)	Attention	57.0M	0.3488 ± 0.0002	0.4	0.2852	0.3724	0.867
Spatial-only (baseline) (w256, d8)	Baseline	56.8M	0.3488 ± 0.0001	0.4	0.2853	0.3724	0.867

useful temporal conditioning, but they behave like single-step corrections to the spatial state and saturate near 0.21–0.22 RMSE. The transformer variants instead alternate cross-modal attention and nonlinear refinement, allowing temporal information to be integrated repeatedly before decoding. This extra interaction depth is what separates the top transformer configurations from simpler conditioning mechanisms. We therefore select Transformer Tr-A with width 512 and depth 8 as the TerraNova spatiotemporal fusion module. It is the best

global model, the efficient high-capacity endpoint of the Pareto frontier, and the strongest modality-balanced solution across gridded and country held-out-year generalization.
The chosen configuration is a ***Cross-Attention Transformer decoder with width*** 512 ***and depth*** 8*.*

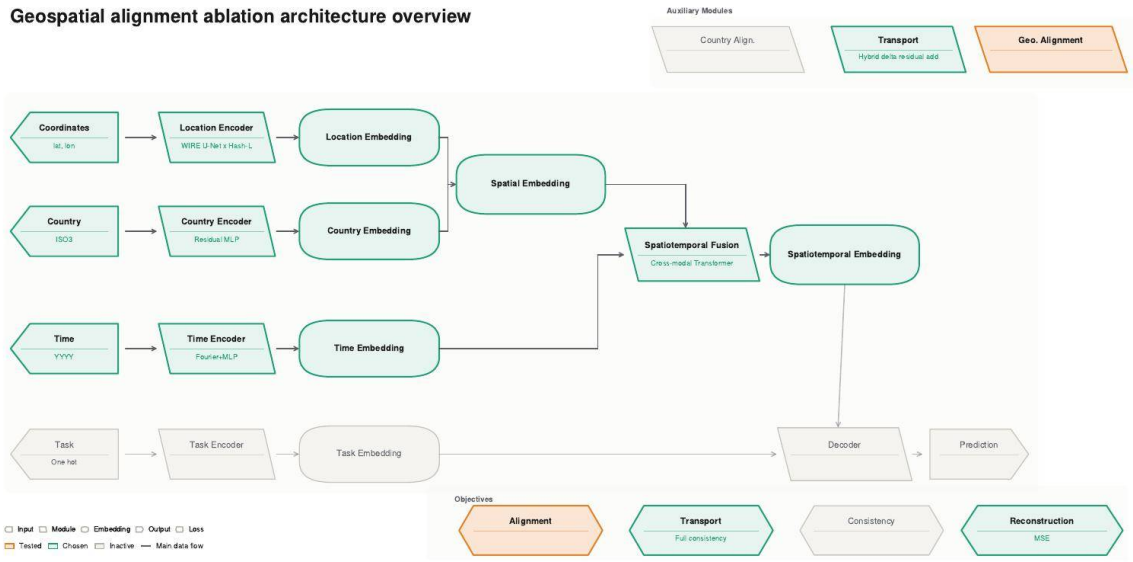

Figure 36: **Geospatial-alignment ablation scope.** The upstream location, country, time, transport, and spatiotemporal-fusion modules are fixed to the selected configurations. The tested component is the geospatial-alignment auxiliary branch and its contrastive alignment loss.

B.1.19 Alignment with geospatial embeddings

Motivation and scope

The previous ablations select the coordinate, country, time, transport, and spatiotemporal-fusion modules used by the TerraNova backbone. This ablation instead tests a different question: whether the learned spatiotemporal representation can be explicitly aligned to external geospatial embedding spaces while preserving reconstruction accuracy of the multitask model. The ablated component is an auxiliary contrastive supervision family applied to the shared spatiotemporal embedding.

The backbone is fixed to the selected upstream configuration (Fig. 36): the location encoder is the WIRE U-Net × Hash-L model, the country encoder is the residual MLP country encoder, the time encoder is Fourier+MLP, the transport module is the selected full-consistency transport configuration, and the spatiotemporal fusion block is the selected cross-modal Transformer. Only the geospatial-alignment objective and its sampling budget are varied. The alignment family has three equally weighted targets: SatCLIP, GeoCLIP, and Copernicus-embed. Each target provides contrastive auxiliary supervision over coordinates; none is treated as a scalar prediction task.

Protocol

Training follows a randomized micro-epoch structure used by the final multitask training loop. Each epoch contains reconstruction steps, transport steps, and geospatial-alignment steps in randomized order. The reconstruction objective remains the standard regression objective for gridded and country-level variables. The transport budget is fixed at $M = 512$ transport steps per epoch, using the selected transport configuration. The geospatial-

alignment budget is varied over $G \in \{0, 128, 512\}$, where $G = 0$ is the no-alignment baseline. For $G > 0$, alignment batches are sampled from the three target embedding families with equal probability: $p(\text{SatCLIP}) = p(\text{GeoCLIP}) = p(\text{Copernicus}) = \frac{1}{3}$.
The screening phase sweeps the alignment loss family and alignment loss weight at seed 42. The tested loss families are inverse-distance weighting and linear-distance weighting; these choices only apply when $G > 0$. The no-alignment baseline is therefore reported simply as $G = 0$, not as a member of either distance-weighting family. Screening uses the same selection score as the training callback:$S = \text{Top1}_{\text{align}} - \lambda_{\text{rec}} \, \text{RMSE}_{\text{val}}$, so configurations must improve geospatial retrieval without collapsing held-out-year reconstruction. Confirmation reruns the best screened structural configurations across three seeds, with 3 times more training epochs and decaying learning rates.

Variant	Loss	G	Weight	Score↑	Top-1↑	RMSE↓	Grid r ↑	Ctry r ↑
Inverse M=512 G=512 w1	inverse	512	1	0.636	0.705	0.697	0.724	0.611
Inverse M=512 G=128 w1	inverse	128	1	0.424	0.490	0.663	0.767	0.623
Inverse M=512 G=512 w0.1	inverse	512	0.1	0.324	0.395	0.714	0.710	0.595
Inverse M=512 G=512 w0.05	inverse	512	0.05	0.271	0.342	0.705	0.718	0.607
Inverse M=512 G=512 w0.01	inverse	512	0.01	0.184	0.255	0.706	0.712	0.603
Inverse M=512 G=128 w0.1	inverse	128	0.1	0.183	0.250	0.666	0.768	0.627
Inverse M=512 G=128 w0.05	inverse	128	0.05	0.163	0.230	0.665	0.767	0.623
Inverse M=512 G=128 w0.01	inverse	128	0.01	0.129	0.195	0.656	0.771	0.637
Linear M=512 G=512 w1	linear	512	1	0.020	0.091	0.713	0.695	0.603
Linear M=512 G=512 w0.05	linear	512	0.05	0.017	0.087	0.693	0.722	0.625
Linear M=512 G=512 w0.01	linear	512	0.01	0.016	0.085	0.687	0.724	0.629
Linear M=512 G=512 w0.1	linear	512	0.1	0.006	0.077	0.710	0.717	0.612
Linear M=512 G=128 w0.05	linear	128	0.05	0.005	0.071	0.661	0.773	0.633
Linear M=512 G=128 w0.01	linear	128	0.01	0.003	0.068	0.646	0.763	0.640
Linear M=512 G=128 w0.1	linear	128	0.1	-0.005	0.061	0.657	0.768	0.635
Linear M=512 G=128 w1	linear	128	1	-0.013	0.053	0.665	0.754	0.633
No alignment M=512 G=0	–	0	0	-0.061	0.005	0.657	0.760	0.639

Phase 1 screening: fixed ST-fusion winner backbone, fixed LR 10^{-4}, M=512 transport steps per epoch, seed 42.

Table 17: **Screening leaderboard for the geospatial-alignment study.** The table reports the alignment/reconstruction selection score, alignment top-1 retrieval, validation RMSE, and gridded/country held-out-year correlations. The $G = 0$ row is the no-alignment baseline and is not assigned to a distance-weighting family.

B.1.20 Results

The confirmed winner is the inverse-distance alignment loss with $M = 512$, $G = 512$, and alignment weight 1. It reaches an alignment top-1 of 0.8753±0.0015, compared with

0.0076±0.0024 for the no-alignment baseline, a gain of +0.8677. The corresponding selection score is 0.8329±0.0020.
The main reconstruction-preservation diagnostic is gridded held-out-year correlation, because it isolates the WorldTensor side of the benchmark where coordinate-level spatial structure is most directly tested. On this metric the winner obtains 0.8946±0.0395, compared with 0.8428±0.0258 for the no-alignment baseline ($\Delta = +0.0517$). Country-level held-out-year correlation is 0.7841±0.0108, compared with 0.8192±0.0258 for the no-alignment baseline ($\Delta = -0.0352$).

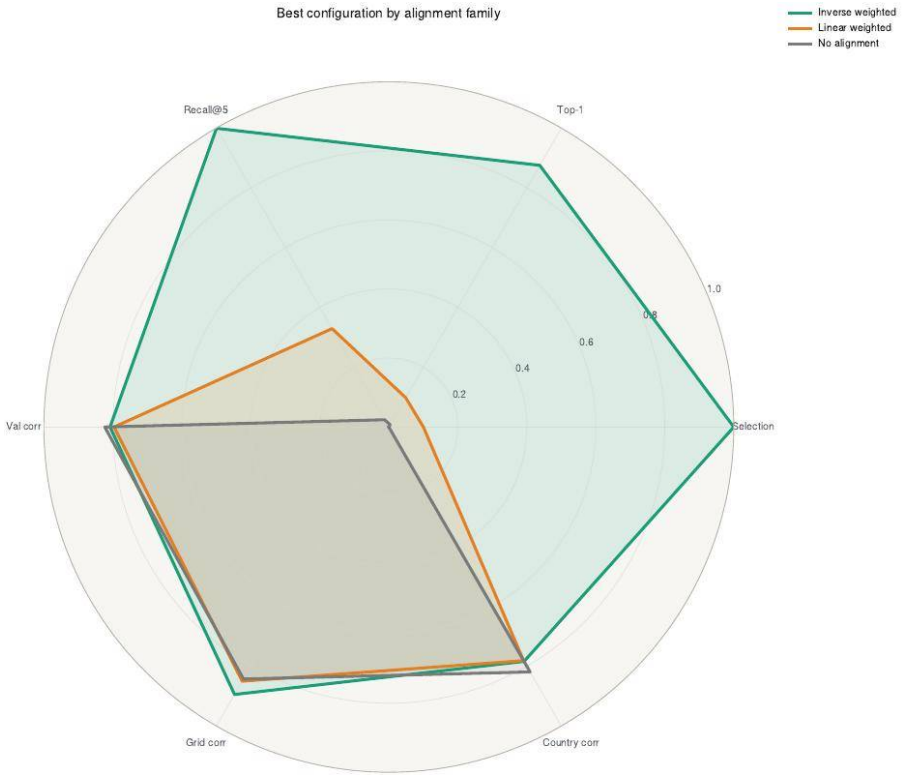

Figure 37: **Best configuration by alignment family.** The radar plot compares the best no-alignment, inverse-distance, and linear-distance configurations across alignment retrieval and reconstruction-preservation metrics. The inverse-distance family dominates the alignment axes while retaining comparable validation correlations.

Overall, the study supports using geospatial embedding alignment as an auxiliary representation objective rather than as a prediction task. The alignment branch substantially improves retrieval into the three external geospatial embedding spaces. Without losing significant reconstruction accuracy, we can align our spatiotemporal embeddings with existing geospatial embeddings, thereby providing our model with information from the visual domain, both from satellite images (Copernicus-embed and SatCLIP) and with images taken from the ground (GeoCLIP). In terms of correlation on gridded-data prediction, this alignment produces a significant increase of $\Delta \geq 0.05$, which suggests that it allows our embedding to be regularised and better suited for spatiotemporal generalisation. Furthermore, both an

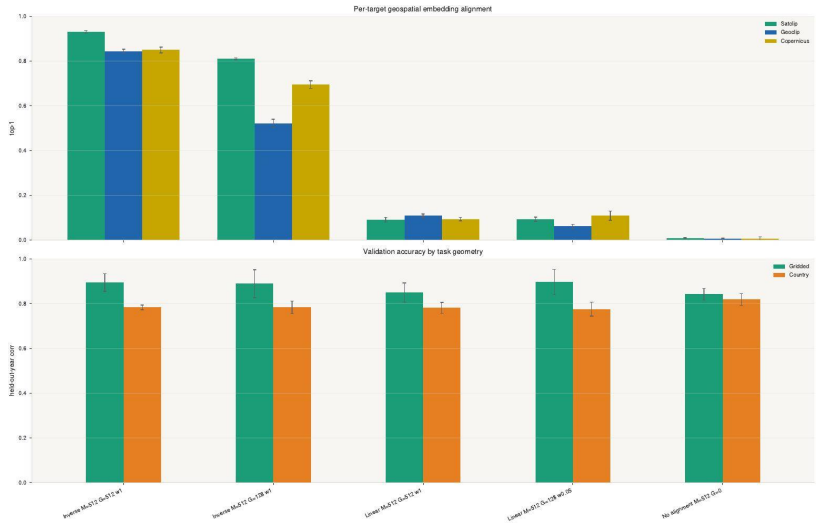

Figure 38: **Target-level alignment and validation accuracy.** Top: SatCLIP, GeoCLIP, and Copernicus top-1 retrieval with seed error bars. Bottom: gridded and country held-out-year correlations with seed error bars. The winning configuration improves all three external embedding targets simultaneously, with the clearest gain on SatCLIP and strong gains on GeoCLIP and Copernicus.

Variant	Family	G	Weight	Score↑	Top-1↑	ΔTop-1	RMSE↓	Corr↑	Grid r ↑	Ctry r ↑	Vis. errs
Inverse M=512 G=512 w1	Inverse weighted	512	1	**0.8329 ± 0.0020**	**0.8753 ± 0.0015**	+0.8677	0.4244	0.8082	0.8946	0.7841	3
Inverse M=512 G=128 w1	Inverse weighted	128	1	0.6343 ± 0.0006	0.6756 ± 0.0023	+0.6681	0.4137	0.8074	0.8899	0.7838	3
Linear M=512 G=512 w1	Linear weighted	512	1	0.0533 ± 0.0054	0.0980 ± 0.0050	+0.0904	0.4471	0.7965	0.8502	0.7814	3
Linear M=512 G=128 w0.05	Linear weighted	128	0.05	0.0454 ± 0.0109	0.0883 ± 0.0097	+0.0807	0.4283	0.8021	0.8968	0.7759	3
No alignment M=512 G=0	No alignment	0	0	-0.0335 ± 0.0038	0.0076 ± 0.0024	+0.0000	0.4109	0.8236	0.8428	0.8192	3

Phase 2 confirmation: fixed ST-fusion winner backbone, M=512 transport steps per epoch, fixed LR 10^{-4}, three seeds. Visualization errors were non-fatal and did not remove saved validation/retrieval metrics.

Table 18: **Confirmation leaderboard for the geospatial-alignment study.** The table reports the alignment/reconstruction selection score, alignment top-1 retrieval, validation RMSE, and gridded/country held-out-year correlations. The $G = 0$ row is the no-alignment baseline and is not assigned to a distance-weighting family.

increased number of steps per epoch and a strong alignment-loss weight provide the best configuration. Country-level tasks see a modest reduction in accuracy as these alignment losses are not designed for helping in country-level tasks. The purpose of the next ablation study will therefore be to align the country and location embeddings so they can benefit one another.

The chosen configuration is ***inverse-distance location sampling with a loss weight of 1, at the largest alignment-step budget tested.*** The production run allocates 256 geospatial-alignment steps per epoch (the ablation's largest budget was 512), reflecting the fixed per-epoch step budget shared across all step families.

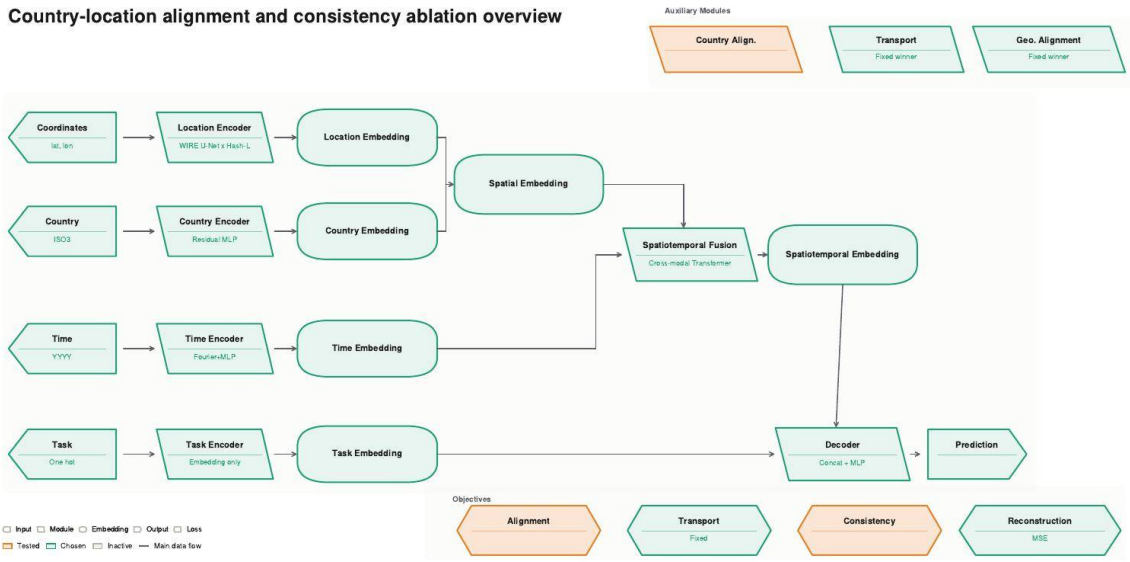

Figure 39: **Country-location ablation scope.** The backbone and external geospatial-alignment branch are fixed. Orange components denote the internal country-location objectives tested in this ablation. The task encoder is an embedding lookup and the decoder is the fixed concat+MLP decoder.

B.2 Country-location alignment and consistency

B.2.1 Motivation and scope

The geospatial-alignment ablation tests alignment to external coordinate-based embedding spaces. This study instead tests a different representation question inside TerraNova: whether the model should explicitly align its native coordinate and administrative-country geometries, and whether country-level predictions should be regularized to match aggregated coordinate-level predictions sampled from the same country.

The upstream architecture is fixed to the selected sequential-ablation recipe. The coordinate branch uses the WIRE U-Net × Hash-L location encoder, the administrative branch uses the residual-MLP country encoder, the temporal branch uses Fourier+MLP, the transport module is the selected full-consistency transport configuration, the external geospatial-alignment branch is the selected inverse-distance configuration, the fusion block is the selected cross-modal Transformer, the task encoder is embedding-only, and the decoder is the selected concat+MLP task-conditioned decoder. In this experiment, the tested components are only the internal country-location alignment objective, the country-location prediction-consistency objective, and the within-country coordinate sampling density.

B.2.2 Protocol

Training follows the same randomized micro-epoch structure as the final multitask loop. Each mini-epoch mixes reconstruction steps, transport steps, fixed external geospatial-alignment steps, and optionally internal country-location steps. We reduce the number of geospatial alignment and transport steps per epoch to 256, to keep computation tractable

in the experiment. Let Ω_c denote the polygonal support of country c, let $x = (\phi, \lambda)$ be a coordinate (latitude, longitude), and let t be the queried year. The model produces a coordinate embedding $h_x = f_\theta^{\text{loc}}(x, t)$ and a country embedding $h_c = f_\theta^{\text{ctry}}(c, t)$ in the selected spatiotemporal source space. The reconstruction objective remains the standardized held-out-year prediction loss used in the previous WorldTensor/CountryTensor studies:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{B}|} \sum_{(u,v,t,y)\in\mathcal{B}} \left(\hat{y}_\theta(u, v, t) - y\right)^2,$$

where u is either a coordinate or a country identifier and all targets are standardized by the training split. In this experiment, we increased the number of both country-level and gridded tasks.

For the internal alignment objective, a country c is sampled uniformly across all countries, a set of positive coordinates $P_c = \{x_j^+\}$ is sampled from Ω_c, and negative coordinates $N_c = \{x_k^-\}$ are sampled from other countries (randomly selected). The positive coordinate embeddings are mean-pooled,

$$\bar{h}_c^+ = \frac{1}{|P_c|} \sum_{x\in P_c} h_x,$$

and a sampled-negative InfoNCE contrastive loss aligns the country embedding with the within-country coordinate embedding:

$$\mathcal{L}_{\text{cl}} = -\log \frac{\exp(s(h_c, \bar{h}_c^+)/\tau_{\text{T}})}{\exp(s(h_c, \bar{h}_c^+)/\tau_{\text{T}}) + \sum_{x^-\in N_c} \exp(s(h_c, h_{x^-})/\tau_{\text{T}})},$$

where $s(a, b) = \hat{a}^\top \hat{b}$ is cosine similarity of L2-normalised embeddings ($\hat{v} = v/\|v\|_2$) and $\tau_{\text{T}} = 0.07$ is the temperature.

The coordinate sampler tests two within-country densities:

$$q_{\text{uni}}(x \mid c) \propto \mathbf{1}\{x \in \Omega_c\}, \qquad q_{\text{pop}}(x \mid c) \propto \log(1 + P(x))\, \mathbf{1}\{x \in \Omega_c\},$$

where $P(x)$ is the population-density at a particular location and moment in time used by the sampler. Country selection itself remains uniform; only the coordinates sampled inside each country differ. We use $\log(1 + P(x))$ to compress the distribution of population densities so highly-populated regions don't dominate the sampling process. This was inspired by high-dynamic-range and linear-RGB specular-range compression algorithms in computer graphics.

The consistency objective is a prediction-space regularizer. For a country task v, country c, and year t, the country prediction is compared with the mean coordinate prediction over sampled coordinates in the same country:

$$\bar{y}_\theta(c, v, t) = \frac{1}{K} \sum_{j=1}^{K} \hat{y}_\theta(x_j, v, t), \qquad x_j \sim q(x \mid c),$$

$$\mathcal{L}_{\text{cons}} = \left(\hat{y}_\theta(c, v, t) - \bar{y}_\theta(c, v, t)\right)^2.$$

The total training objective in an active mini-epoch is therefore

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{tr}}\mathcal{L}_{\text{tr}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{cl}}\mathcal{L}_{\text{cl}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}},$$

with the transport and external geospatial-alignment terms held fixed from the previous ablations. We search for the best configuration using this grid:

$$A \in \{0, 256\}, \qquad C \in \{0, 256\}, \qquad \lambda_{\text{cl}}, \lambda_{\text{cons}} \in \{0, 0.25\},$$

where A and C are the internal alignment and consistency steps per mini-epoch. The no-country-location-loss baseline has $A = C = 0$; the uniform duplicate of this baseline is omitted because the coordinate sampler is unused when both internal losses are disabled. All confirmation runs use three seeds $(42, 314, 2718)$, 210,000 target training steps, fixed learning rate 1.5×10^{-4}, 256 fixed transport steps per mini-epoch, and 256 fixed external geospatial-alignment steps per mini-epoch. The primary selection metric is the equal-weight mean of held-out-year correlations on the two native evaluation geometries:

$$S = \frac{1}{2}\left(r_{\text{grid}} + r_{\text{ctry}}\right),$$

with auxiliary diagnostics reporting coordinate-to country retrieval, cosine similarity, margins, and preservation of the fixed external geospatial-alignment heads.

Family	A	C	Samp.	Score↑	ΔScore	Coord Top-1↑	Cosine↑	Grid r ↑	Ctry r ↑	RMSE↓
Alignment only	256	0	pop	**0.8554 ± 0.0135**	+0.0093	**0.6990 ± 0.0273**	0.3860	0.8774	0.8333	0.4194
Alignment only	256	0	uni	0.8538 ± 0.0154	+0.0077	0.7688 ± 0.0165	0.4183	0.8789	0.8287	0.4244
No country-location loss	0	0	pop	0.8461 ± 0.0065	–	0.0458 ± 0.0237	0.0469	0.8753	0.8169	0.4401
Consistency only	0	256	uni	0.8453 ± 0.0162	-0.0008	0.3406 ± 0.0236	0.2576	0.8740	0.8166	0.4448
Consistency only	0	256	pop	0.8434 ± 0.0140	-0.0026	0.3427 ± 0.1116	0.2658	0.8735	0.8134	0.4517
Alignment + consistency	256	256	uni	0.8365 ± 0.0043	-0.0095	0.8188 ± 0.0573	0.4876	0.8582	0.8149	0.4475
Alignment + consistency	256	256	pop	0.8351 ± 0.0096	-0.0110	0.6802 ± 0.0375	0.4215	0.8550	0.8151	0.4457

Table 19: **Country-location confirmation leaderboard.** Score, country-location retrieval, reconstruction, and per-modality validation metrics are reported as mean ± s.d. across confirmation seeds. ΔScore is relative to the no-loss baseline; −− denotes the reference row itself. A and C are internal country-location alignment and prediction-consistency steps per mini-epoch. Active internal losses use weight 0.25. The external geospatial-alignment branch, transport module, backbone, and reconstruction objective are fixed.

B.2.3 Results

The confirmed winner is Align A=256, w=0.25 + pop. It reaches a score of 0.8554±0.0135 and coordinate-to-country top-1 retrieval of 0.6990±0.0273. Relative to the no-country-location-loss baseline, the score change is +0.0093 and the coordinate-to-country top-1 change is +0.6531. The corresponding held-out-year correlations are 0.8774±0.0242 for gridded tasks and 0.8333±0.0106 for country tasks. The gridded-correlation delta is +0.0022, while the country-correlation delta is +0.0164. The predictive gain is therefore concentrated in the country geometry rather than in the coordinate-grid geometry, indicating that this alignment allows the relatively sparse country-level tasks to benefit from the higher-density information found in the gridded datasets. This matters because WorldTensor is not merely a gridded

version of the country-level dataset; it contains distinct Earth-system variables, including climate and air-pollution fields. That coordinate and country-level tasks improve one another under joint learning is a central result: the two native geometries are mutually informative. The consistency objective is not selected. Alignment-only increases country held-out-year correlation by +0.0164 with population sampling and +0.0118 with uniform sampling. Adding consistency to the matched alignment-only variant changes the confirmation score by -0.0203 for population sampling and -0.0173 for uniform sampling. Although consistency can raise some retrieval diagnostics, it lowers the predictive selection metric and is therefore excluded from the carried-forward recipe. This consistency score seems to be too strong of a regulariser and we discard it as there are no quantitative benefits in reconstruction tasks. Furthermore, its combination with the alignment loss is even more detrimental to reconstruction performance in both data modalities.

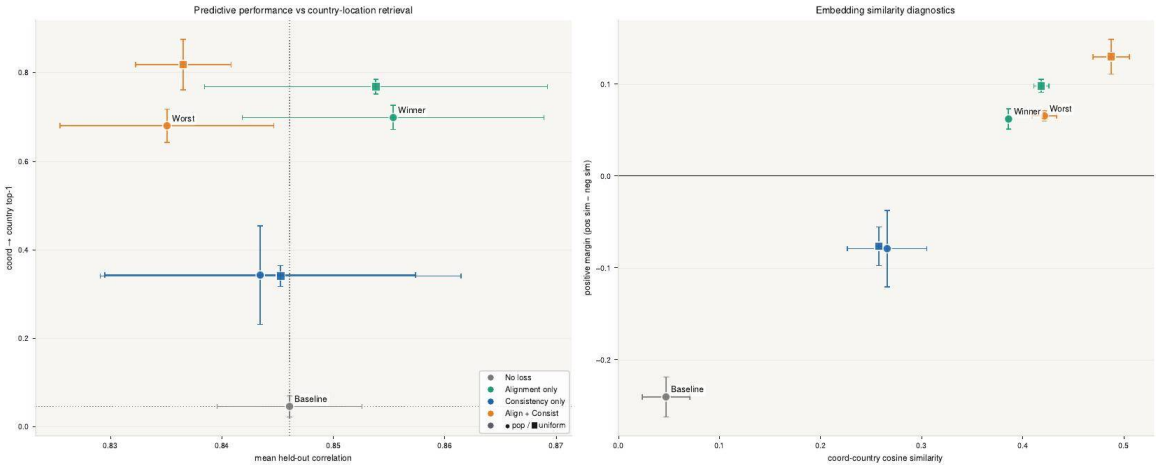

Figure 40: **Country-location retrieval trade-off.** The horizontal axis reports held-out predictive correlation and the vertical axis reports coordinate-to-country retrieval. Error bars show seed variation.

B.2.4 Conclusion

This ablation selects internal country-location alignment as an auxiliary representation objective and rejects the prediction-consistency objective for the current TerraNova recipe. The selected configuration is the population sampled, alignment-only objective with $A = 256$ and $\lambda_{\text{cl}} = 0.25$. The result is conceptually useful: the model already has strong coordinate-level performance from the fixed spatial and geospatial-alignment components, while internal country-location alignment mainly improves the administrative-country side of the joint benchmark. This simple yet effective contrastive loss proves useful for aligning country and gridded representations in a way that is mutually beneficial, allowing for soft-coupling of locations and countries without any architectural enforcement.

The chosen configuration is ***population sampled, alignment-only objective with*** $A = 256$ ***and*** $\lambda_{\text{cl}} = 0.25$*.. The consistency loss is discarded and does not move forward to the next experiment.*

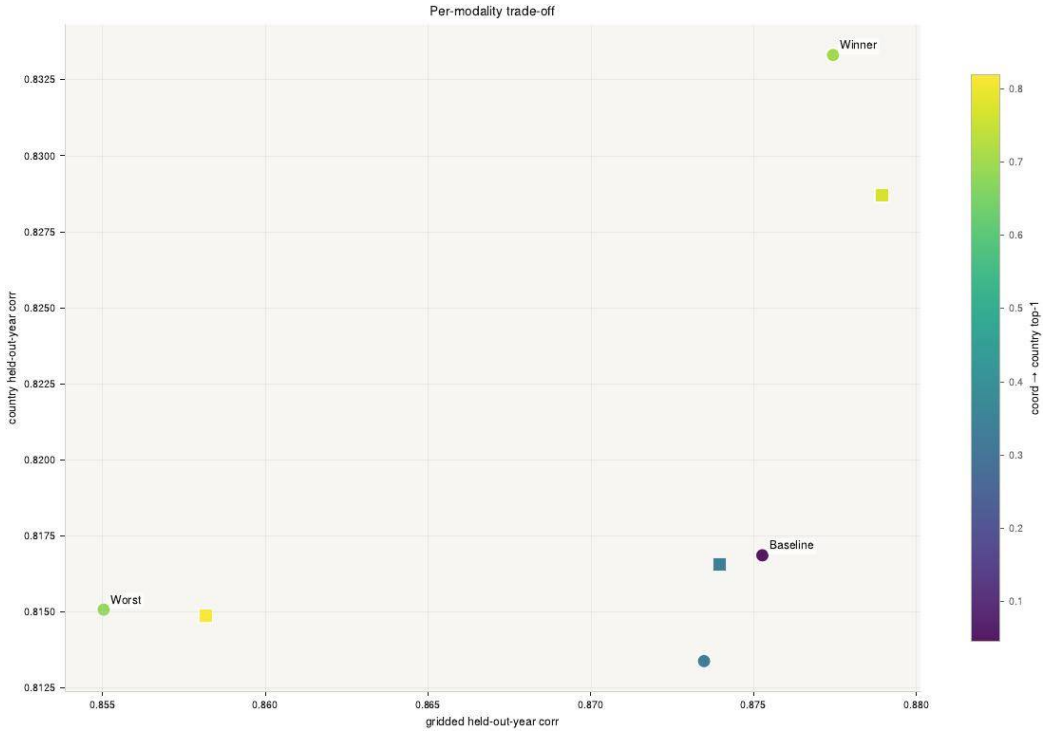

Figure 41: **Per-modality trade-off.** Gridded and country held-out-year correlations show whether the alignment and consistency objectives improve one geometry at the expense of the other.

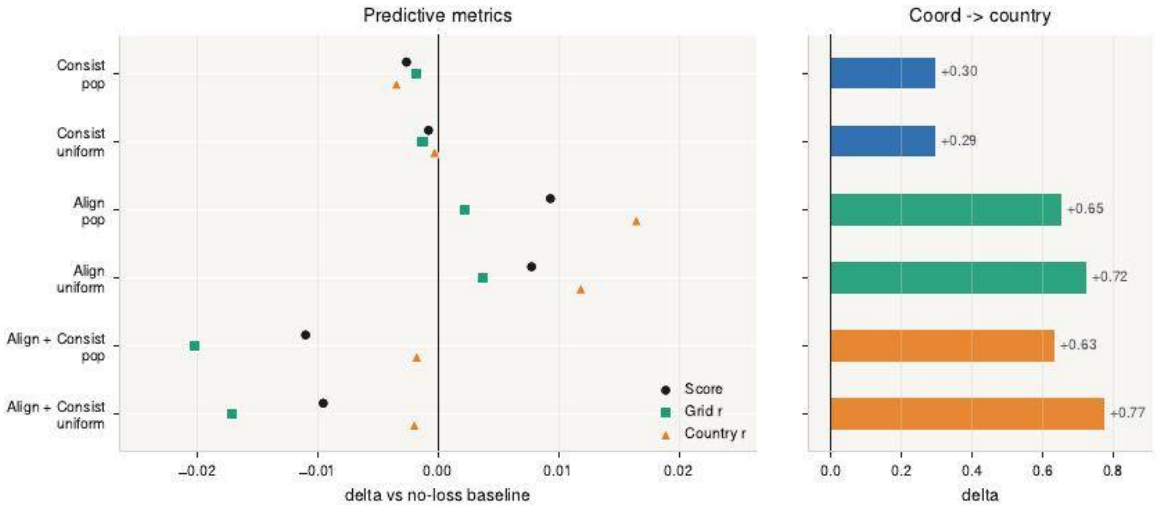

Figure 42: **Baseline-relative country-location effects.** Metric deltas show that the alignment-only variants are the main source of the predictive gain, while consistency does not improve the selected score.

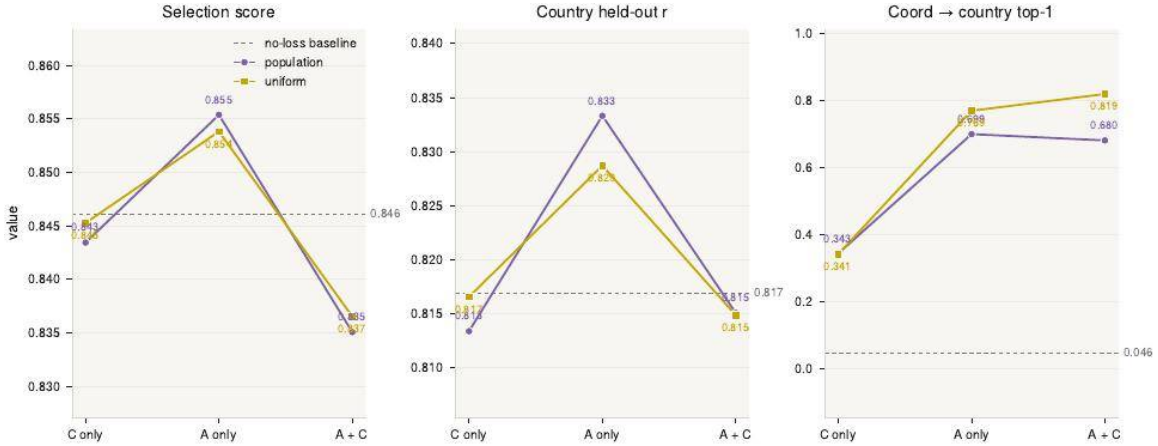

Figure 43: **Alignment and consistency contrasts.** Consistency-only variants, alignment-only variants, and joint variants are shown against the same no-loss baseline; population- and uniformly sampled coordinate negatives show that consistency does not add to the alignment objective.

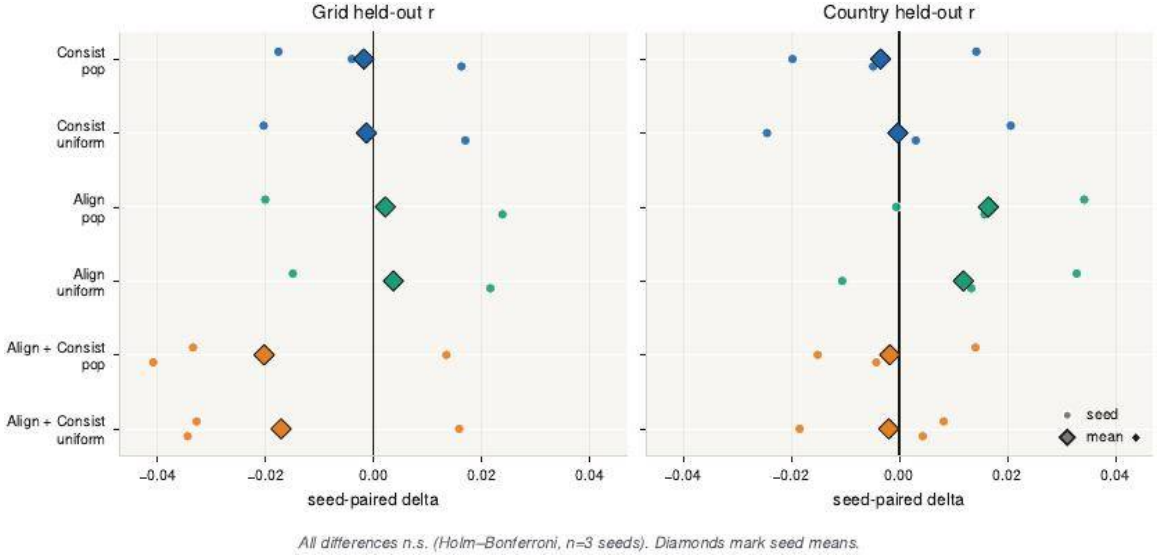

Figure 44: **Seed-paired validation deltas.** Each point is a seed paired against the no-country-location-loss baseline. Diamonds mark seed means.

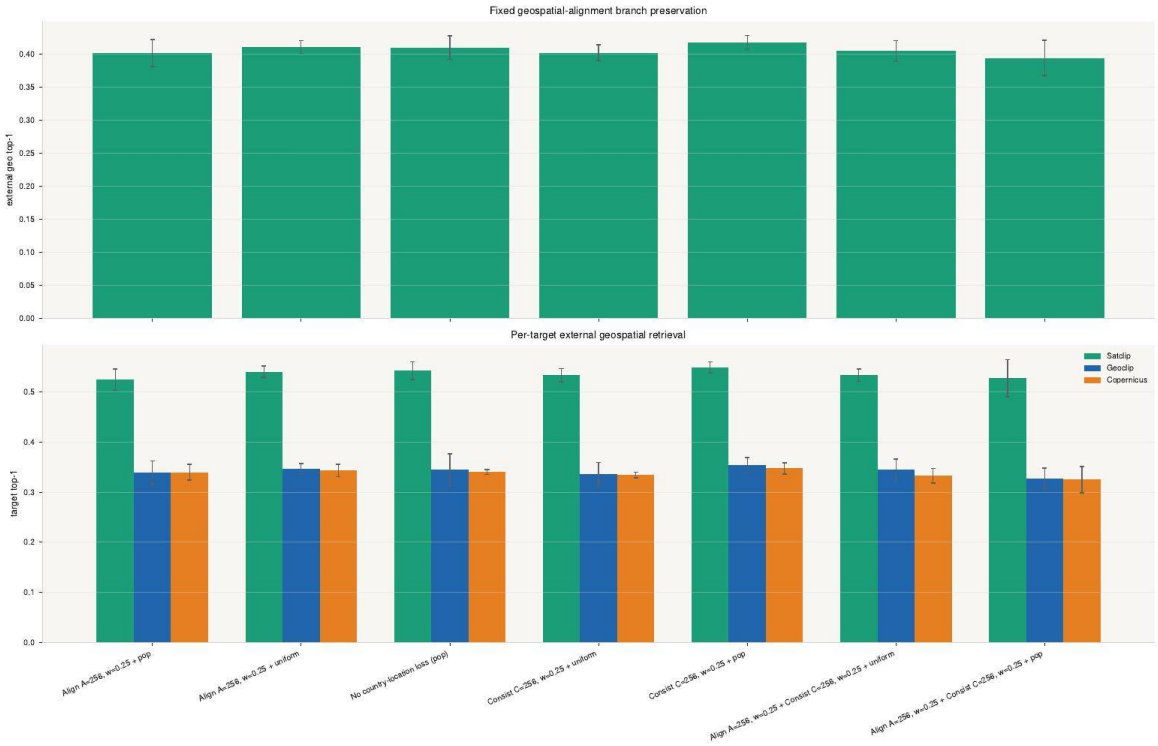

Figure 45: **External geospatial-alignment preservation.** The fixed external geospatial-alignment branch remains monitored so that country-location objectives do not silently destroy the previously selected representation property.

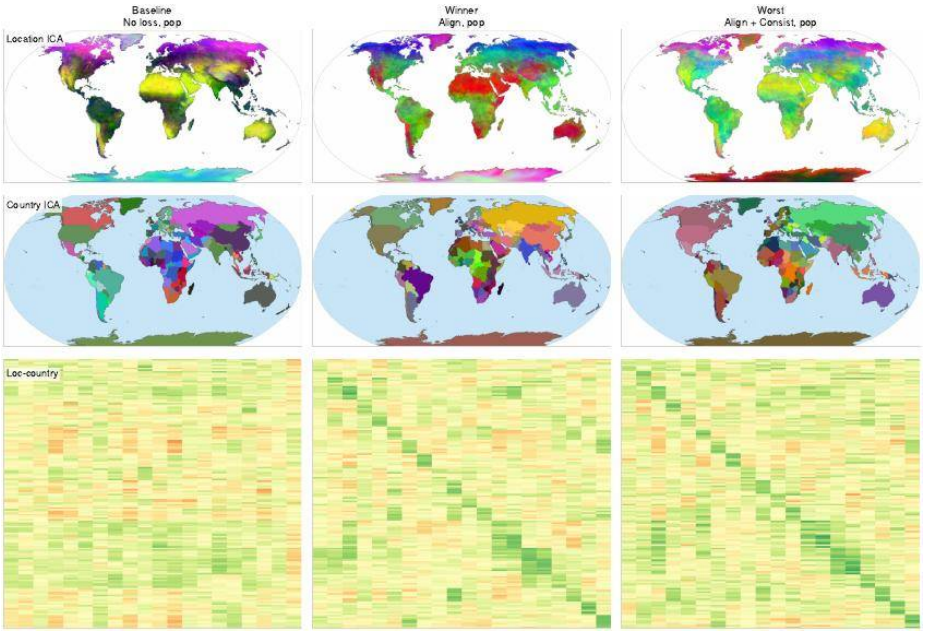

Figure 46: **Representative embedding geometry.** Seed-42 qualitative artifacts compare the no-loss baseline, the confirmed winner, and the lowest-scoring confirmed variant. These panels are diagnostic only and are not used for model selection.

B.3 Decoder & Task encoder

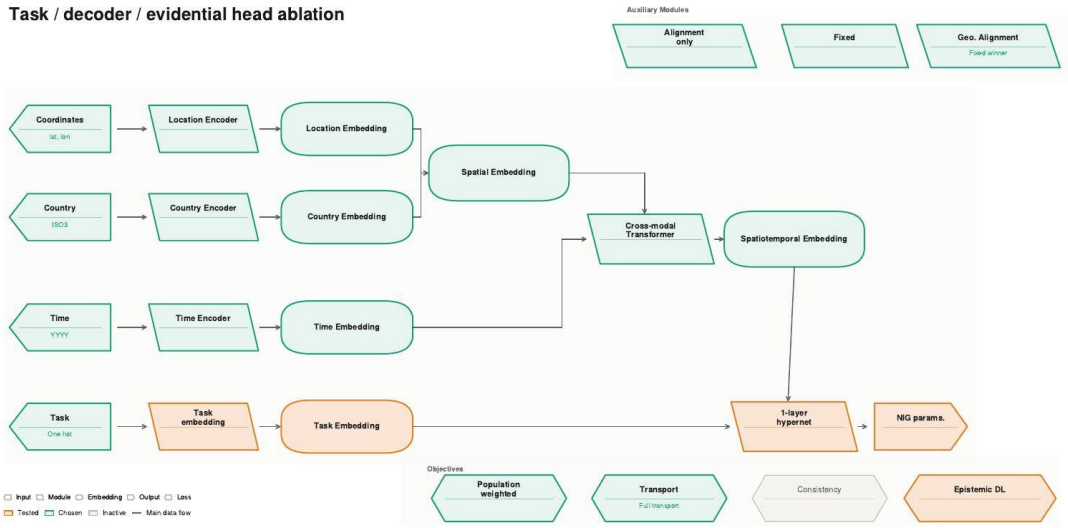

Figure 47: **Task-conditioned decoder and evidential-head ablation scope.** The location, country, time, transport, geospatial-alignment, and country-location alignment pathways are fixed to the selected upstream recipe. The consistency loss is inactive. The carried-forward decoder recipe uses task embeddings, a cross-modal Transformer between the spatiotemporal state and the task embedding, a one-layer hypernetwork decoder, and a single NIG head emitting the four evidential parameters.

B.3.1 Motivation and scope

The preceding ablations select the coordinate, country, temporal, transport, external geospatial-alignment, internal country-location alignment, and spatiotemporal-fusion components of TerraNova. This ablation tests the final prediction interface: how task identity should interact with the selected spatiotemporal representation, and how the decoder should express predictive uncertainty.

The fixed upstream recipe is the selected sequential-ablation backbone. The country-location alignment objective is kept active with population-weighted within-country coordinate sampling, while the prediction-consistency objective is inactive because it was rejected in the country-location study. The external geospatial-alignment branch and temporal transport branch are fixed. The ablated components are the task encoder, the task–spatiotemporal fusion module, the decoder, and the Normal-Inverse-Gamma (NIG) prediction head (Fig. 47). This is different compared to the previous ablation studies: we move beyond a simple MSE reconstruction objective towards a fully evidential deep learning loss function.

The design question is also different from the earlier spatiotemporal-fusion study. There, the model learned how spatial and temporal information should combine before any task-specific prediction. Here, the input to the decoder is already a selected spatiotemporal embedding. We are interested in whether each prediction task should be represented only as a learned identifier, whether it should be processed by additional residual layers, whether it should

condition the decoder through a hypernetwork, and whether the four NIG parameters should be predicted jointly or with separate output heads.

B.3.2 Model family

Let u_i denote the fixed spatiotemporal embedding for sample i, and let k_i denote the requested reconstruction task. The task encoder first maps the task identifier to a vector $e_{k_i} = E_\psi(k_i)$. The embedding-only variant uses this lookup directly. The residual task encoder instead applies a residual MLP R_ψ to the lookup: $e^{\text{res}}_{k_i} = R_\psi(e_{k_i})$. Both variants use the same task registry and output dimension; the difference is whether task identity is allowed to pass through extra nonlinear layers before conditioning the decoder. This was also tested for the country encoder, where we saw that these additional residual layers gave a small but quantifiable encoding advantage.

The task–spatiotemporal fusion block maps the fixed spatiotemporal state and the task embedding to a conditioning vector $c_i = G_\phi(u_i, e_{k_i})$. Three fusion families are tested. The *concat-MLP* baseline uses $c_i = M_\phi([u_i; e_{k_i}])$. The *task-only baseline* removes the spatiotemporal input from the conditioning path and sets $c_i = e_{k_i}$. This does not remove the spatiotemporal state from hypernetwork decoders: it only means that the generated decoder weights depend on task identity alone. Finally, the *cross-modal Transformer* uses the spatiotemporal token as a query and the task token as key/value memory. With $q_0 = W_u u_i, \qquad m = W_e e_{k_i}$, each block applies gated cross-attention followed by a gated feed-forward update, $\tilde{q}_\ell = q_\ell + g^{\text{attn}}_\ell \,\text{Attn}(\text{LN}(q_\ell), m, m)$, $q_{\ell+1} = \tilde{q}_\ell + g^{\text{ffn}}_\ell \,\text{FFN}(\text{LN}(\tilde{q}_\ell))$, and outputs $c_i = W_o q_L$. This type of model lets the spatial-temporal state ask which task information is needed for the current prediction. With it, we enable the model to learn that tasks are spatiotemporally dependent, rather than fixed.

Two decoder families are compared. The simple decoder predicts directly from the fused task–spatiotemporal vector: $a_i = Ac_i + b$, where a_i is the raw output vector. The one-layer hypernetwork decoder is more structured. Its weight-generating core first maps the conditioning vector through a single hidden projection, $h_i = \sigma(W_h c_i + b_h)$, where σ is the GELU nonlinearity. Parameter-generator heads then emit the weights of a sample-specific one-hidden-layer decoder,

$$W_{1,i}, b_{1,i}, W_{2,i}, b_{2,i} = \big(P_{W_1}(h_i), P_{b_1}(h_i), P_{W_2}(h_i), P_{b_2}(h_i)\big),$$

$$a_i = \text{GELU}(u_i W_{1,i} + b_{1,i}) W_{2,i} + b_{2,i}.$$

Thus the task-conditioned path generates the local decoder, but the generated decoder is applied to the fixed spatiotemporal state. This preserves spatial and temporal information at the final prediction step while allowing different tasks to use different local models. The selected configuration uses this single-projection hypernetwork core, the generated one-hidden-layer decoder, an L2 row-normalisation constraint on generated weights, and a single NIG output head.

B.3.3 Evidential prediction and loss

The decoder emits four raw scalars that are transformed into Normal-Inverse-Gamma parameters:

$$\mu_i = a_{\mu,i},$$

$$\nu_i = \text{softplus}(a_{\nu,i}) + 0.5, \qquad \alpha_i = \text{softplus}(a_{\alpha,i}) + 1 + 10^{-4}, \qquad \beta_i = \text{softplus}(a_{\beta,i}) + 10^{-4}.$$

In our implementation, we constrain the potential values for the uncertainties as:

$$\beta_i \leq \sigma^2_{\max}(\alpha_i - 1), \qquad \nu_i \geq \frac{\beta_i}{\sigma^2_{\max}(\alpha_i - 1)}, \qquad \sigma_{\max} = 3.$$

The single-head variant predicts $(\mu, \nu, \alpha, \beta)$ from one shared projection. The four-head variant uses separate output projections for the four parameters. The latter is slightly more flexible, but it also adds an extra degree of freedom to a parameterisation that is already weakly identified by the predictive loss.
The NIG marginal likelihood is the Student-t likelihood obtained by integrating over the latent Gaussian mean and variance. For a standardised target y_i, with

$$\omega_i = 2\beta_i(1 + \nu_i),$$

the per-sample negative log-likelihood is

$$\mathcal{L}_{\text{NIG},i} = \frac{1}{2}\log\frac{\pi}{\nu_i} - \alpha_i \log\omega_i + \left(\alpha_i + \frac{1}{2}\right)\log\big(\nu_i(y_i - \mu_i)^2 + \omega_i\big) + \log\Gamma(\alpha_i) - \log\Gamma\left(\alpha_i + \frac{1}{2}\right).$$

The training objective adds the width-normalised evidential regulariser used throughout this ablation:

$$w_{\text{St},i} = \sqrt{\frac{\beta_i(1 + \nu_i)}{\alpha_i \nu_i}},$$

$$\mathcal{R}_{\text{evi},i} = \left|\frac{y_i - \mu_i}{w_{\text{St},i}}\right|^p (2\nu_i + \alpha_i), \qquad p = 1,$$

and a soft width cap on the Student-t predictive width,

$$\mathcal{R}_{\text{cap},i} = \max(0, w_{\text{St},i} - 3)^2.$$

The reconstructed-target loss is therefore

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{B}|}\sum_{i \in \mathcal{B}} \left[\mathcal{L}_{\text{NIG},i} + 0.1\,\mathcal{R}_{\text{evi},i} + 0.1\,\mathcal{R}_{\text{cap},i}\right].$$

The NIG parameters also define uncertainty diagnostics. The classic heuristic decomposition is

$$\text{Var}_{\text{alea}} = \frac{\beta}{\alpha - 1}, \qquad \text{Var}_{\text{epi}} = \frac{\beta}{\nu(\alpha - 1)}.$$

Because the split is not uniquely identified by the marginal likelihood, model selection in this ablation does not use NLL or the uncertainty decomposition. Selection uses held-out reconstruction RMSE, with calibration and uncertainty reported as diagnostics.

B.3.4 Protocol

The grid we test in this experiment is

$$\{\text{embedding}, \text{residual}\} \times \{\text{concat MLP}, \text{cross modal Transformer}, \text{task only}\} \times \{\text{MLP decoder}, \text{one layer hypernetwork}\} \times \{\text{single NIG head}, \text{four NIG heads}\}.$$

The task-only/simple-decoder combination is excluded because the simple decoder ignores the spatiotemporal state at the head level; with task-only conditioning it would reduce to a task lookup with no location or time signal. This leaves 20 NIG variants. The recovered set contains 59 summaries from the planned 20×3 seed grid. The selected model and the raw-RMSE rank-one H4 model both have all three confirmation seeds $(42, 314, 2718)$. Each confirmation run uses a fixed learning rate of 8×10^{-5}, weight decay 0.02, and a target budget of 450,000 optimisation steps. Each mixed mini-epoch contains 32,768 reconstruction steps, 256 transport steps, 256 external geospatial-alignment steps, and 256 country-location alignment steps. The internal prediction-consistency step budget is zero.
The primary selection statistic is validation RMSE on standardised held-out-year reconstruction:

$$\overline{\text{RMSE}} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \sqrt{\frac{1}{|\mathcal{V}_\tau|} \sum_{i \in \mathcal{V}_\tau} (\hat{y}_i - y_i)^2},$$

where $\mathcal{T}$ is the set of evaluated reconstruction tasks and $\mathcal{V}_\tau$ is the held-out validation set for task τ. We also report mean held-out correlation, gridded held-out correlation, country held-out correlation, and expected calibration error.

Arch	TE	Dec	H	RMSE↓	ΔRMSE	Corr↑	Grid r	Ctry r	ECE↓
Tr-cross	emb	hyper	4	0.3974 ± 0.0119	-0.0015	0.6369 ± 0.0185	0.453	0.821	0.154
Tr-cross	emb	hyper	1	**0.3988 ± 0.0092**	0.0000	0.6414 ± 0.0249	0.464	0.818	0.173
Tr-cross	res	hyper	4	0.4001 ± 0.0106	0.0012	0.6306 ± 0.0317	0.455	0.807	0.185
Tr-cross	emb	MLP	4	0.4044 ± 0.0130	0.0056	0.6436 ± 0.0160	0.462	0.825	0.182
Tr-cross	emb	MLP	1	0.4052 ± 0.0113	0.0064	0.6478 ± 0.0251	0.469	0.826	0.211
Tr-cross	res	hyper	1	0.4077 ± 0.0138	0.0089	0.6368 ± 0.0181	0.458	0.816	0.174
Tr-cross	res	MLP	4	0.4130 ± 0.0122	0.0142	0.6257 ± 0.0268	0.444	0.808	0.209
Tr-cross	res	MLP	1	0.4135 ± 0.0071	0.0146	0.6383 ± 0.0212	0.465	0.812	0.191
Concat	emb	MLP	1	0.4300 ± 0.0152	0.0312	0.6274 ± 0.0201	0.465	0.789	0.182
Concat	emb	MLP	4	0.4314 ± 0.0146	0.0326	0.6259 ± 0.0138	0.463	0.789	0.173
Task	emb	hyper	1	0.5081 ± 0.0182	0.1093	0.5533 ± 0.0185	0.393	0.714	0.163
Concat	emb	hyper	4	0.5105 ± 0.1001	0.1117	0.5568 ± 0.1065	0.406	0.707	0.203
Task	res	hyper	4	0.5230 ± 0.0212	0.1242	0.5351 ± 0.0056	0.425	0.645	0.186
Task	res	hyper	1	0.5280 ± 0.0136	0.1292	0.5410 ± 0.0309	0.396	0.686	0.155
Task	emb	hyper	4	0.5287 ± 0.0212	0.1299	0.5423 ± 0.0066	0.431	0.654	0.184
Concat	emb	hyper	1	0.6403 ± 0.0825	0.2415	0.4622 ± 0.0707	0.354	0.570	0.206
Concat	res	MLP	1	0.6680 ± 0.1552	0.2692	0.3931 ± 0.1476	0.289	0.497	0.189
Concat	res	MLP	4	0.8226 ± 0.0168	0.4238	0.1959 ± 0.0661	0.115	0.277	0.216
Concat	res	hyper	1	0.8587 ± 0.0278	0.4599	0.1608 ± 0.0090	0.164	0.158	0.227
Concat	res	hyper	4	0.8622 ± 0.0286	0.4634	0.1512 ± 0.0056	0.149	0.154	0.226

Table 20: **Decoder ablation study results.**

B.3.5 Results

The best raw-RMSE model is the task-embedding, cross-modal-Transformer, one-layer-hypernetwork decoder with four NIG heads: $\overline{\text{RMSE}}_{\text{H4}} = 0.3974 \pm 0.0119$. The corresponding single-head model reaches $\overline{\text{RMSE}}_{\text{H1}} = 0.3988 \pm 0.0092$, only 0.0015 RMSE units worse. In the seed-paired comparison against H4, the single-head model has a mean RMSE difference of $+0.0015$ with a 95% interval from -0.0027 to $+0.0057$ when expressed as H1 minus H4, and the Holm-adjusted paired test gives $p = 1.0$. The two variants are therefore indistinguishable at the available seed budget. We carry forward the single-head model because it has the same task encoder, fusion block, and hypernetwork decoder as H4, but a simpler evidential prediction head. It also has the higher mean correlation (0.6414 versus 0.6369).

Where the error lives. Figure 48 separates the two evaluation modalities. The selected model reconstructs country-level signals very well ($r_{\text{country}} = 0.82$) but is weaker on the fine gridded targets ($r_{\text{grid}} = 0.46$). The decoder ablation has therefore essentially saturated the country modality: no variant in the grid moves the country correlation much. The remaining error budget is in high-frequency spatial detail, not in the task-conditioning interface this ablation studies, so once the cross-modal Transformer is in place the decoder is no longer the bottleneck.

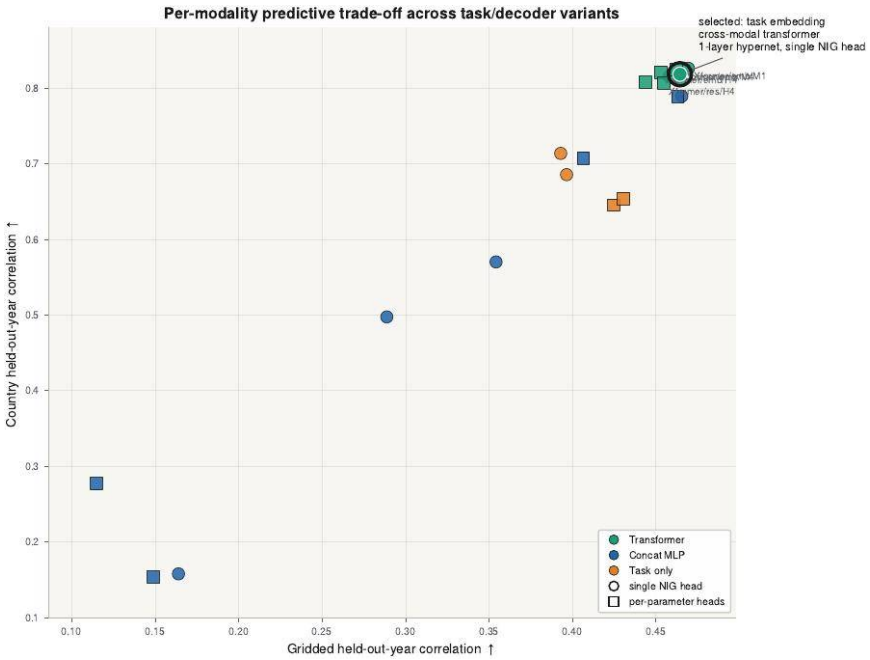

Figure 48: **Per-modality predictive trade-off.** The horizontal axis is gridded held-out-year correlation and the vertical axis is country held-out-year correlation. The selected single-head hypernetwork model is highlighted. It lies on the top Transformer cluster and preserves the same gridded–country balance as the slightly lower-RMSE H4 variant: the spread between variants is almost entirely horizontal, confirming that the decoder choice is decided on gridded reconstruction while country performance is already saturated.

Effect sizes against the selected model. Figure 49 compares every confirmed variant against the selected model on matched seeds. The large positive differences come from concat-MLP and task-only variants, especially when the residual task encoder is combined with weak task–spatiotemporal fusion. The Transformer family is much closer to the selected model, and the H4 comparison sits near zero. This pattern supports two conclusions: the cross-modal Transformer is the important architectural change, and the H1/H4 distinction is a small head-level choice within the winning family.

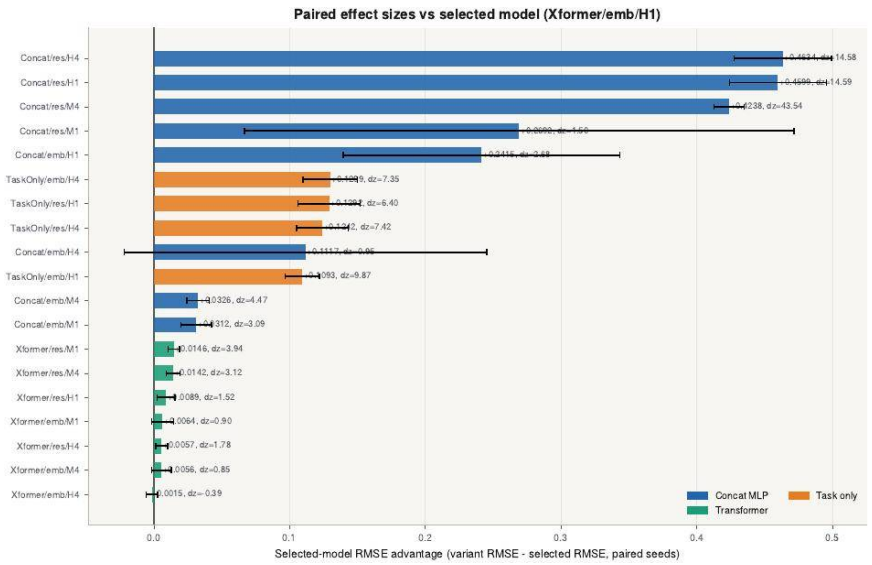

Figure 49: **Seed-paired RMSE effect sizes relative to the selected model.** Positive bars mean the alternative has higher RMSE than the selected model on matched seeds. The H4 variant is effectively tied with the selected H1 variant, whereas weaker fusion and task-encoder combinations are separated by large RMSE margins.

Task-level behaviour. Figure 50 decomposes held-out R^2 by reconstruction task. No single architecture wins every task, but the Transformer variants dominate the high-performing columns. The selected H1 model is not chosen because it wins the largest number of individual tasks; it is chosen because it lies in the best global RMSE cluster while retaining the simpler and more standard NIG head.

B.3.6 Conclusion

This ablation selects a task-embedding, cross-modal Transformer, hypernetwork decoder with a single NIG head. The raw RMSE winner is the otherwise identical four-head NIG variant, but the gap is only 0.0015 RMSE units and is not significant with three matched seeds. The single-head model is therefore the preferred carried-forward recipe: it keeps the decisive architectural components, removes an unnecessary output-head over-parameterisation, and preserves the same gridded–country performance balance.

The broader result is that task conditioning must happen before the decoder and must be allowed to interact repeatedly with the spatiotemporal state. Residual task encoders do not

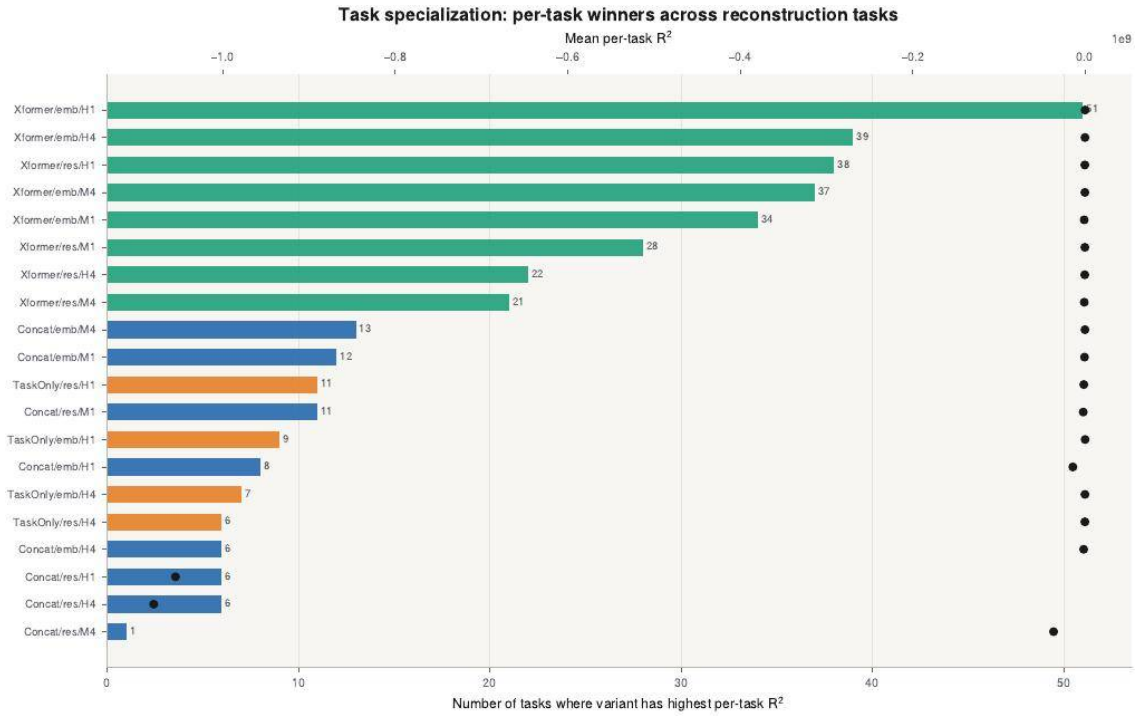

Figure 50: **Task-level held-out R^2 across variants.** Black boxes mark per-task winners; the selected global model is marked separately. The figure shows that task-level winners are distributed across the Transformer family, while weaker fusion families lose many tasks simultaneously.

help, task-only conditioning is insufficient, and concat MLP fusion is not competitive with cross-modal attention. A one-layer hypernetwork is a natural final readout because it lets task identity generate a local decoder for the selected spatiotemporal state, while the NIG head provides calibrated uncertainty diagnostics without changing the RMSE-based selection criterion.

The chosen configuration is the ***task embedding, cross-modal Transformer, Hypernetwork decoder with a single NIG head****.*

B.4 Training configuration: Multi-scale curriculum, active learning, and regularisation

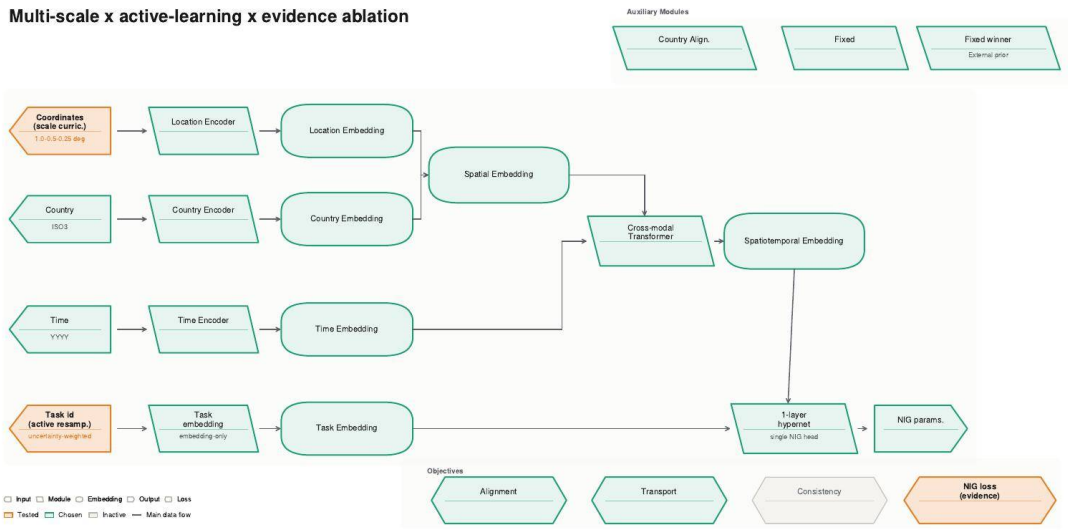

Figure 51: **Multi-scale, active-learning, and evidence ablation scope.** The entire model architecture, data diet, optimiser, and evaluation protocol are fixed to the carried-forward evidential recipe selected by the preceding decoder and task-encoder ablation. Orange components mark the three training-time interventions tested in this study, each shown on the pipeline stage it acts upon: the scale curriculum coarsens the coordinate input, active resampling reweights the task input, and the evidence coefficient scales the Normal-Inverse-Gamma (NIG) reconstruction loss. The consistency objective is inactive. The task encoder is an embedding lookup, the fusion block is the selected cross-modal Transformer, and the decoder is the fixed single-head hypernetwork NIG decoder.

B.4.1 Motivation and scope

The preceding ablations select the architecture of TerraNova: the coordinate, country, temporal, transport, external geospatial-alignment, internal country-location alignment, spatiotemporal-fusion, task-encoder, decoder, and evidential-head components are all fixed by the sequential studies that come before this one. This ablation tests a different set of design choices. Rather than asking *what* the model should be, it asks *how* the model should be trained: whether a coarse-to-fine spatial-scale curriculum, an uncertainty-driven active-resampling schedule, and the strength of the evidential regulariser improve the evidential reconstruction model under a fixed optimiser budget. These are training-procedure interventions, not architectural ones, so they are studied jointly on a single frozen hypothesis class.

The upstream recipe is held fixed at the carried-forward winner of the decoder ablation (Fig. 47). The coordinate branch, country branch, temporal branch, transport module, external geospatial-alignment branch, and country-location alignment objective are all kept at their selected configurations; the prediction-consistency objective is inactive because it was rejected in the country-location study. The task encoder is embedding-only, the fusion block

is the cross-modal Transformer, and the decoder is the single-head one-layer hypernetwork emitting the four NIG parameters. The data, the optimiser family and base learning rate, the precision, the effective batch size, the step-family mixture, the geo-alignment configuration, the modality-balance fraction, and the evaluation protocol are constant across all runs. Only the three training-time interventions vary, which makes the design a clean estimate of training-procedure effects on a fixed model.
The scale curriculum $A \in \{\text{off}, \text{on}\}$ trains coarse-to-fine over resolutions $1.0° \to 0.5° \to 0.25°$ with a per-scale learning-rate restart; active learning $B \in \{\text{off}, \text{on}\}$ resamples tasks by their evidential uncertainty; and the evidence coefficient $C = \lambda_{\text{ev}} \in \{0.1, 0.3\}$ scales the width-normalised evidence regulariser. This is a $2 \times 2 \times 2$ grid, replicated over three seeds $(42, 314, 2718)$ for 24 runs in total. As in previous runs, we tested against the held-out reconstruction error and calibration diagnostics.

Hypothesis 1 (Curriculum)**.** The scale curriculum reduces reconstruction error relative to native-resolution training at equal optimiser budget; the main effect of A on the selection metric is negative.

Hypothesis 2 (Active learning)**.** Uncertainty-driven resampling reduces error and improves calibration; the main effect of B is favourable on both the accuracy and the calibration axes.

Hypothesis 3 (Evidence strength)**.** Increasing λ_{ev} trades sharpness for calibration: larger λ_{ev} regularises the evidential head toward better-calibrated but less sharp predictives, with a non-monotone effect on point error.

Hypothesis 4 (Interaction)**.** The curriculum and active learning interact: both reshape the effective training distribution, so their joint effect departs from additivity.

B.4.2 Fixed evidential backbone

We summarise here the fixed backbone to fix notation; it is identical in every run. A query is a triple (τ, s, t) of a task identifier τ, a spatial location $s = (\text{lon}, \text{lat})$ (or a country token), and a time t (year, optionally month and day). The location, time, task, and country encoders map these to embeddings $\mathbf{e}_s, \mathbf{e}_t, \mathbf{e}_\tau, \mathbf{e}_c \in \mathbb{R}^d$. The task encoder is embedding-only, a learned lookup $\mathbf{e}_\tau = E[\tau]$ with no residual MLP branch, which is the configuration that won the decoder ablation. A spatiotemporal token $\mathbf{z}_{\text{st}}$ is formed from $\mathbf{e}_s, \mathbf{e}_t$ (and $\mathbf{e}_c$ when a country query is active) and fused with the task token by a cross-modal Transformer, in which the spatiotemporal token attends to the task token as key/value memory, producing a fused context $\mathbf{h} = \text{Fuse}(\mathbf{e}_\tau, \mathbf{z}_{\text{st}}) \in \mathbb{R}^{d_f}$. A one-layer hypernetwork g_ϕ then maps the task token to the weights of a small per-task predictor, which is applied to $\mathbf{h}$ to emit the four NIG parameters of a single shared head,

$$(\mu, \nu, \alpha, \beta) = f_\theta(\tau, s, t), \qquad \mu \in \mathbb{R},\ \nu > 0,\ \alpha > 1,\ \beta > 0. \tag{107}$$

Positivity is enforced with softplus offsets, $\nu = \text{softplus}(\cdot) + \frac{1}{2}$, $\alpha = 1 + \text{softplus}(\cdot) + 10^{-4}$, $\beta = \text{softplus}(\cdot) + 10^{-4}$, and the loss applies numerical floors in `fp32` for the log and Γ terms. The single-head choice, rather than four independent parameter heads, is the decoder ablation's winner and is held fixed here.

B.4.3 Evidential prediction and loss

Following deep evidential regression, we place a conjugate Normal-Inverse-Gamma prior on the unknown mean γ and variance σ^2 of a Gaussian observation model, $y \sim \mathcal{N}(\gamma, \sigma^2)$ with $\gamma \sim \mathcal{N}(\mu, \sigma^2/\nu)$ and $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, so that $(\gamma, \sigma^2) \sim \mathrm{NIG}(\mu, \nu, \alpha, \beta)$. The network emits the four hyperparameters per query. Marginalising γ and σ^2 gives a Student-t posterior predictive,

$$p(y \mid \mu, \nu, \alpha, \beta) = \mathrm{St}\Big(y;\ \mu,\ \tfrac{\beta(1+\nu)}{\nu\,\alpha},\ 2\alpha\Big), \tag{108}$$

whose exact negative log-density, writing $\Omega = 2\beta(1+\nu)$, is the per-sample NIG negative log-likelihood

$$\mathcal{L}^{\mathrm{NLL}}(y) = \tfrac{1}{2}\log\frac{\pi}{\nu} - \alpha\log\Omega + \big(\alpha + \tfrac{1}{2}\big)\log\big(\nu(y-\mu)^2 + \Omega\big) + \log\Gamma(\alpha) - \log\Gamma\big(\alpha + \tfrac{1}{2}\big). \tag{109}$$

The classic evidence regulariser $|y-\mu|(2\nu+\alpha)$, is known to drive $\nu \to 0$ and to confound aleatoric noise with epistemic uncertainty because ν and α are not separately identifiable from eq. (109). We therefore use the width-normalised regulariser built on the identifiable Student-t predictive width

$$w_{\mathrm{St}} = \sqrt{\frac{\beta(1+\nu)}{\alpha\,\nu}}, \tag{110}$$

which is the scale of the marginal in eq. (108). The regulariser penalises the standardised residual, weighted by the total evidence $2\nu + \alpha$,

$$\mathcal{L}^{\mathrm{R}}(y) = \left|\frac{y-\mu}{w_{\mathrm{St}}}\right|^{p}(2\nu+\alpha), \qquad p = 1, \tag{111}$$

so that only errors that are large relative to the predicted scale accrue an evidence penalty, and a genuinely high-noise region is not mistaken for an epistemically uncertain one. Because targets are per-task standardised to unit marginal variance, an admissible predictive width is $O(1)$; eq. (111) leaves a degenerate escape hatch ($w_{\mathrm{St}} \to \infty$ makes the standardised residual vanish), which we close with a smooth one-sided cap

$$\mathcal{L}^{\mathrm{cap}}(y) = \mathrm{relu}\big(w_{\mathrm{St}} - W_{\max}\big)^2, \qquad W_{\max} = 3, \tag{112}$$

a soft, differentiable bound at $\approx 3\sigma$ on standardised data that never hard-clamps. The reconstruction objective for one observation is therefore

$$\boxed{\mathcal{L}(y) = \mathcal{L}^{\mathrm{NLL}}(y) + \lambda_{\mathrm{ev}}\,\mathcal{L}^{\mathrm{R}}(y) + \kappa\,\mathcal{L}^{\mathrm{cap}}(y), \quad \kappa = 0.1,} \tag{113}$$

reduced by the mean over the batch. The evidence coefficient λ_{ev} is the third ablation factor; the width-cap coefficient κ is fixed. For ranking and for active learning we use the identifiable Meinert proxies: the predictive width w_{St} of eq. (110) as the aleatoric scale and $1/\sqrt{\nu}$ as the epistemic evidence, with total uncertainty $U = w_{\mathrm{St}} + 1/\sqrt{\nu}$.

B.4.4 Configuration 1: spatial-scale curriculum

The curriculum visits resolutions $r_1 > r_2 > \cdots > r_S$ in degrees, coarse to fine, with native resolution $r_S = \min_s r_s$; the study uses $(r_1, r_2, r_3) = (1.0°, 0.5°, 0.25°)$. Given a total of N epochs derived from the target optimiser-step budget and per-scale fractions $f_1, \ldots, f_{S-1}$ with the final scale taking the remainder, the per-scale epoch counts are

$$c_s = \max\big(1,\ \mathrm{round}(f_s N)\big)\ \ (s < S), \qquad c_S = N - \sum_{s<S} c_s, \tag{114}$$

with a finest-first repair: if c_S would be non-positive, epochs are returned from coarse scales (decrementing c_s while $c_s > 1$) until $c_S \geq 1$, and if $N \leq S$ the coarsest scales are dropped so the finest scales are always trained. By construction $\sum_s c_s = N$, and scale s is active over epochs $[\sum_{s'<s} c_{s'},\ \sum_{s'\leq s} c_{s'})$. Coarsening acts only on gridded reconstruction batches when the active resolution exceeds native ($r > r_S$); country tasks carry a single representative coordinate and are untouched. Each coordinate is snapped to the r-grid, $q_r(x) = \mathrm{round}(x/r) \cdot r$, and labels are mean-pooled within each occupied cell $(\hat{s}, \mathrm{year})$,

$$\tilde{y}_i = \frac{1}{|\mathcal{C}(i)|} \sum_{j \in \mathcal{C}(i)} y_j, \qquad \mathcal{C}(i) = \{\, j : (\hat{s}_j, \mathrm{year}_j) = (\hat{s}_i, \mathrm{year}_i) \,\}, \tag{115}$$

so the model sees a smoothed, lower-frequency target field early in training and the full-resolution field once the native scale is reached. At each scale transition the optimiser schedule is restarted with a fresh warmup-cosine cycle sized to that scale's epoch count c_s (Loshchilov and Hutter, 2017): warmup over a proportional prefix and cosine decay to the floor over the remainder of c_s. This prevents a coarse scale's decay tail from suppressing learning on the next finer scale, and ensures the native scale receives a full, fresh decay to the minimum learning rate.

Algorithm 8 Scale-curriculum control (per epoch)

Require: resolutions $r_1 > \ldots > r_S$, fractions $f_{1:S-1}$, epochs N
  $c_{1:S} \leftarrow$ epoch budget by eq. (114); cumulative starts σ_s
  for each training epoch $e = 0, 1, \ldots, N-1$ **do**
    $s \leftarrow \max\{\, i : c_i > 0 \text{ and } e \geq \sigma_i \,\}$ ▷ active scale index
    if s changed from previous epoch **then**
      restart warmup-cosine schedule for c_s epochs ▷ if enabled
      set current resolution r_s; log `train/scale_resolution`
    for each reconstruction batch **do**
      if $r_s > r_S$ **and** batch is gridded **then**
        coarsen labels to grid r_s by mean-pooling

B.4.5 Configuration 2: evidential active resampling

Every K epochs (study: $K = 2$) and for every task t in the live sampler $\mathcal{T}$, a callback draws n held-out validation points (study: $n = 64$), evaluates the model to obtain $(\mu, \nu, \alpha, \beta)$ per point, and computes the mean identifiable total uncertainty $u_t = \frac{1}{n}\sum_i (w_{\mathrm{St}}^{(i)} + \nu^{(i)-1/2})$. Coverage is

per-task by construction: no task is left at its default weight. The per-task signal is smoothed with an exponential moving average (decay d; study $d = 0.5$), $\bar{u}_t \leftarrow d\,\bar{u}_t + (1-d)\,u_t$ initialised at the first measurement, and converted to a multiplicative adjustment of each task's base sampling probability p_t^0. With mean signal $\langle \bar{u} \rangle = |\mathcal{T}|^{-1} \sum_t \bar{u}_t$, strength exponent s (study $s = 0.5$), and clamps $[m_{\min}, m_{\max}] = [0.2, 5.0]$,

$$m_t = \mathrm{clip}\left(\left(\frac{\bar{u}_t}{\langle \bar{u} \rangle} \right)^s,\, m_{\min},\, m_{\max} \right), \qquad w_t = p_t^0\, m_t, \qquad p_t = \frac{w_t}{\sum_{t'} w_{t'}}. \tag{116}$$

The exponent s tunes aggressiveness ($s = 0$ disables reweighting; smaller is gentler) and the clamps bound how far any task may drift from its prior.

Algorithm 9 Evidential active resampling (every K epochs)

Require: base probabilities p^0, decay d, strength s, clamps $m_{\min}, m_{\max}$, budget n

for each task $t \in \mathcal{T}$ **do**
  draw n validation points; evaluate $(\mu, \nu, \alpha, \beta)$
  $u_t \leftarrow \frac{1}{n} \sum_i (w_{\mathrm{St}}^{(i)} + \nu^{(i)-1/2})$
  $\bar{u}_t \leftarrow d\,\bar{u}_t + (1-d)\,u_t$ ▷ EMA; init $\bar{u}_t = u_t$
$\langle \bar{u} \rangle \leftarrow \mathrm{mean}_t\, \bar{u}_t$
for each task t **do**
  $m_t \leftarrow \mathrm{clip}\left((\bar{u}_t / \langle \bar{u} \rangle)^s, m_{\min}, m_{\max} \right)$; $w_t \leftarrow p_t^0 m_t$
$p_t \leftarrow w_t / \sum_{t'} w_{t'}$; write $\{p_t\}$ to sampler; append audit log

B.4.6 Configuration 3: evidence-regularisation strength

The third factor is the coefficient λ_{ev} on the width-normalised evidence regulariser in eq. (113). It governs how strongly the head is pushed to report evidence commensurate with its standardised error: small λ_{ev} lets the NLL dominate, giving sharper but potentially over-confident predictives, while larger λ_{ev} inflates the penalty on under-supported confidence, widening predictives toward better coverage at some cost in sharpness (hypothesis 3). The study sweeps $\lambda_{\mathrm{ev}} \in \{0.1, 0.3\}$; all other loss coefficients, including the width-cap weight $\kappa = 0.1$, are fixed.

B.4.7 Protocol

All runs share the protocol below; only the three factors vary. Training is mixed-precision `bf16` with a fixed base learning rate and weight decay, and the warmup-cosine schedule is restarted per scale when the curriculum is active and run as a single cycle otherwise; gradient accumulation realises the target effective batch size. Each epoch draws a fixed mixture of step families (reconstruction, geo-alignment, country-alignment, transport), with the reconstruction family carrying the NIG objective of eq. (113), and a modality-balance fraction pins the country/gridded split of reconstruction steps so the data diet is constant across cells. The spatiotemporal alignment objective is held at the winner configuration of the preceding alignment study, so alignment is a constant rather than a factor.

The primary, pre-registered selection statistic is the reconstruction validation root-mean-squared error on standardised held-out-year targets,

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\mu_i - y_i)^2}, \tag{117}$$

where the point prediction is the NIG location μ; lower is better. We report it overall and split by modality (gridded versus country), with the prediction-target correlation as a secondary diagnostic. Because the model is evidential, calibration is reported as a co-equal family of diagnostics that does not enter selection: the negative log-likelihood (NLL), expected calibration error (ECE), the Kolmogorov-Smirnov statistic of the probability-integral-transform residuals against uniformity (KS-PIT), sharpness as the mean predictive width, and the empirical coverage of the nominal 50/95% predictive intervals. Each metric is summarised per cell as a mean ± standard deviation over the three seeds.

B.4.8 Results

All 24 runs completed and the design is fully balanced (8 cells × 3 seeds) with no missing or errored runs. The eight cells of the $2 \times 2 \times 2$ design are shown over their three factor axes in Fig. 52. The confirmed winner is **no curriculum, active learning on,** $\lambda_{\mathrm{ev}} = 0.3$, reaching a selection RMSE of 0.307 ± 0.048 and a calibration NLL of 1.50 ± 0.06 (Table 21). It is simultaneously the most accurate and has by far the lowest calibration NLL (the next-lowest is 18.1 and the worst exceeds 330), though on the sharper ECE and KS–PIT statistics it trades some quality for this NLL and coverage advantage (discussed below). It is rank-one on RMSE in all three seeds.

			RMSE		corr		NLL		RMSE by modality	
SC	AL	ε	mean±sd	Δ^*	mean±sd	Δ^*	mean±sd	Δ^*	grid	ctry
off	on	0.3	0.307 ± 0.048	–	0.738 ± 0.031	–	1.50 ± 0.06	–	0.303	0.316
on	on	0.3	0.329 ± 0.056	+0.021	0.721 ± 0.042	-0.017	18.08 ± 23.38	+16.58	0.335	0.326
off	on	0.1	0.331 ± 0.051	+0.024	0.740 ± 0.038	+0.002	85.71 ± 66.46	+84.21	0.300	0.366
off	off	0.3	0.369 ± 0.054	+0.062	0.703 ± 0.032	-0.034	88.19 ± 64.04	+86.69	0.360	0.382
on	on	0.1	0.371 ± 0.053	+0.063	0.700 ± 0.037	-0.037	204.44 ± 149.66	+202.94	0.352	0.392
off	off	0.1	0.381 ± 0.051	+0.074	0.695 ± 0.027	-0.043	332.35 ± 236.97	+330.84	0.364	0.401
on	off	0.3	0.392 ± 0.055	+0.084	0.688 ± 0.035	-0.050	71.79 ± 69.12	+70.28	0.374	0.413
on	off	0.1	0.405 ± 0.057	+0.098	0.684 ± 0.037	-0.053	208.45 ± 184.89	+206.94	0.380	0.434

∗ Δ vs. best config (shaded); lower RMSE/NLL and higher corr are better.

Table 21: **Summary leaderboard.** Per cell, RMSE, correlation, and calibration NLL are reported as mean ± s.d. over the three seeds, each with the increment Δ^* relative to the best (shaded) configuration, followed by the RMSE split by modality (gridded versus country). Rows are ordered by RMSE; lower RMSE and NLL and higher correlation are better.

Main effects. Active learning is the dominant beneficial factor, with $\widehat{\Delta}_{\mathrm{RMSE}} = -0.052$ (95% CI $[-0.061, -0.043]$), and it lowers the mean calibration NLL from 175 to 77. The

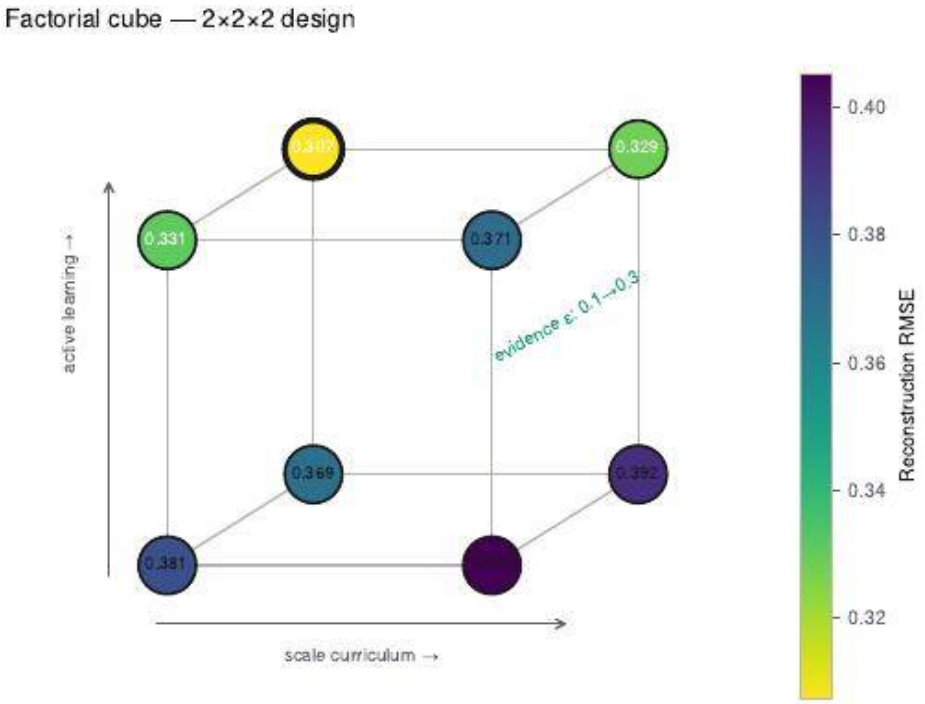

Figure 52: **Factorial design cube.** The eight vertices of the $2 \times 2 \times 2$ design over the scale-curriculum, active-learning, and evidence-coefficient axes, coloured by mean selection RMSE (darker is lower). The best vertex (`SC off, AL on,` ε=0.3) is ringed; moving along the active-learning axis lowers RMSE, while activating the curriculum raises it.

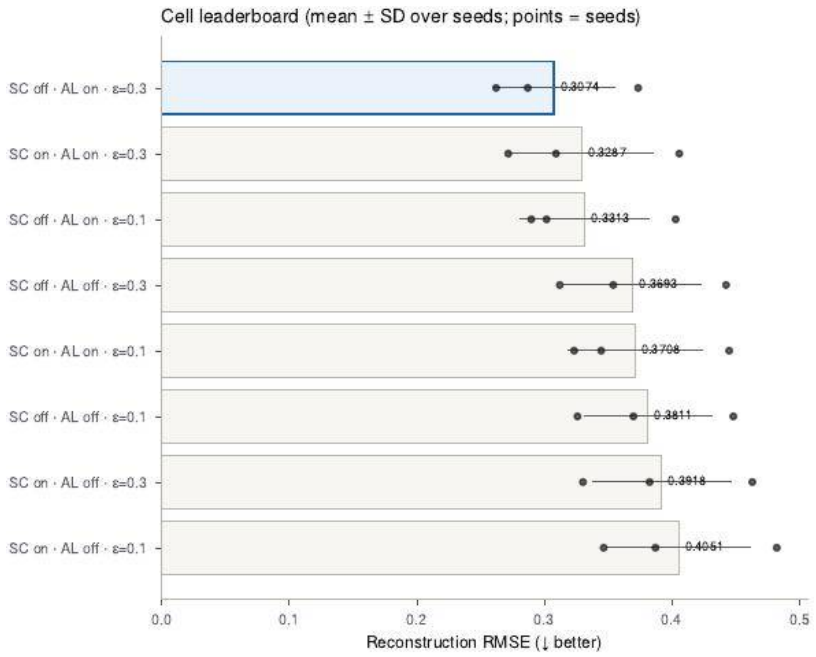

Figure 53: Mean selection RMSE ± s.d. over seeds, with the individual seed points overlaid. The best cell (`SC off, AL on,` ε=0.3) is separated from the field, and every active-learning cell outranks its no-active-learning counterpart.

evidence coefficient improves accuracy modestly ($\widehat{\Delta}_{\mathrm{RMSE}} = -0.023$) and is the principal driver of calibration. The scale curriculum, *increases* error ($\widehat{\Delta}_{\mathrm{RMSE}} = +0.027$; marginal RMSE 0.374 on versus 0.347 off).

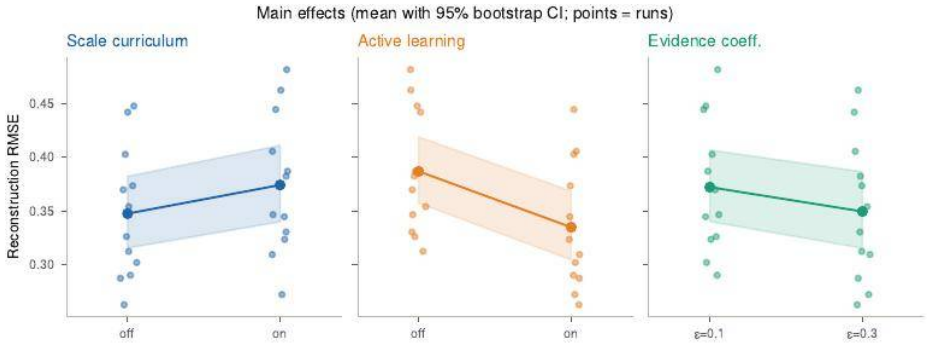

Figure 54: **Main effects.** Per-factor level means with 95% bootstrap confidence intervals and run points overlaid. Active learning and the evidence coefficient lower RMSE; the curriculum raises it.

Calibration and uncertainty. The accuracy-calibration relationship (Fig. 56) places the best cell in the lower-left corner (low RMSE *and* low NLL), with the Pareto front populated entirely by active-learning cells. The mechanism is visible in the calibration diagnostics (Table 22) and the uncertainty decomposition (Fig. 57): raising λ_{ev} from 0.1 to 0.3 widens

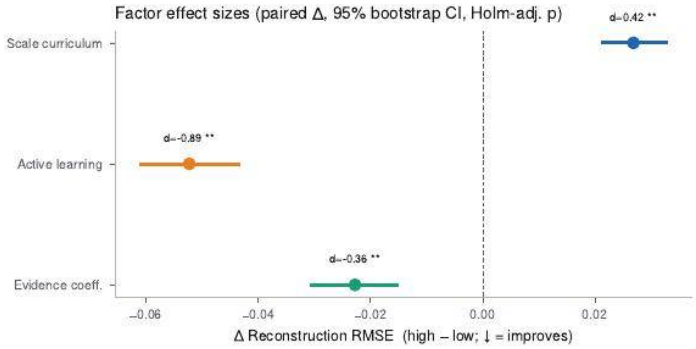

Figure 55: **Paired factor effect sizes.** Mean paired high−low RMSE contrast per factor with 95% bootstrap confidence intervals and Holm-adjusted significance. Active learning has the largest favourable effect; the curriculum effect is unfavourable.

the predictive (mean width $0.50 \to 1.47$) and raises both the aleatoric and epistemic proxies, removing the catastrophic over-confidence responsible for the large NLLs. This is a trade-off rather than a uniform improvement, as anticipated in hypothesis 3: expected calibration error is essentially flat across factors (ECE ≈ 0.17–0.22), the KS-PIT statistic mildly worsens at $\lambda_{\text{ev}} = 0.3$ ($0.34 \to 0.44$), and 95% coverage drifts from 0.980 to 0.997, that is, toward over-coverage. The NLL gains come from suppressing the over-confident tail, not from uniformly sharper-and-correct predictions. This may partly reflect unconverged training, as computational constraints did not allow these three-seed configurations to run for a full training budget.

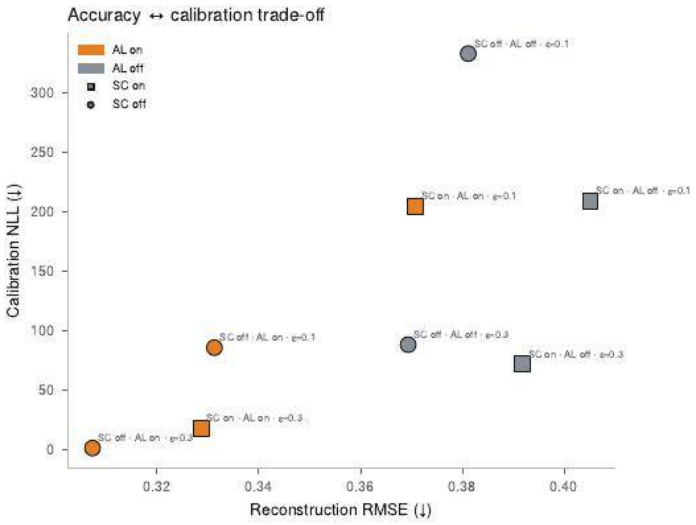

Figure 56: **Accuracy-calibration trade-off.** Selection RMSE against calibration NLL across cells. The Pareto front is composed entirely of active-learning cells and the best cell dominates the lower-left corner.

Cell	NLL	ECE	KS-PIT	Sharp	Cov50	Cov95
SC- AL+ ec=0.3	1.50	0.216	0.458	1.80	0.99	1.00
SC+ AL+ ec=0.3	18.08	0.212	0.433	1.44	0.98	1.00
SC- AL+ ec=0.1	85.71	0.188	0.365	0.81	0.88	0.99
SC- AL- ec=0.3	88.19	0.211	0.425	1.24	0.98	1.00
SC+ AL+ ec=0.1	204.44	0.168	0.323	0.47	0.88	0.97
SC- AL- ec=0.1	332.35	0.172	0.328	0.28	0.87	0.97
SC+ AL- ec=0.3	71.79	0.211	0.426	1.41	0.99	1.00
SC+ AL- ec=0.1	208.45	0.180	0.348	0.42	0.91	0.99

Table 22: **Calibration diagnostics per cell.** NLL, ECE, KS-PIT, sharpness (mean predictive width), and 50/95% coverage. The winning cell has a NLL more than an order of magnitude below the next-best cell, while ECE is nearly constant across the grid.

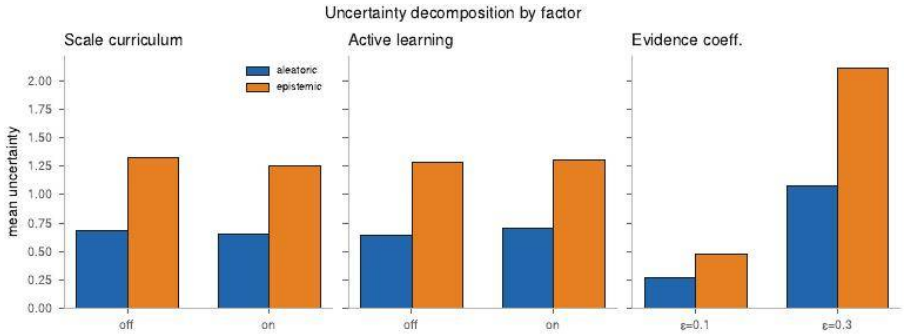

Figure 57: **Uncertainty decomposition by factor.** Aleatoric (w_{St}) and epistemic ($1/\sqrt{\nu}$) proxies aggregated over each factor level; the evidence coefficient is the dominant control on predictive width.

Ranking robustness and mechanism. The winning cell is rank-one in all three seeds and the cell ordering is stable across seeds (Fig. 58).

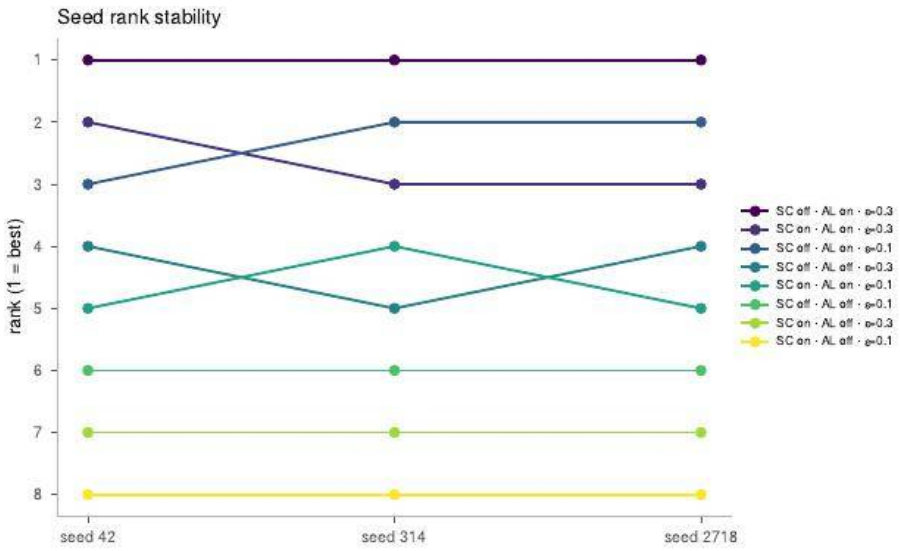

Figure 58: **Per-seed rank trajectories.** Each line tracks a cell's rank across the three seeds; the best cell is rank-one in every seed and the overall ordering is stable.

B.4.9 Conclusion

This ablation selects the training procedure for the fixed evidential reconstruction model. Active resampling is the single most valuable lever, lowering reconstruction error ($\widehat{\Delta}_{\mathrm{RMSE}} = -0.052$, $d = -0.89$) and roughly halving the calibration NLL, and it supports hypothesis 2 on both axes. A larger evidence coefficient ($\lambda_{\mathrm{ev}} = 0.3$) is a cheap and consistent calibration gain that also modestly improves accuracy, supporting hypothesis 3, with the nuance that the benefit is concentrated in the NLL tail rather than in ECE or PIT uniformity. The two configurations act additively, so the predicted active-learning×curriculum interaction of hypothesis 4 is not supported. The scale curriculum, by contrast, significantly *hurts* reconstruction ($\widehat{\Delta}_{\mathrm{RMSE}} = +0.027$), refuting hypothesis 1; candidate explanations are that the mean-pool coarsening discards high-frequency signal the fixed budget cannot recover at native resolution, that the per-scale learning-rate restart interacts poorly with the remainder of the schedule, or that the standardised-target regime renders the coarse stages nearly redundant. Another explanation is that the hybrid location encoder architecture, which includes a multi-resolution hash encoding, is designed to handle high resolution data from the beginning, and expending training budget in lower resolutions ultimately hurts the capacity of the model to encode high-resolution signals.

The chosen configuration is ***active learning on, evidence coefficient*** $\lambda_{\mathrm{ev}} = 0.3$***, and no scale curriculum****.*

B.5 Backward confirmation: the two alignment objectives

B.5.1 Motivation and scope

Every ablation up to this point is a step of a *forward*, greedy selection: a component is isolated, compared against alternatives while the rest of the model is held fixed or disabled, and the winner is promoted into the next stage. Forward selection is the only tractable way to search a space this large, but it has a known blind spot. A component chosen while its collaborators were absent may be redundant once they are present, and a component that looked decisive in isolation may contribute nothing at the end of the chain. Greedy selection cannot see this, because it never re-tests an early choice against the final model. This study closes that gap for the two components on which the paper's central architectural claim rests: the two *alignment* objectives. TerraNova carries two native geometries and two losses that tie its representation:

Internal alignment The country–location objective (§A.9.3; forward study in §B.2) is the *only* route by which a country query is matched to coordinates. Nothing else in the architecture connects the two geometries.

External alignment The geospatial objective (§A.9.2; forward study in §B.1.19) ties the location representation to three pretrained geo-foundation embeddings: SatCLIP, GeoCLIP and Copernicus.

The first is a claim about *multimodality*: two geometries, one bridge. The second is a claim in the spirit of the platonic-representation hypothesis (Huh et al., 2024): that a representation improves by being pulled towards the geometry other strong models have already found. Forward selection established that each objective helps when added. It cannot say whether either still earns its place in the finished model, nor what exactly each one buys.

A crossed pair. We train two knockouts from an otherwise identical production recipe:

	internal (country–location)	external (geo)
`-country`	**off**	on
`-geo`	on	**off**

The two runs agree on every other hyperparameter and differ only in which alignment objective each drops. This is deliberate, and it is what makes the study self-contained: **each knockout *retains* the objective the other removes, so each is the other's matched control.** No model in which nothing is removed is required, and the shared background cancels exactly in the contrast.

The estimator. For a capability c measured on tasks t, we form the single contrast

$$\widehat{\Delta}(c) \;=\; \mathrm{agg}_t\Big[\, y_{\texttt{-geo}}(t,c) \;-\; y_{\texttt{-country}}(t,c) \,\Big], \tag{118}$$

where $y_m(t,c)$ is the score of run m, oriented so that higher is better. Because the two runs are identical apart from which alignment objective they drop, *any* departure of $\widehat{\Delta}(c)$ from zero is attributable to the alignment objectives, and its sign says which one:

- $\widehat{\Delta}(c) > 0$ — the model lacking *geo* alignment is the better one, so c depends on the *country–location* objective;
- $\widehat{\Delta}(c) < 0$ — the model lacking *country* alignment is the better one, so c depends on the *geo* objective;
- $\widehat{\Delta}(c) \approx 0$ — c depends on neither, which is the specificity control.

One contrast, read across capabilities, therefore tests both objectives at once. This is a 2×2 factorial observed at its two off-diagonal cells. It does not identify either objective's main effect against a both-on model but it does identify a *double dissociation*, which is precisely what the hypotheses below predict and a strictly stronger causal statement than "removing it makes things worse".

B.5.2 Hypotheses

Hypothesis 5 (Coupling)**.** The country–location objective is the only route by which a country query is matched to coordinates. Removing it should not merely degrade the cross-geometry capabilities but *sever* them: coordinate-to-country retrieval should fall to chance, and national-to-gridded downscaling to the flat-national null (a within-country correlation of zero, which is by construction the score obtained by painting each country its own national mean). Purely within-geometry skill should be untouched.

Hypothesis 6 (Support)**.** The three external teacher banks are *land-only by construction*: they are precomputed at 500k land coordinates each (`satclip_land_500k`, `geoclip_land_500k`, `copernicus_land_500k`), so the geospatial objective is never evaluated at an ocean coordinate. If external alignment works by importing structure that the teachers possess, its benefit must appear exactly on the teachers' support and vanish off it. Removing it should therefore cost representation quality on *land* targets and nothing on *ocean* targets. The ocean targets are a support-matched placebo that the design supplies for free.

Together the two hypotheses predict a double dissociation: each objective's removal should damage the capability family that objective serves, and leave the other family alone.

B.5.3 What we measure — and one probe we deliberately do not

Cross-geometry coupling (head-free). The training-time validation harness records the country–location diagnostics directly from the model's *native* embeddings: a country embedding and the mean location embedding of that country are compared by cosine similarity, with no learned projection anywhere in the path. We report coordinate-to-country top-1 retrieval, country-to-location top-1 retrieval, and the raw country–location cosine, over a fixed pool of 22 countries (so chance is $1/22 = 0.045$). These are aggregate diagnostics over a fixed evaluation set rather than per-task quantities, and are reported as point estimates without intervals.

Downstream capabilities. Each checkpoint is then evaluated on the paper's own protocols. *Representation quality*: held-out field reconstruction from a frozen embedding probe at label budgets $\{256, 1024, 4096\}$, over 11 targets (9 land, 2 ocean), each averaged over three

evaluation seeds. *Cross-geometry transfer*: national-to-gridded downscaling on four leakage-free targets, scored by the recovered within-country structure, over two evaluation seeds.

A probe we deliberately do not use. The obvious way to score the geo-alignment knockout is the alignment retrieval accuracy against the external targets, and it does collapse: from 0.42 to 0.000. We do *not* report it, because it is tautological. That probe projects the representation through a learned alignment head, and the alignment head is trained by nothing except the geospatial-alignment loss. With the loss switched off the head remains at its random initialisation, so retrieval is at chance *by construction*, whatever the representation underneath is like. The number measures the probe, not the model. Every quantity we do report is read either from head-free native embeddings or from a downstream task that never touches an alignment head.

Statistics. Compute permitted a single training seed (42) per knockout, so a seed-paired test is unavailable and we take the *task* as the unit of analysis. For the reconstruction probe we report the per-task contrast of Eq. 118 at every budget, the mean over targets within each domain, and the count of targets on which the sign holds; the domain split is pre-registered by Hypothesis 6 and is not chosen after the fact. For downscaling we report the per-target mean over evaluation seeds. Given 9 land targets, 4 downscaling targets and 2 ocean targets, we lean on effect sizes, sign consistency and the distance from a *known null* (chance retrieval; zero within-country correlation) rather than on p-values, which at these sample sizes would be uninformative either way.

B.5.4 Results

Cutting the internal bridge severs both cross-geometry capabilities. Removing the country–location objective drives coordinate-to-country retrieval from 0.730 to **0.043**, against a chance level of 0.045; that is, exactly to chance (Table 23, Fig. 59a). Country-to-location retrieval falls from 0.955 to 0.091, and the raw country–location cosine from 0.761 to 0.198. The two geometries are not degraded; they are *decoupled*. National-to-gridded downscaling collapses in step (Fig. 59b, Table 24): the within-country correlation falls to -0.02, $+0.04$, $+0.03$ and -0.00 on GPP, LAI, fAPAR and tree cover respectively; zero on all four, meaning the model recovers no sub-national structure whatsoever and does no better than assigning every cell its country's national mean. The control retains $+0.23$ to $+0.39$ on the same four targets. This supports Hypothesis 5 in its strong form. Crucially, the same model is untouched *within* a geometry: reconstruction RMSE is 0.276 against the control's 0.272, and calibration ECE 0.173 against 0.172.

External alignment pays off exactly on its teachers' support. Removing the geospatial objective costs $\Delta R^2 = -0.149$ averaged over the nine *land* targets at the largest label budget, and it does so on *all nine* (Fig. 59c, Table 25). from -0.084 on surface $PM_{2.5}$ up to -0.219 on photovoltaic potential and -0.211 on leaf-area index. On the two *ocean* targets, which lie off the teachers' support, the same knockout costs -0.014: nothing. The dissociation is not an artefact of a single operating point. It holds at every label budget, land $-0.122/-0.148/-0.149$ at 256/1024/4096 points, 9/9 negative each time, against ocean $+0.036/+0.009/-0.014$, whose sign is not even stable (Fig. 60). This supports Hypothesis 6:

alignment to external representations improves this model's own representation precisely where those representations have something to say, and not elsewhere.

The two effects are dissociated. Each objective's removal damages its own capability family and spares the other's. The model without country alignment sits at chance on cross-geometry retrieval but is *indistinguishable* from its control on land reconstruction; the model without geo alignment loses 0.149 R^2 on land but retains coordinate-to-country retrieval at 0.730, sixteen times chance. Because the two runs share everything except which objective they drop, this double dissociation cannot be explained by a difference in recipe, capacity, or training budget. The two alignment objectives do genuinely distinct work.

Probe	−geo align	−country align	chance
cross-geometry (head-free native embeddings)			
coordinate → country	0.730	**0.043**	0.045
country → location	0.955	**0.091**	0.045
country ↔ location cosine	0.761	**0.198**	–
within-geometry (unaffected)			
reconstruction RMSE	0.272	0.276	–
calibration ECE	0.172	0.173	–

Table 23: **Cross-geometry coupling probes.** Measured on the model's native country and location embeddings by cosine similarity, with no learned projection in the path, over a fixed pool of 22 countries (chance = 1/22). Removing the country–location objective takes coordinate-to-country retrieval to chance, while leaving within-geometry reconstruction and calibration unchanged. `-geo`, which retains the objective, is the control.

Target	−geo align	−country align	Δ
GPP	+0.376	**−0.023**	−0.399
LAI	+0.385	**+0.039**	−0.346
fAPAR	+0.339	**+0.031**	−0.308
Tree cover	+0.228	**−0.004**	−0.232

Table 24: **National-to-gridded downscaling.** Within-country correlation, averaged over two evaluation seeds; 0 is the flat-national null, the score obtained by assigning every cell its country's national mean. Removing the country–location objective removes the capability entirely. This contrast is if anything *conservative*: the control, `-geo`, is itself somewhat weakened on this task by the loss of external alignment.

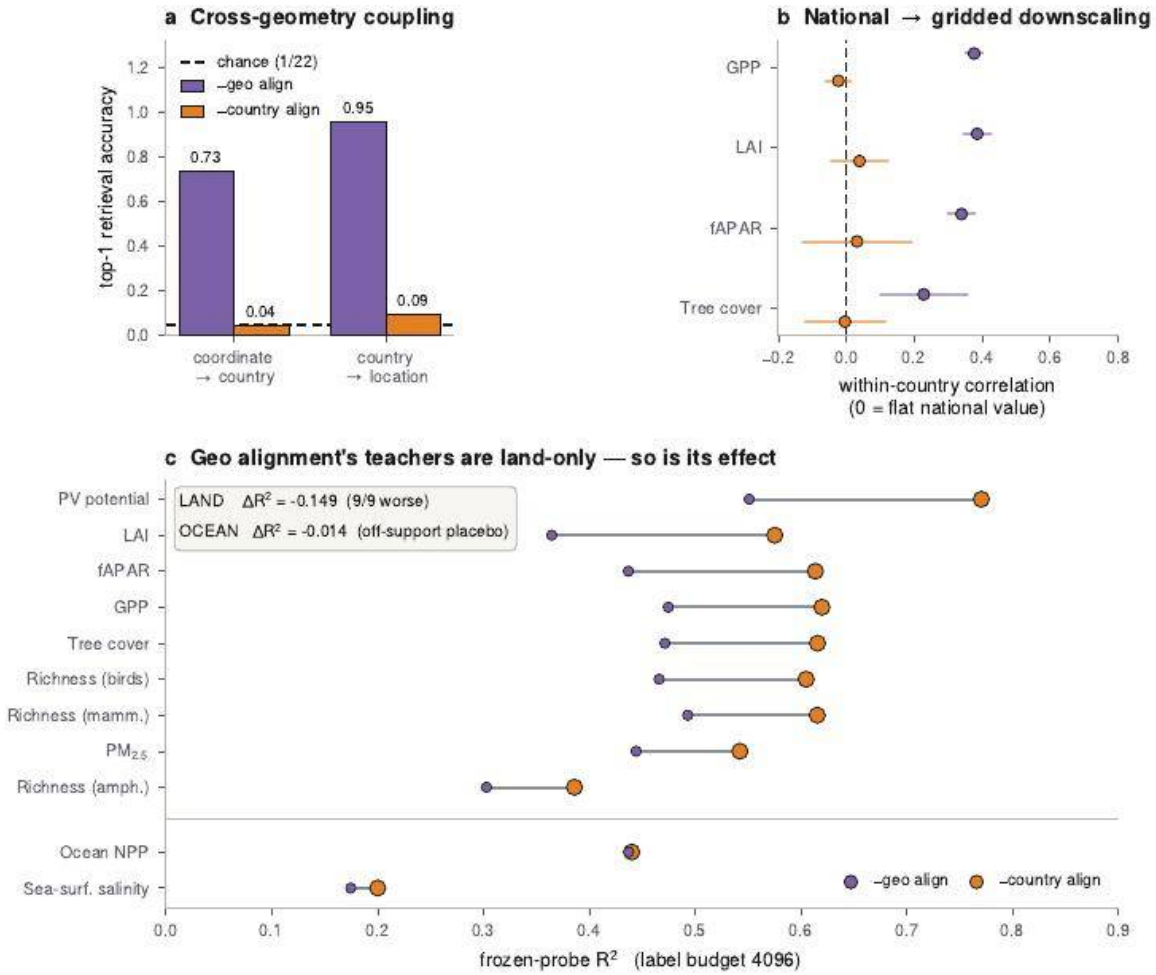

Figure 59: **The two alignment objectives do distinct work.** A crossed pair of knockouts: `-country` drops the internal country–location objective and keeps the external one; `-geo` does the reverse. They are identical in every other respect, so each is the other's control. **(a)** Cross-geometry retrieval, measured on the model's native embeddings with no learned probe. Removing the country–location objective drops coordinate-to-country retrieval to 0.043 against a chance level of $1/22 = 0.045$: the geometries are decoupled, not degraded. **(b)** National-to-gridded downscaling. The same knockout falls to a within-country correlation of zero on all four targets while the control retains 0.23 to 0.39. **(c)** Frozen-probe reconstruction R^2. The geo-alignment knockout loses 0.149 R^2 on all nine *land* targets and nothing (-0.014) on the two *ocean* targets. The three external teacher banks are land-only by construction, so the ocean targets act as a support-matched placebo: external alignment helps exactly where its teachers have support. Each objective damages its own capability family and spares the other's.

	budget 256			budget 1024			budget 4096		
Task	−geo align	−country align	Δ	−geo align	−country align	Δ	−geo align	−country align	Δ
land									
fAPAR	0.399	0.539	−0.140	0.409	0.588	−0.179	0.437	0.614	−0.177
LAI	0.277	0.475	−0.197	0.331	0.553	−0.222	0.365	0.575	−0.211
PV potential	0.526	0.721	−0.195	0.549	0.758	−0.209	0.551	0.771	−0.219
Richness (amph.)	0.318	0.319	−0.001	0.249	0.310	−0.061	0.303	0.386	−0.083
Richness (birds)	0.425	0.516	−0.090	0.446	0.586	−0.140	0.466	0.605	−0.139
Richness (mamm.)	0.407	0.535	−0.127	0.481	0.600	−0.119	0.493	0.616	−0.122
$PM_{2.5}$	0.373	0.445	−0.072	0.420	0.526	−0.106	0.444	0.542	−0.098
GPP	0.397	0.553	−0.156	0.460	0.612	−0.151	0.475	0.620	−0.145
Tree cover	0.431	0.555	−0.123	0.470	0.610	−0.140	0.472	0.616	−0.144
mean Δ			**−0.122**			**−0.148**			**−0.149**
ocean									
Ocean NPP	0.387	0.331	+0.056	0.423	0.397	+0.025	0.437	0.440	−0.003
Sea-surf. salinity	0.102	0.086	+0.016	0.169	0.175	−0.007	0.175	0.200	−0.025
mean Δ			**+0.036**			**+0.009**			**−0.014**

Table 25: **Frozen-probe reconstruction R^2 by domain and label budget.** Δ = `-geo` − `-country`, so a negative value means removing geo alignment is worse. Each entry is a mean over three evaluation seeds. Removing external alignment costs R^2 on every land target at every budget, and nothing at sea.

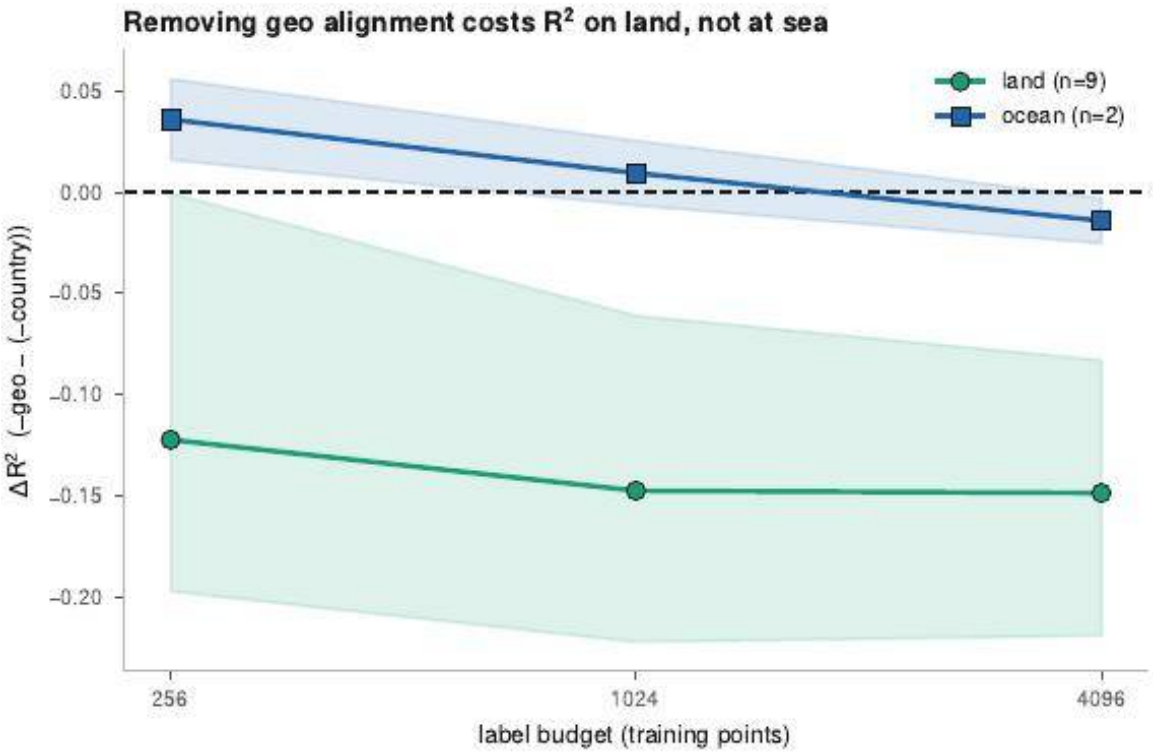

Figure 60: **The land/ocean dissociation is not an artefact of one operating point.** Mean ΔR^2 (Eq. 118) for the geo-alignment knockout, by domain, at each label budget; the band spans the per-target range. The land band lies wholly below zero at every budget and never overlaps the ocean band, whose sign is not even stable.

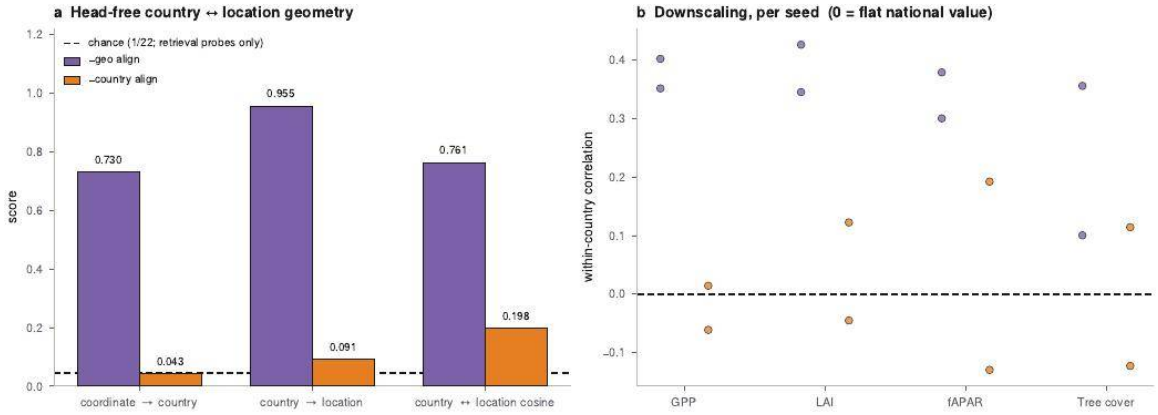

Figure 61: **Coupling probes and downscaling, in full. (a)** All three head-free country–location probes, including the raw cosine similarity, which needs no retrieval pool and therefore no chance baseline: it falls from 0.761 to 0.198. **(b)** Downscaling, per evaluation seed rather than averaged.

B.5.5 Limitations

First, the crossed design identifies the *contrast* between the two alignment objectives, not either one's absolute marginal contribution: no model was trained in which neither objective is removed, so we report relative degradations and never an absolute effect on the finished model. Second, each configuration was trained *once* (seed 42); the spread we report is over tasks and evaluation seeds, not over training seeds, and a difference consistent across tasks may still reflect a single lucky or unlucky initialisation. Third, the off-support placebo rests on only *two* ocean targets; the direction is unambiguous and holds at all three budgets, but the ocean estimate is correspondingly imprecise. Fourth, the coupling probes are aggregate diagnostics over a fixed 22-country pool, so country-to-location top-1 is quantised in steps of 1/22; the coordinate-to-country probe, which is computed over many locations, and the raw cosine, which is continuous, carry the weight of the claim.

B.5.6 Conclusion

Backward confirmation on the final model supports both alignment objectives, and shows that they earn their place for different reasons. The **internal country–location alignment is load-bearing**: it is the only bridge between TerraNova's two native geometries, and cutting it does not weaken the cross-geometry capabilities but abolishes them while leaving everything within a geometry untouched. The **external geospatial alignment improves the representation itself**, and it does so precisely on the support of the models it borrows from: removing it costs R^2 on every land target at every label budget and nothing at all on the ocean targets its teachers never saw. The first result is what makes TerraNova multimodal rather than two models in a trenchcoat; the second is a concrete, falsifiable instance of the platonic-representation hypothesis, complete with the placebo condition that a weaker version of the claim would fail.

Appendix C. Extended Results

C.1 Static reconstruction benchmark

Figure 62 showseleven held-out targets, nine read-outs and seven label budgets, the six of the main text together with a 128-label column below its range, each cell a mean over three seeds. At the largest budget TerraNova's task-adapted read-out is at or above the best baseline on every target, so its mean rank is 1.00, ahead of RANGE+ at 2.18 and GT-Loc at 4.27; a linear probe on the same frozen embedding is seventh of the nine, at a mean rank of 6.00, below five of the published encoders. The comparison is not capacity-matched: the adapted read-out fits a fresh task row and rank-4 MiSS residuals at 416,000 trainable parameters per task, whereas the published encoders are read by a probe on frozen features, so the margin is a statement about task-conditioned decoding rather than about the linear separability of the embedding.

Splitting the regression targets by domain (Fig. 63) separates a property of a representation from a property of its training data. Every image-trained encoder loses skill when it moves from land to ocean, by between 0.12–0.60 R^2, whereas TerraNova loses 0.01 with a seed interval that spans zero, and the random-Fourier null loses nothing at all because it has no learned land prior to lose. The ordering is the same at every label budget and no image-trained encoder closes its deficit at the largest one, so it is a property of what those encoders saw in training rather than a small-sample artefact.

The same protocol run on categorical targets asks a harder question, since a class map has to be right about the boundaries and not merely about the level. Reconstructing the 847 terrestrial ecoregions from coordinates alone (Fig. 64), and under that figure's own sampling, the task-adapted read-out reaches 0.68 accuracy, level with RANGE+, against 0.59 for the frozen probe of the same model and 0.53 for the random-Fourier positional null.

C.2 Temporal transfer

The temporal battery is 29 leakage-clean national indicators, each split by holding out its earliest and latest years, and Figure 65 reads it indicator by indicator and horizon by horizon rather than pooled. The two directions behave differently. Backcasting, TerraNova matches the fitted linear trend throughout, sitting a little above it at every horizon within overlapping bands and falling from 0.81 one year before the training window to 0.60 six years before it while the trend falls from 0.73 to 0.59. Forecasting, it leads at one year ahead (0.78 against 0.70) and the trend overtakes it from two years out, reaching 0.61 against 0.49 at six years ahead, and at that horizon the trend is ahead on 21 of the 29 indicators. Averaged over the battery the adaptation-free probe on the frozen embedding never exceeds 0.38 in either direction, so what carries the temporal skill is the adapted read-out and not the raw embedding. Rolling the state forward through frozen dynamics is worse than predicting the target year directly (0.605 against 0.625 pooled, better on 10 of 29 indicators); adapting the dynamics along with the read-out recovers a small gain (0.639, better on 18 of 29); fitting the two jointly and decoding directly lands between them (0.628, better on 13 of 29). We therefore report the operator as a component the architecture supports rather than as a source of the temporal result. Fertiliser use per cropland turns sharply negative two and three years into the forecast, tourist arrivals does the same at three and four years, and at

Figure 62: **The full benchmark cube.** **(a)** Target × method skill at 8,192 labelled points, R^2 above the rule and accuracy below it, with the two TerraNova read-outs outlined. **(b)** Mean rank across targets against label budget. **(c)** TerraNova's task-adapted read-out against the best baseline for each target, with the identity of that baseline and the margin in the right-hand gutter. Dots are means over three seeds and whiskers are 95% intervals; the three species-richness targets are leakage-gated and excluded throughout.

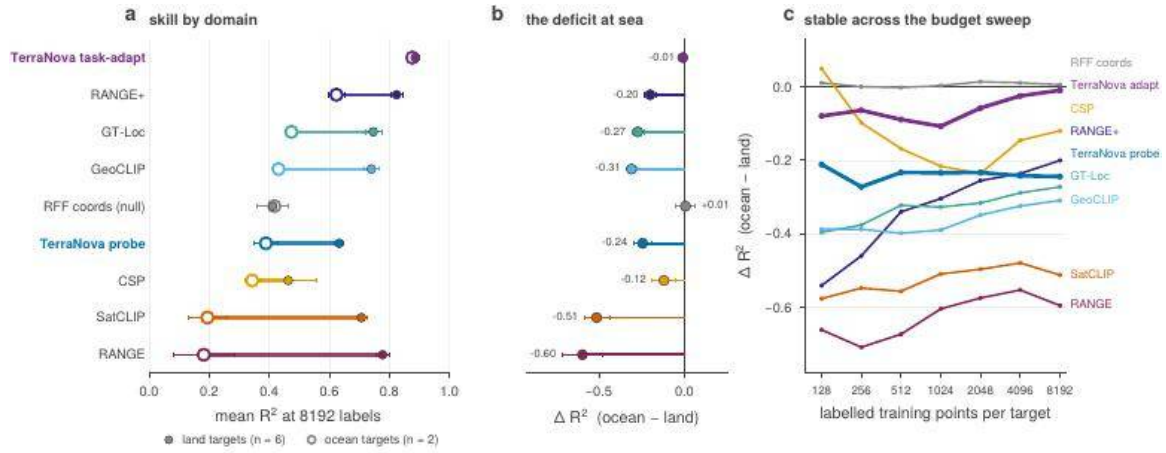

Figure 63: **Land against ocean.** Regression targets only, R^2 being the one metric both domains share. **(a)** Per-method land mean against ocean mean. **(b)** The signed gap Δ = ocean − land with its seed interval. **(c)** That gap against label budget. The random-Fourier read-out is a positional null with no learned land prior, which is why it shows no deficit at all; TerraNova's task-adapted read-out is the only learned read-out whose deficit is also indistinguishable from zero.

six years ahead the fitted trend is well in front of the model on both. Those are indicators that move sharply and idiosyncratically within countries over the period, which is the regime in which a model that infers the level from geography and covariates has least to work with.

C.3 Dense fields from sparse measurements

The main text pools the gridded targets into one degradation curve, which answers how adaptation compares with interpolation on average but not on which field and at which measurement density. Figure 67 separates the two. Where measurements are dense the two approaches are indistinguishable, with every margin inside ±0.03 and classical interpolation ahead on the two smoothest ocean fields; as the lattice thins the margins open in the order of how much structure a field carries, reaching +0.26 on tree cover, +0.16 on leaf area index and +0.14 on fAPAR at one measured cell in 4,096. One leakage-free target ends up on the other side, sea-surface salinity at −0.08, a nearly featureless field of which a 1-in-4,096 lattice leaves too little to anchor an adaptation. The per-target margins here are Pearson correlations over every valid native cell, the measured ones included, rather than the held-out R^2 of Fig. 14b: interpolation reproduces a measured cell exactly, so this scoring is generous to it and most generous at the dense end, and the pooled held-out numbers of §6.3 of the main text are the ones to read as skill.

Figures 68 to 75 are the reconstructions themselves, one page per variable and one row per density, so that the numbers above can be checked against what the fields look like. Held-out R^2 over the ladder from one cell in sixteen to one in 4,096 falls from 1.00 to 0.95 for photovoltaic potential, 0.96 to 0.91 for fAPAR, 0.95 to 0.84 for leaf area index, 0.96 to 0.83 for terrestrial GPP, 0.98 to 0.79 for ocean net primary productivity, 0.91 to 0.71 for tree cover, 0.99 to 0.60 for surface $PM_{2.5}$ and 0.99 to 0.28 for sea-surface salinity. The error maps

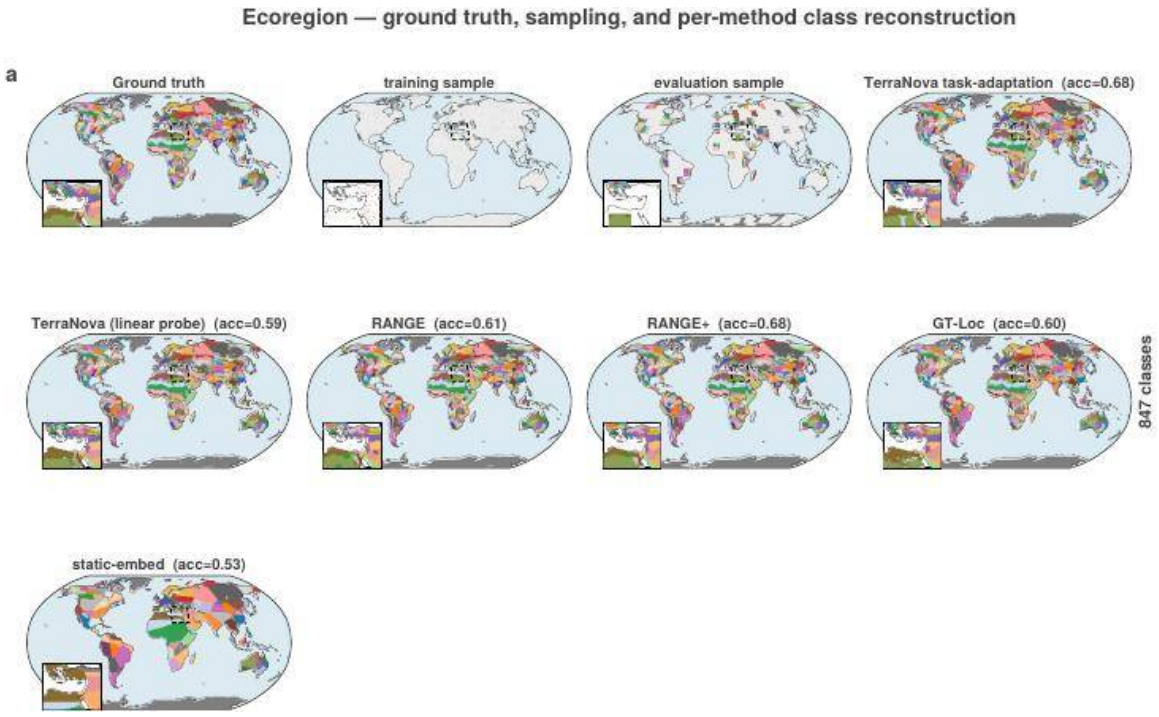

Figure 64: **Categorical reconstruction.** The 847 terrestrial ecoregions reconstructed from coordinates: ground truth, the training and evaluation samples, and the argmax class map of six read-outs with its held-out accuracy, namely TerraNova's task-adapted and frozen-probe read-outs, RANGE, RANGE+, GT-Loc and the random-Fourier positional null. Colours are arbitrary class identities, so the figure is read by comparing boundaries against the ground-truth panel. This is a separate sweep from the benchmark cube of Fig. 62, with its own sampling, so the accuracies here differ from that figure's ecoregion row.

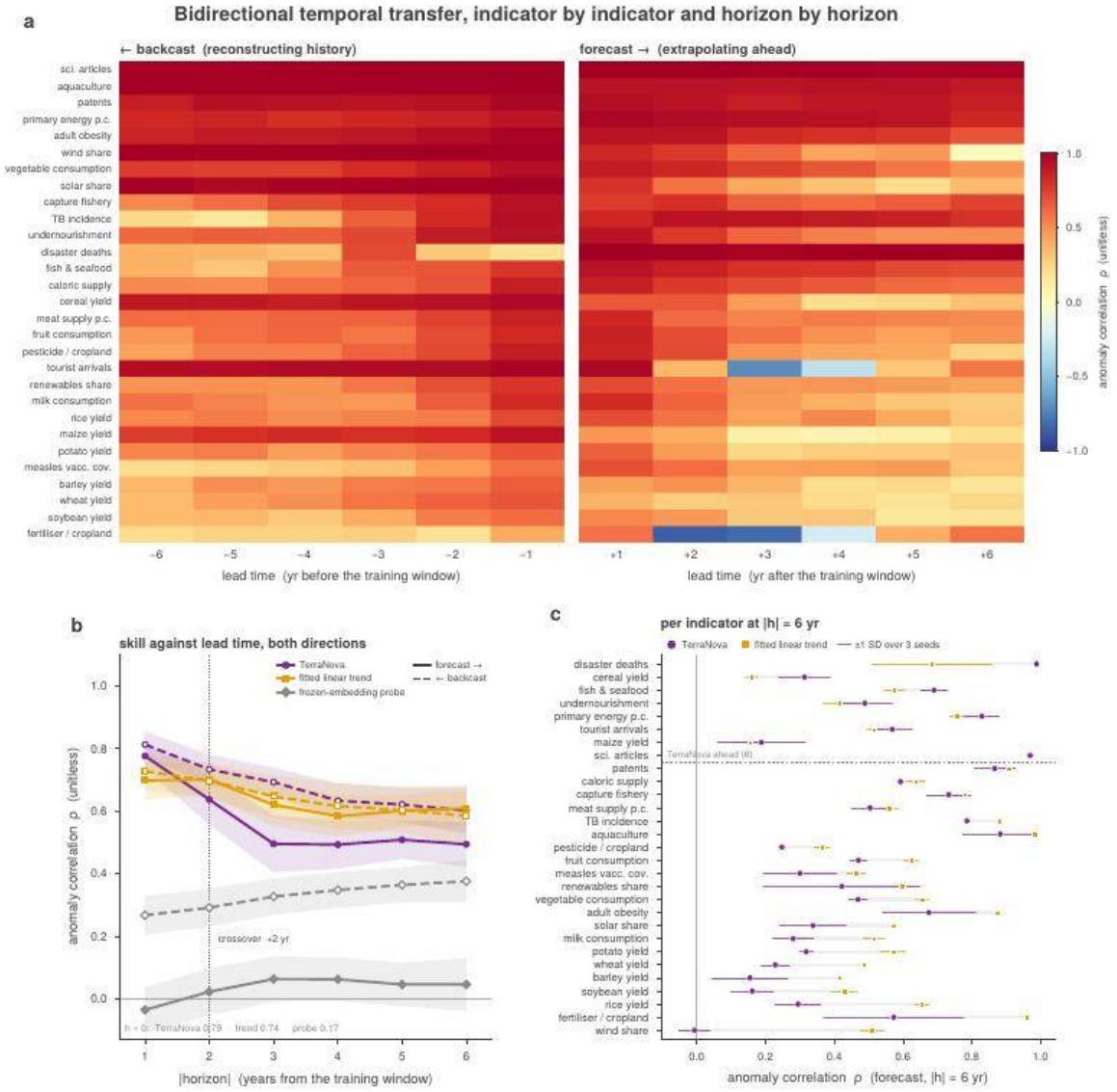

Figure 65: **Temporal skill, per indicator and per horizon. (a)** Anomaly correlation of the adapted read-out for each of the 29 indicators at each signed horizon, backcast and forecast side by side. **(b)** The pooled decay curves for TerraNova, a fitted linear trend and a frozen-embedding probe in both directions, with a 95% band over indicators and seeds; the dotted rule marks where the trend overtakes the model on the forecast side. **(c)** Per-indicator comparison with the trend at six years ahead.

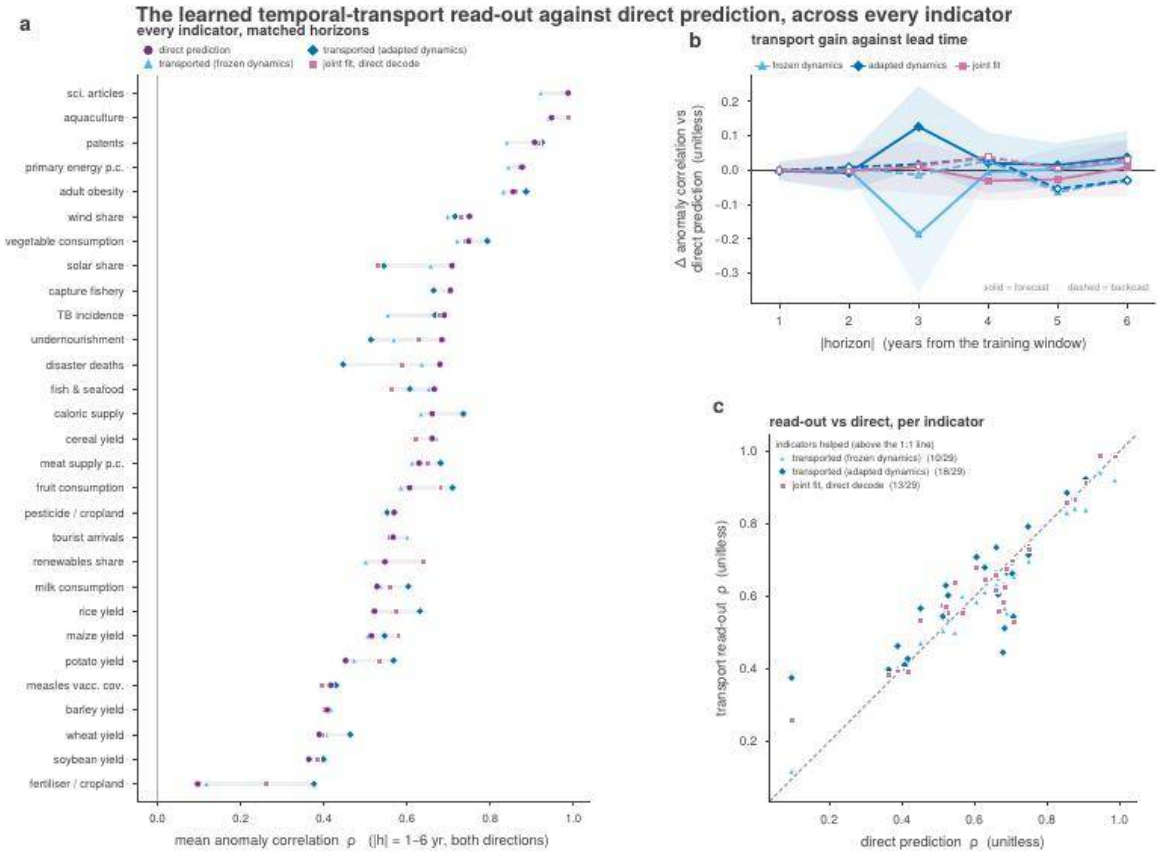

Figure 66: **The transport operator against direct prediction.** Three transported read-outs, frozen dynamics, adapted dynamics and a joint fit decoded directly, set against plain direct prediction over every indicator of the battery rather than a single task. **(a)** Mean anomaly correlation per indicator, pooled over horizons of one to six years in both directions. **(b)** The gain over direct prediction against lead time, forecast solid and backcast dashed; above zero the operator helps. **(c)** Each read-out against direct prediction indicator by indicator, with the number of indicators it helps listed beside the 1:1 line.

are where the degradation is legible: the reconstructions lose fine texture first and coherent large-scale structure last, and the fields that survive the sparsest lattice are the ones whose structure the representation already carries.

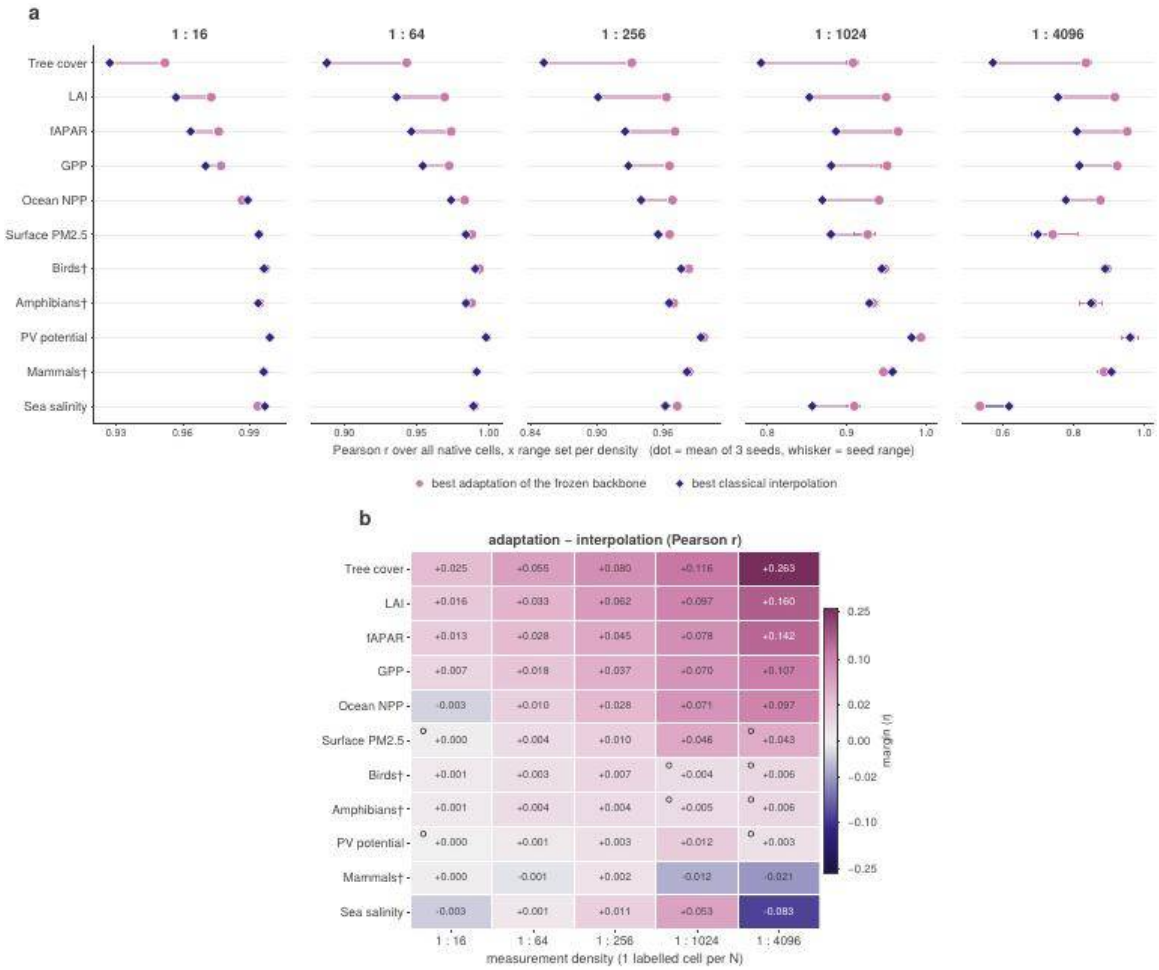

Figure 67: **Sparse reconstruction, per target and per density.** **(a)** Best classical interpolation against best adaptation of the frozen backbone at each measurement density, whiskers spanning three seeds, targets ordered by the margin at the sparsest lattice. **(b)** The margin as a target × density map, on a signed square-root scale symmetric about zero; open circles mark cells whose sign is not shared by all three seeds. Correlation is scored over all valid native cells, the measured ones included, which favours interpolation; the held-out R^2 of Fig. 14b is the skill measure quoted in the main text. Daggers mark the three species-richness targets, whose labels overlap the training corpus; they are shown because reconstructing a field from its own sparse lattice is not the transfer setting the gate concerns, and they are not counted as held-out skill.

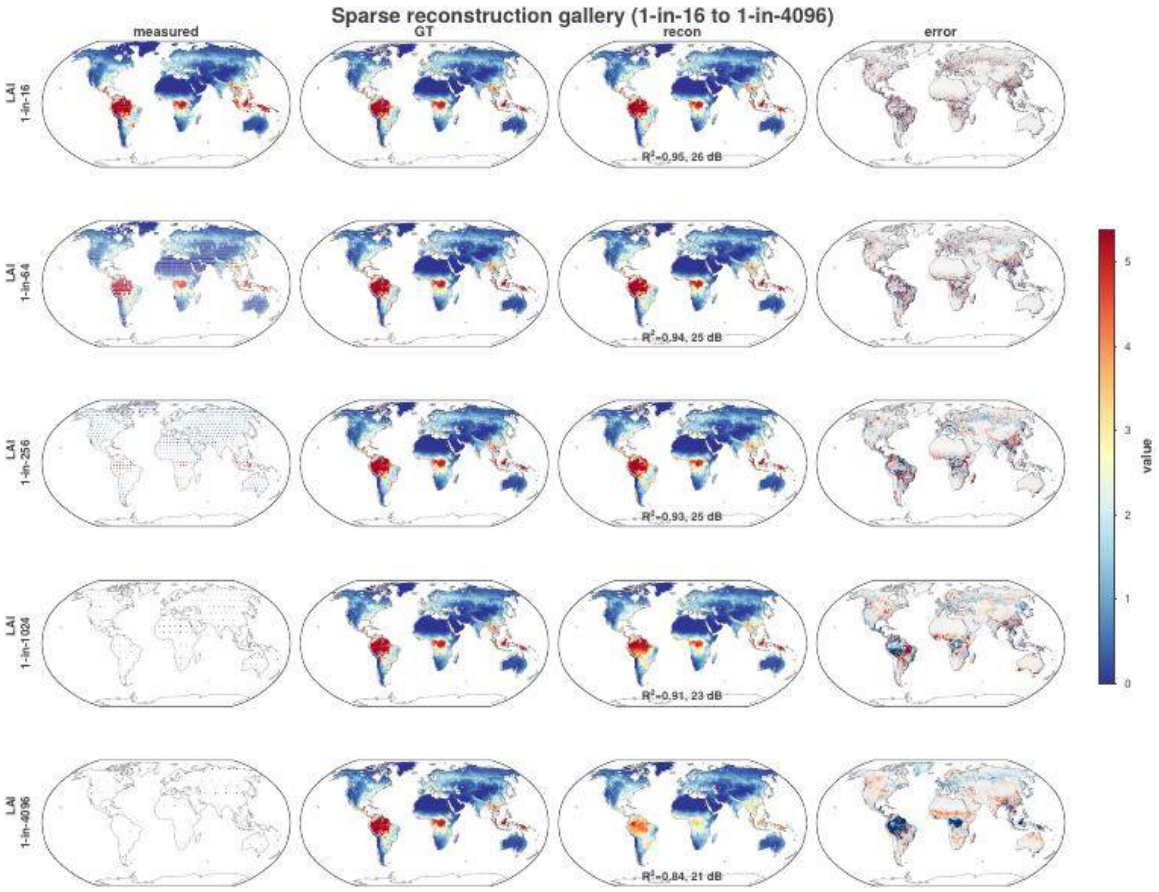

Figure 68: **Leaf area index reconstructed from a thinning lattice.** One row per measurement density, from one cell in sixteen at the top to one in 4,096 at the bottom: the measured cells, the ground truth, the reconstruction with its held-out R^2 and PSNR, and the error. One value scale is shared by the measured, truth and reconstruction columns of the whole figure; the error maps use their own diverging scale centred on zero.

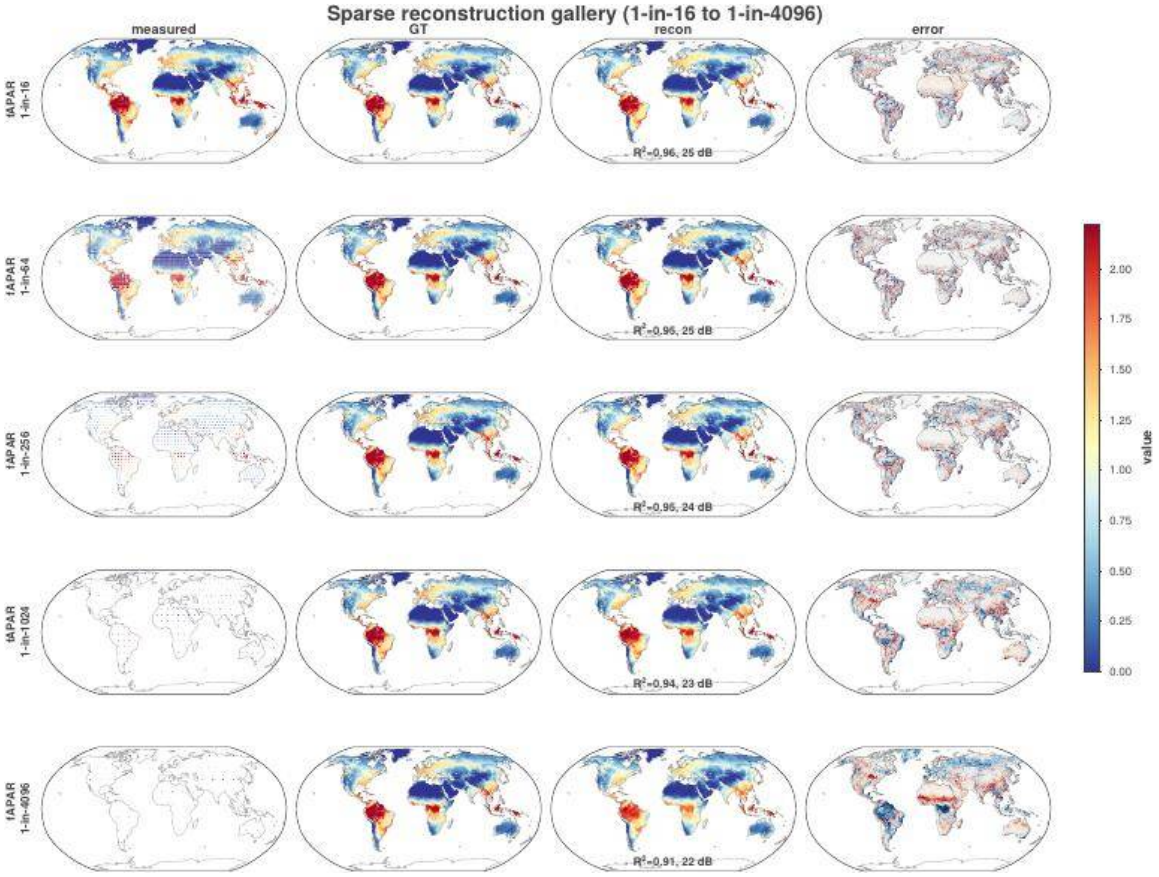

Figure 69: **Fraction of absorbed photosynthetically active radiation.** Rows and columns as in Fig. 68.

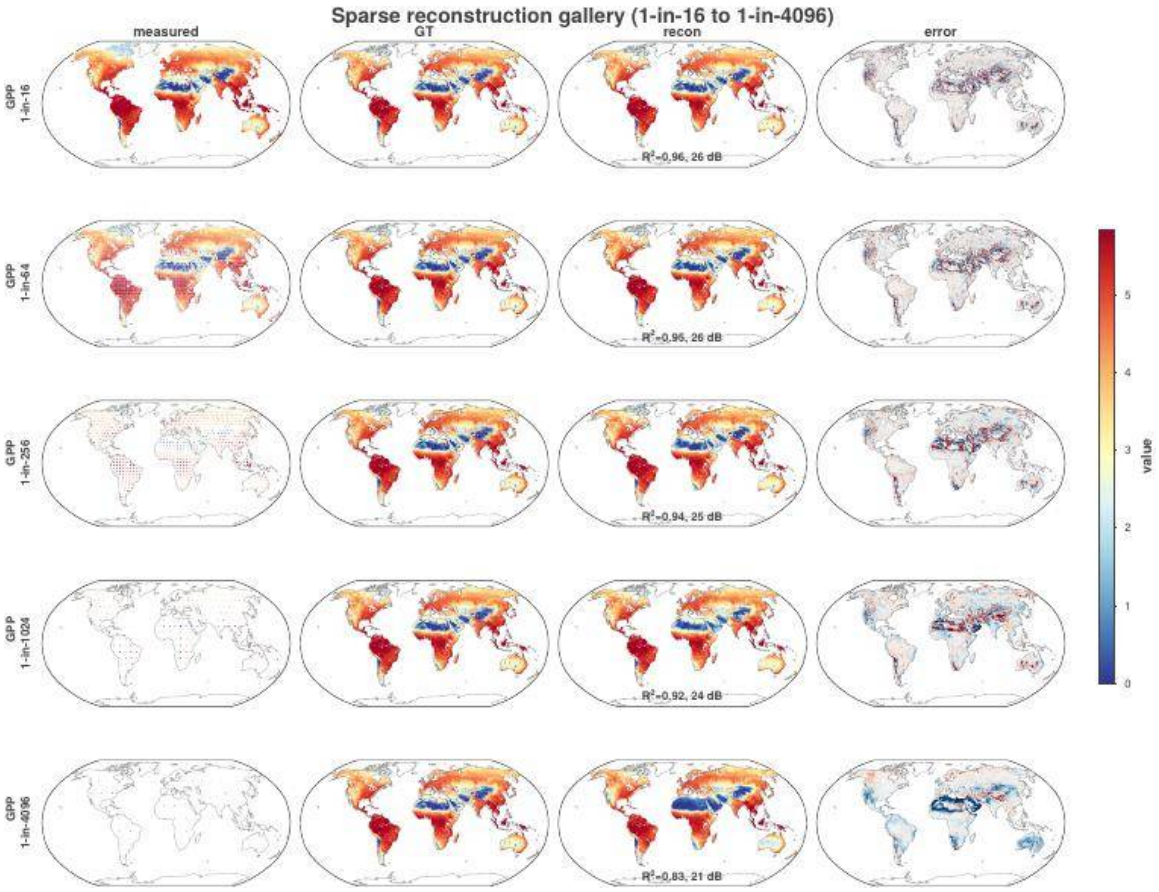

Figure 70: **Terrestrial gross primary productivity.** Rows and columns as in Fig. 68.

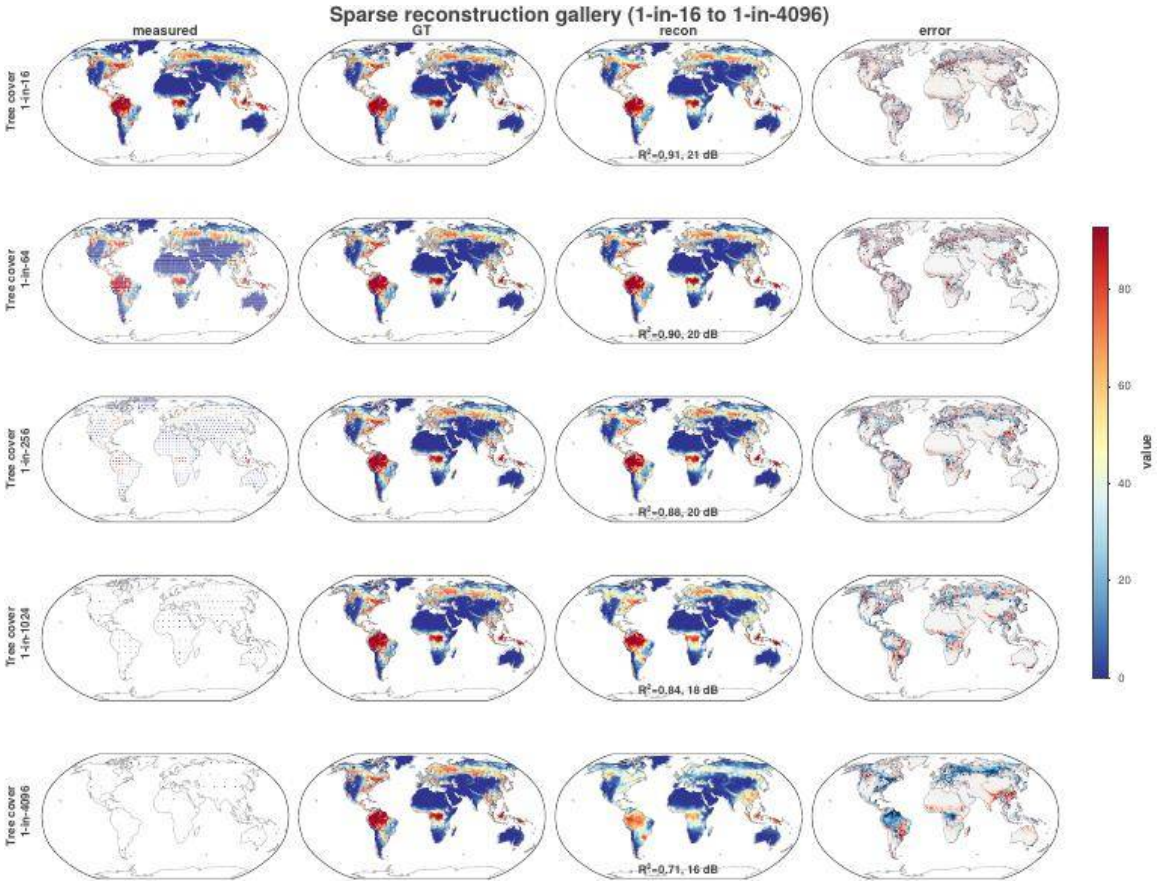

Figure 71: **Tree cover fraction.** Rows and columns as in Fig. 68. This is the target on which adaptation gains most over interpolation at the sparsest lattice.

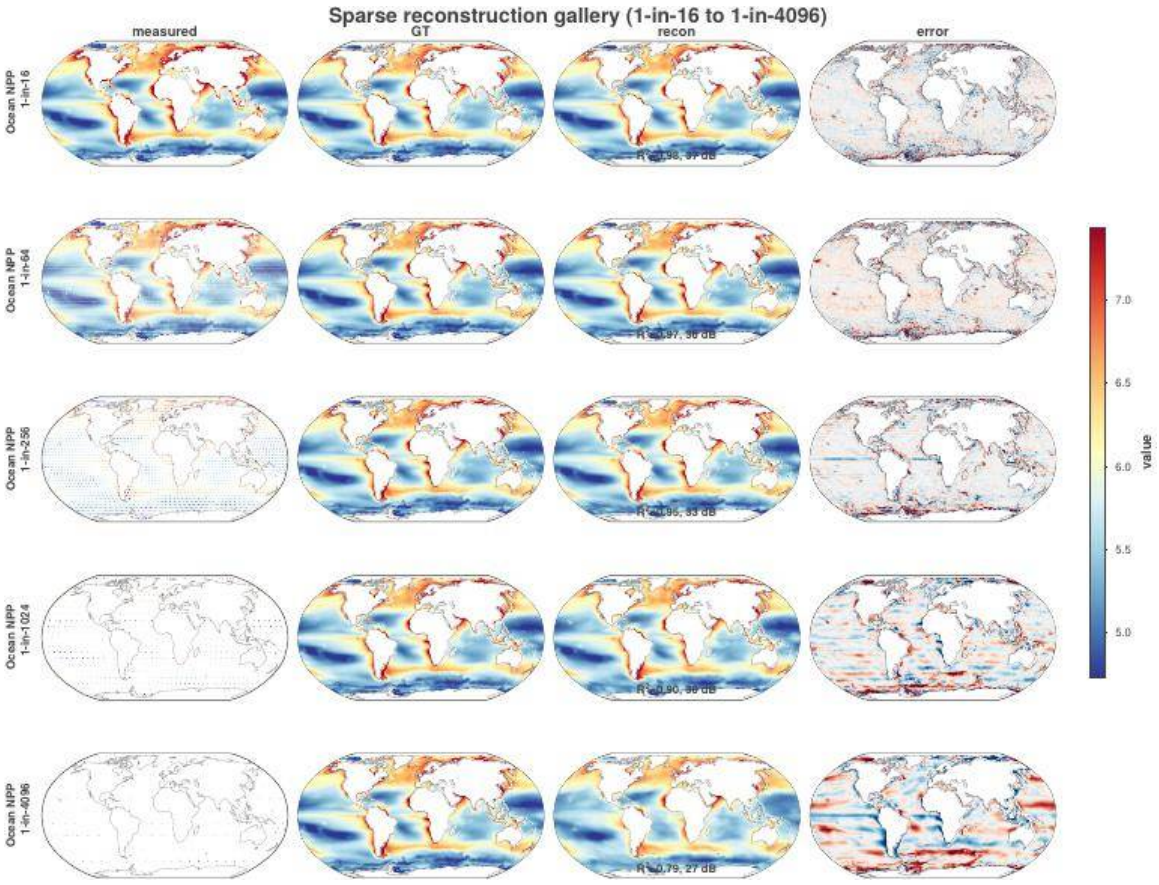

Figure 72: **Ocean net primary productivity.** Rows and columns as in Fig. 68. A marine target, on which the land-image encoders of §C.1 have no training signal at all.

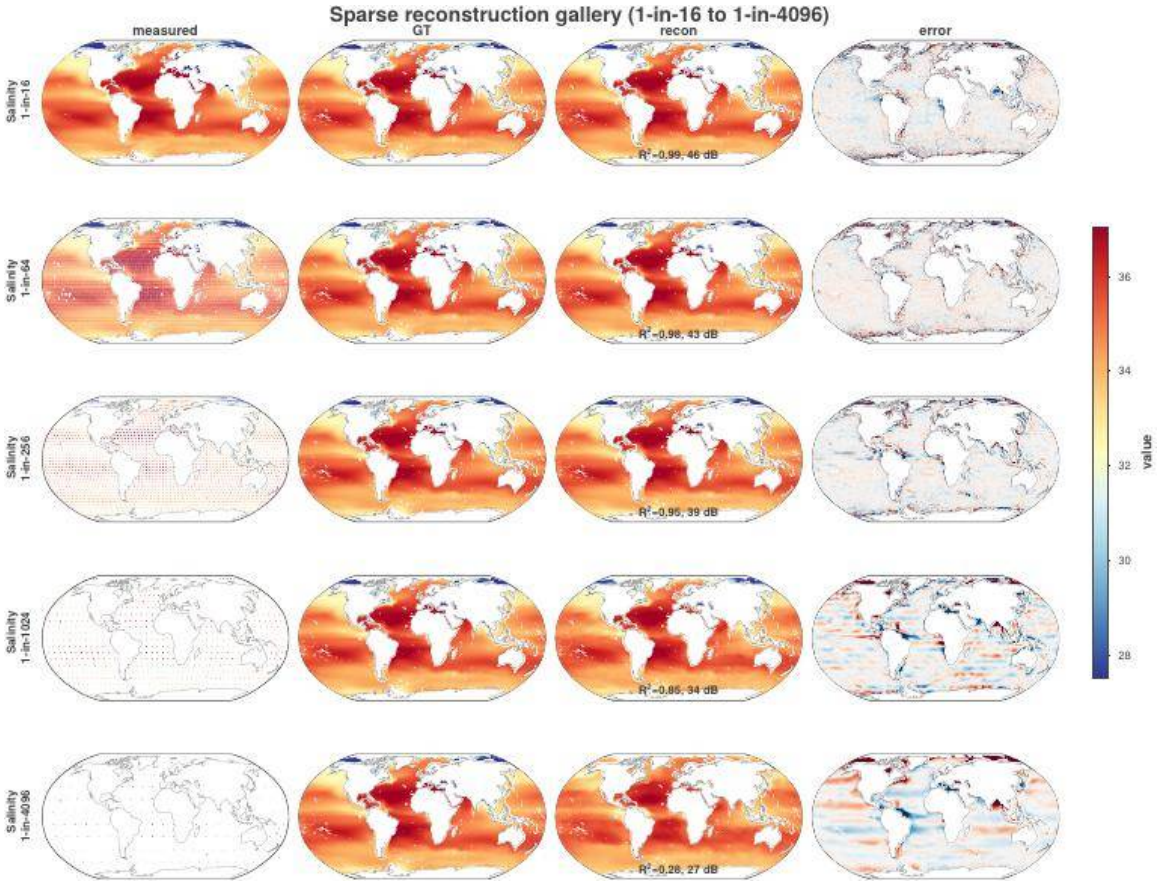

Figure 73: **Sea-surface salinity.** Rows and columns as in Fig. 68. The weakest target of the eight, and the one on which interpolation is ahead at the sparsest lattice.

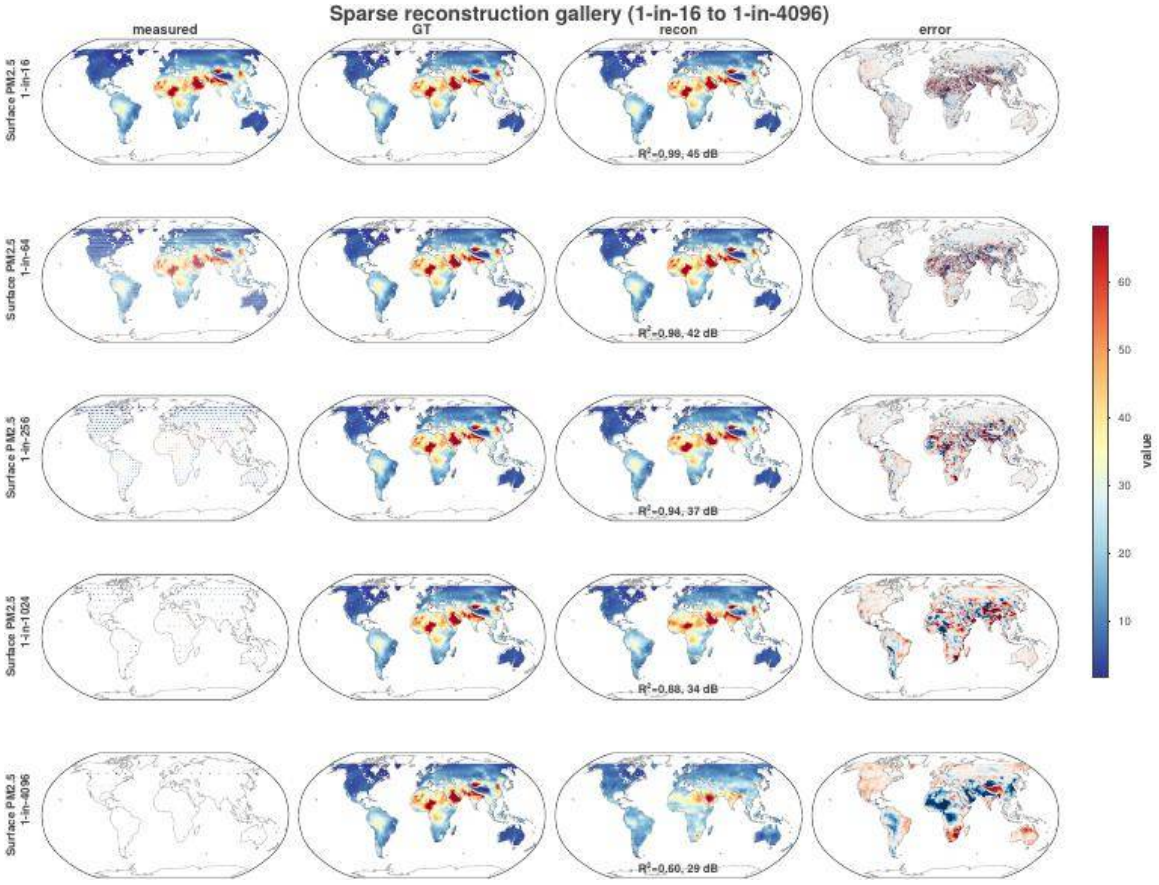

Figure 74: **Surface $PM_{2.5}$.** Rows and columns as in Fig. 68.

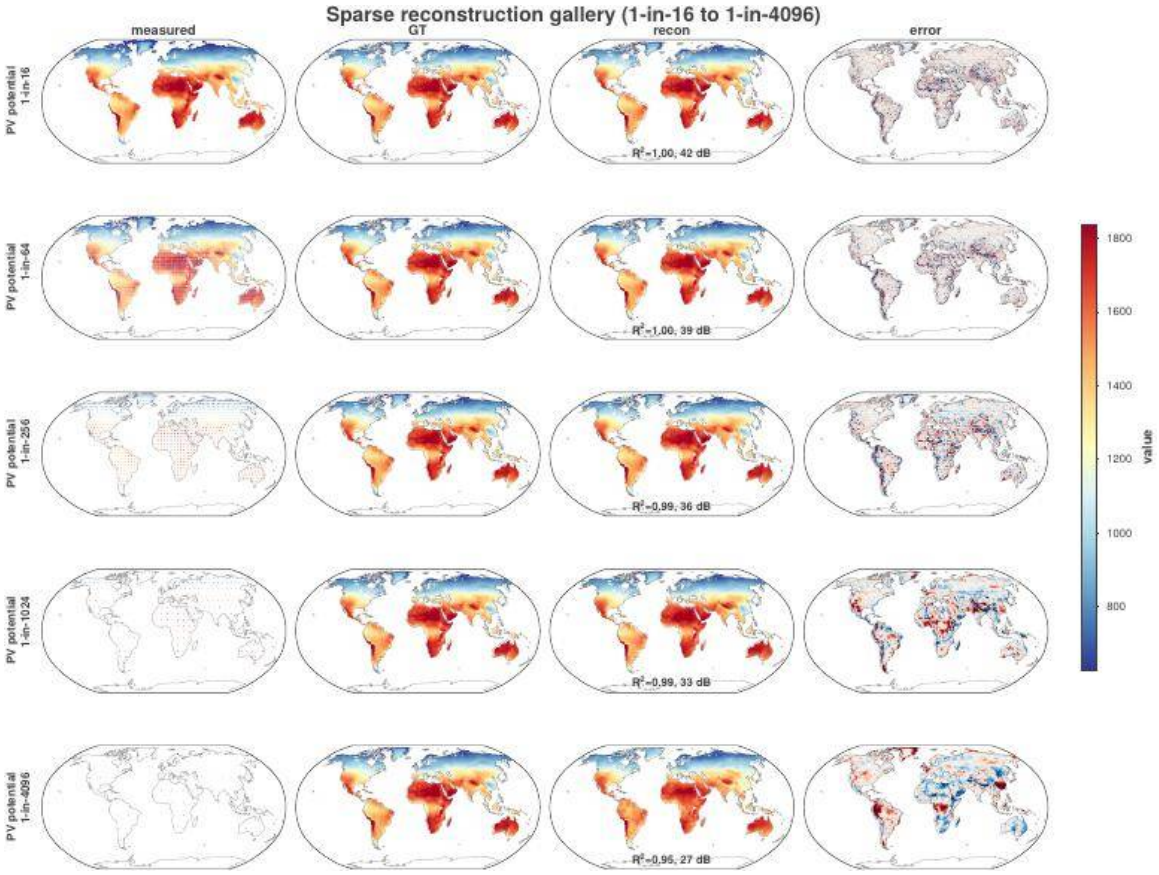

Figure 75: **Photovoltaic potential.** Rows and columns as in Fig. 68.

C.4 National cross-sections from a few reporting countries

The protocol is as follows. For each leakage-clean country indicator we keep its best-covered year, hold out 30% of countries, reveal K of the rest, fit a fresh task row with rank-4 MiSS residuals on the country route, and predict the held-out national level. The reference is a geostatistical ladder over country centroids, comprising Gaussian-process kriging with a Matérn-3/2 kernel (Rasmussen and Williams, 2006; Cressie, 1993), k-nearest neighbours, inverse-distance weighting (Shepard, 1968) and the global mean, together with a frozen-embedding probe, all on the same split and all recalibrated the same way. 31 indicators, three seeds. Inverse-distance weighting is the strongest of the four references, ahead of the fitted Gaussian process at every budget and the best baseline on 15 of the 31 indicators at $K = 32$, because a handful of points spread over the globe is too little to fit a length scale and a Gaussian process reverts to its training mean away from its observations, which at intercontinental distances is nearly everywhere. Regression kriging, the textbook way of combining a trend model with a spatial one, is not testable here: the MiSS fit nearly interpolates the revealed countries, so the residual field left for kriging is numerically zero, and the equal-weight blend reported in the main text is the honest alternative. Calibration behaves as it does elsewhere, with mean coverage at the 90% level of 0.99 for the raw head, 0.92 after σ-scaling and 0.87 after country-split conformal recalibration (Vovk et al., 2005; Lei et al., 2018). Figures 76 and 77 put the dissociation of the main text on the page, for the single seed those maps draw. Tourist arrivals are not predictable from a country's neighbours, and on that seed the reconstruction climbs from 0.04 at four reporting countries to 0.92 at 128, against a three-seed mean of 0.13 to 0.79, as the model is given more of the cross-section to condition on. Adult obesity varies smoothly with geography, and there the reconstruction reaches only 0.51 on that seed at the largest reporting budget, and reaches it unsteadily, while inverse-distance weighting over the same centroids reaches 0.82 as a three-seed mean.

C.5 Country-to-gridded downscaling

Downscaling adapts a read-out from the location embedding to a nationally reported target and is scored within countries, so that reproducing the national mean earns nothing and a flat national field scores exactly zero. Figure 78 records every target against its own label-shuffle null. Terrestrial GPP recovers a within-country correlation of 0.66 ± 0.10 over five seeds against a null of 0.04, tree cover 0.51 ± 0.15 against a null below zero, fAPAR 0.61 ± 0.08 against 0.10 and leaf area index 0.55 ± 0.13 against 0.07; surface $PM_{2.5}$ recovers little at 0.18 ± 0.14, and photovoltaic potential is negative at -0.06 ± 0.08 with its null above it. What separates them is not the model but the data: skill tracks the share of a target's spatial variance that sits within countries rather than between them ($r = 0.79$ across the nine targets), and photovoltaic potential has 84% of its variance between countries, which leaves nothing for a within-country score to find.

The maps behind those numbers are worth seeing side by side (Figs. 79 and 80). For GPP the recovered field and the flat national-mean null are visibly different objects, and the recovery holds in 91% of the 151 countries scored. For photovoltaic potential the truth is a smooth latitudinal band that national means already capture, and the decoded field paints sub-national texture where the data has none. The three species-richness targets in Figure 78 are marked because their labels overlap the training corpus; they are reported

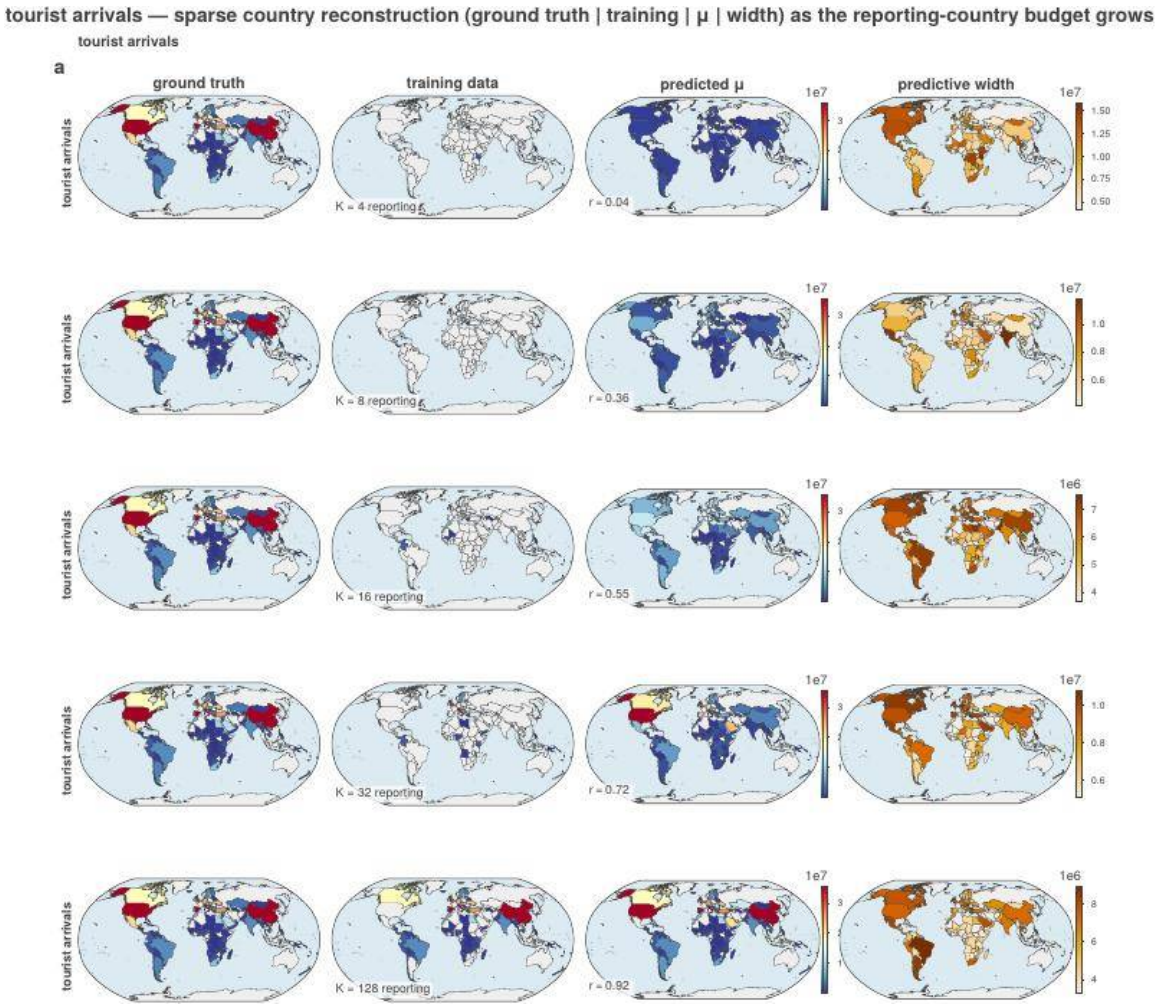

Figure 76: **Tourist arrivals, against the reporting budget.** Each row is one value of K: truth, the revealed countries, the predicted level and the predictive width. This is an indicator a neighbour-based interpolator cannot reach, and it is where the representation gains most.

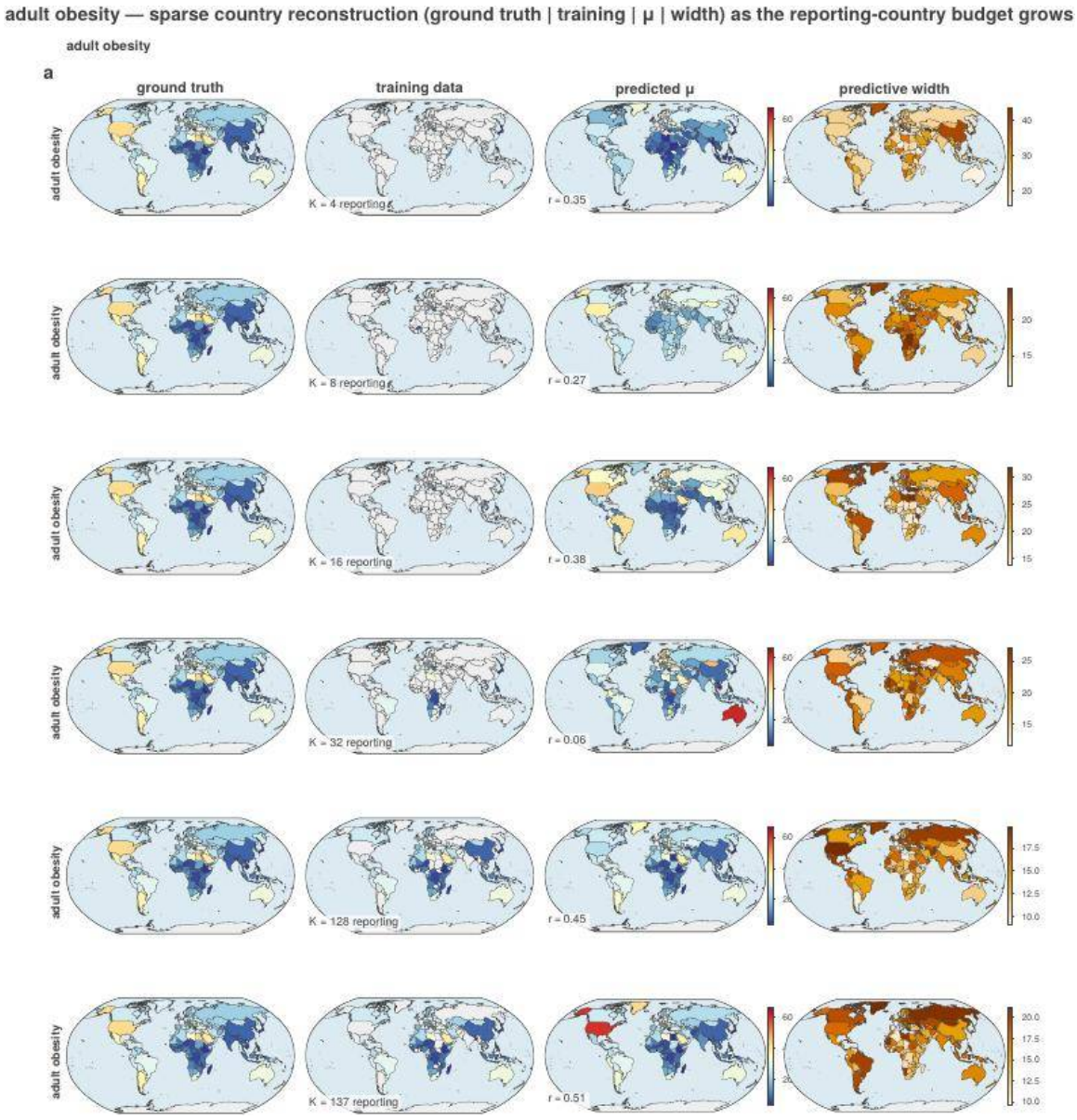

Figure 77: **Adult obesity, against the reporting budget.** Columns as in Fig. 76, with a sixth row at $K = 137$, the largest reporting budget this indicator supports. A spatially smooth indicator, on which interpolation over country centroids is well ahead of the learned read-out, and on which the learned read-out is also unsteady: held-out level correlation runs 0.35, 0.27, 0.38, 0.06, 0.45 and 0.51 across the six budgets.

for completeness and are excluded from every held-out skill claim in §6 of the main text, appearing elsewhere only inside the matched knockout contrast of §5.2, where both arms carry the same overlap and it therefore cancels. As in the main text we report this as the spatial disaggregation the coupled representation makes possible and not as a validated product, and the operation is not always meaningful: a governance or institutional indicator has no sub-national field for a within-country score to recover, so a plausible-looking decoded surface for one should not be read as an estimate.

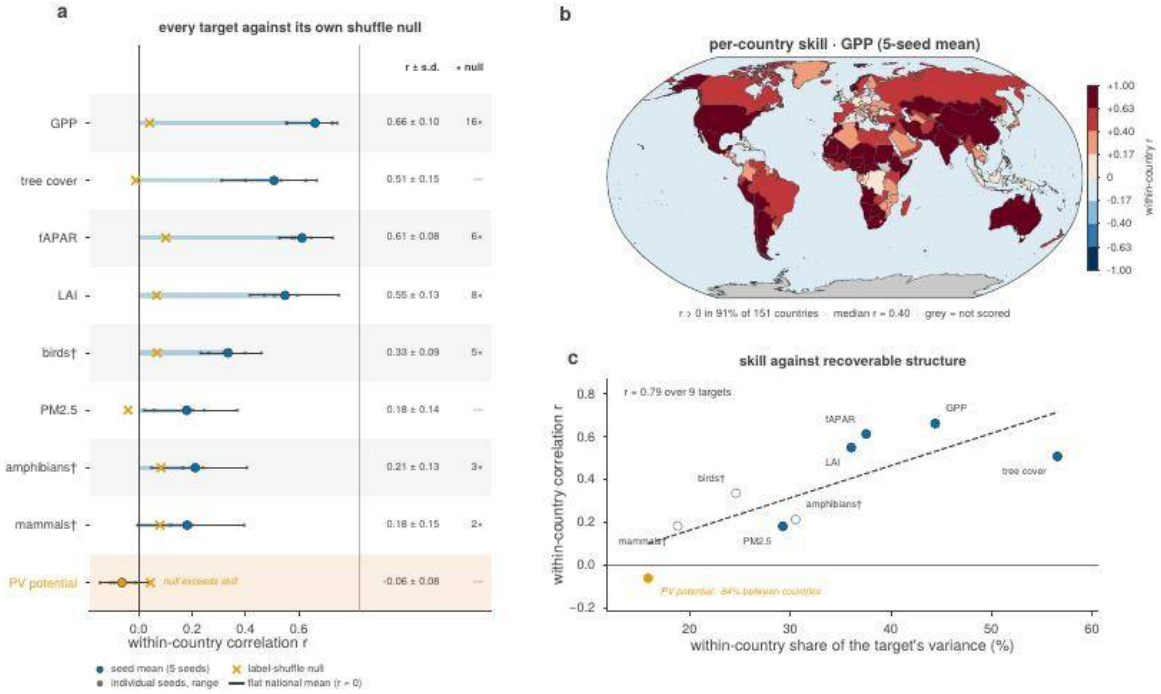

Figure 78: **Downscaling, per target. (a)** Every target against its own label-shuffle null, five-seed mean with the seed range, ordered by the margin over that null; the zero rule is the flat national mean. **(b)** Per-country within-country correlation for the strongest target. **(c)** Skill against the share of the target's variance that sits within countries, a property of the data rather than of the model. Daggers mark targets whose labels overlap the training corpus.

C.6 Anatomy of the representation

Figure 81 draws how much each cell looks like the country it sits in, read two ways. Pushed through the same held-out bridge and compared with its own country's embedding, a cell reaches a median cosine of 0.163 over the land surface; compared instead with the mean of its own territory, with no bridge in between, the median is 0.304. Both readings order countries the same way, and the order is one of size and internal variety rather than of skill: Antarctica at 0.026, the United States at 0.069 and Russia at 0.077 are least like their own country, while Korea at 0.602 and a group of compact African and Baltic states sit above 0.53. The spread within a country is the more informative half, because a small territory scores a high

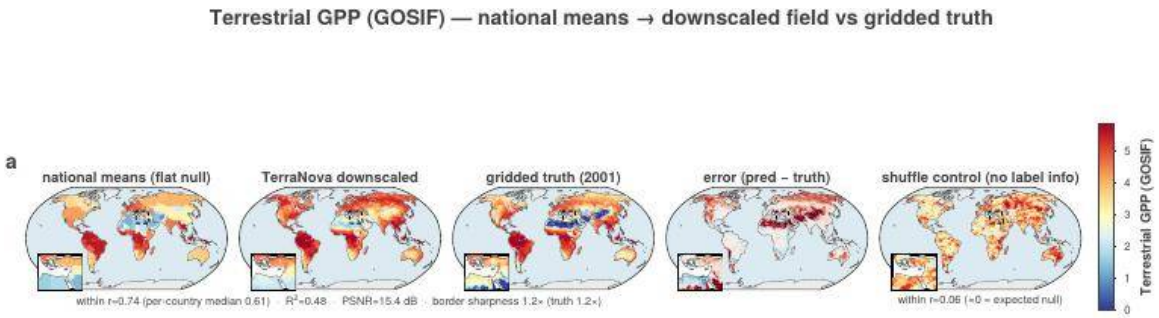

Figure 79: **Terrestrial GPP downscaled from national values.** From left to right: the national means, which are both the input and the flat null; the downscaled field; the gridded truth; the error; and a shuffle control fitted on permuted national labels, which recovers nothing. Insets zoom on the eastern Mediterranean.

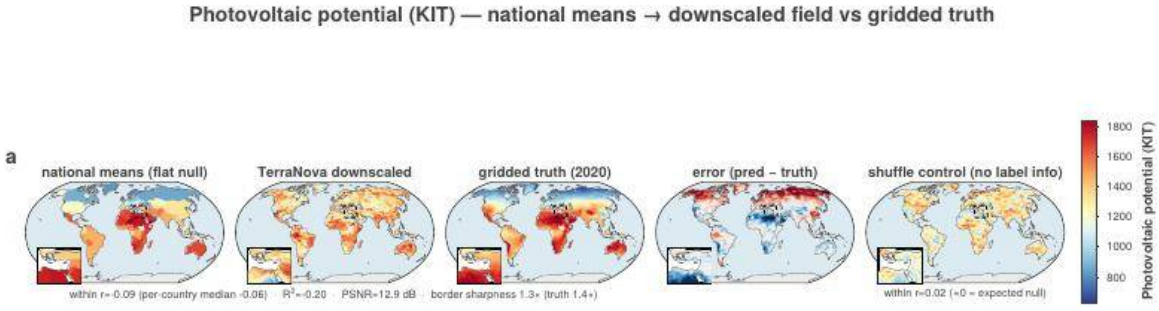

Figure 80: **Photovoltaic potential, the ill-posed case.** Panels as in Fig. 79. The truth is a smooth latitudinal band which national means already carry, so there is no sub-national structure left to recover and the decoded field invents texture.

median almost by construction. The widest interquartile ranges belong to Iraq at 0.406, Chad at 0.364, Morocco at 0.342 and Mali at 0.322, every one of them a country straddling a desert-to-sahel or coast-to-interior transition, so the heterogeneity the representation reports inside a border tracks a real environmental gradient rather than noise.

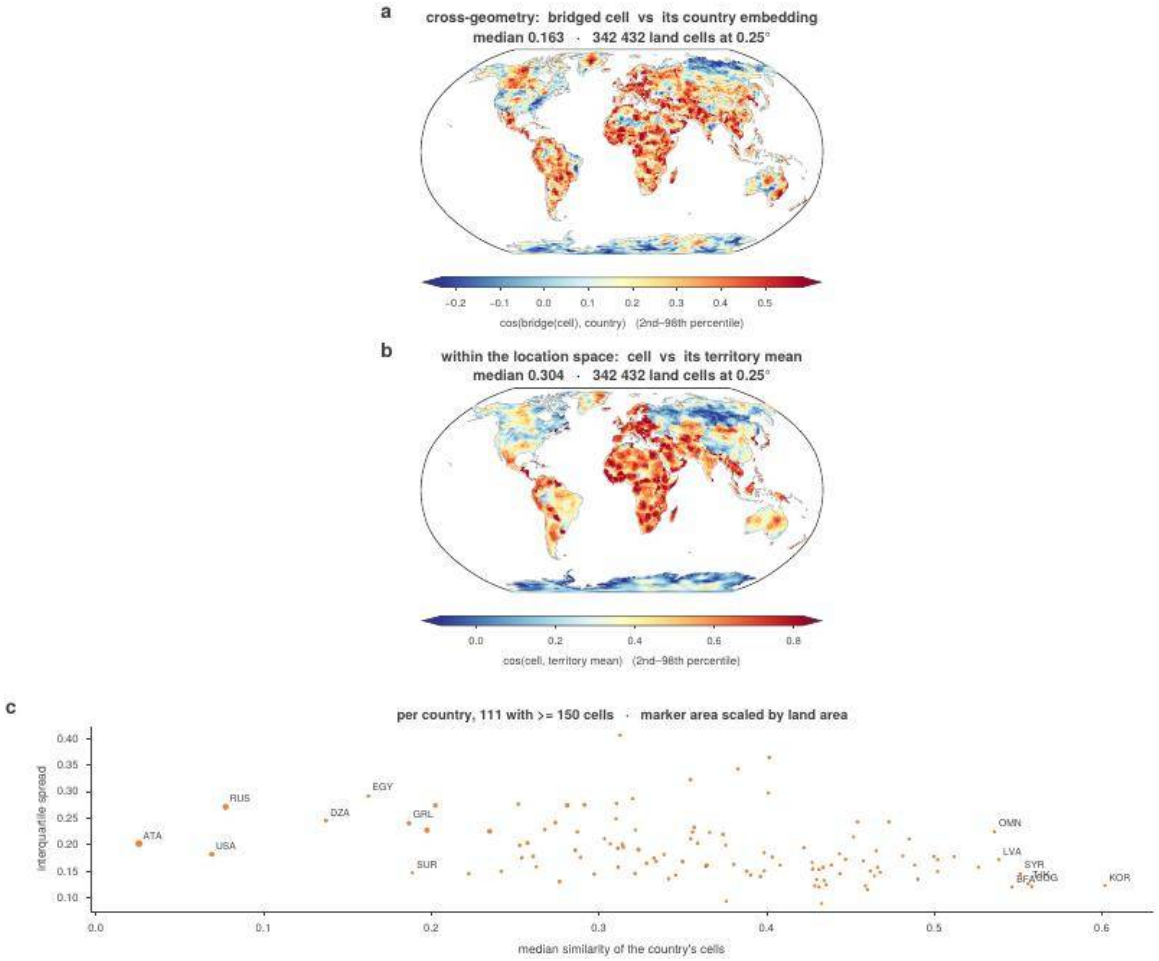

Figure 81: **How much each** 0.25° **cell looks like its own country.** **(a)** Cross-geometry: each cell's location embedding pushed through the held-out ridge bridge and compared with its own country's country-encoder embedding. This is the continuous field that cross-geometry retrieval thresholds. **(b)** Within the location space: the same cells compared with the mean embedding of their own territory, with no bridge in between, which measures how typical a cell is of its country rather than how identifiable it makes it. Both panels are cosines and both colour scales are clipped to the second and ninety-eighth percentiles. **(c)** Per country, the median of **(a)** against its interquartile spread, over the 111 countries with at least 150 land cells, with marker area scaled by land area. A compact country earns a high median almost by construction, so the spread is the more informative axis: the countries at the top of it are those that straddle a sharp environmental transition.

Intrinsic dimension measures how many directions a representation actually uses, and it can be read globally over a set of points or locally around each one. Figure 82 does both. The local map is geographical: dimension concentrates over structurally complex land, along the Andes, the Himalaya and the continental margins, and collapses over the open ocean, at a mean of 6.8 on land against 3.3 at sea. Globally, whole-set dimension tracks downstream static-spatial skill across the benchmark encoders at Spearman $\rho \approx 0.75$. Absolute estimates are estimator-dependent, so we quote FisherS throughout, we do not rank encoders on absolute dimension, and we do not read the association as evidence that this embedding is the more separable one: probed linearly it ranks seventh of the nine read-outs of §C.1. The two embedding spaces use their ambient width very differently: the location space spends 6.5 to 9.9 of its 256 dimensions and the country space 17 to 19 of its 64, by far the denser of the two and consistent with 247 countries being too few points to lie on a thin manifold; separately, the 128-unit penultimate layer of a per-task probe head uses only 2.7 to 5.6.
Both learned spaces support the vector operations the main text describes, and Figures 83 and 84 give more of them than the main figure has room for. Nearest neighbours in the location space are organised by what a place is like rather than by where it is: Nairobi retrieves Addis Ababa, Bogotá and São Paulo, Perth retrieves Cape Town, Melbourne and Santiago, and Tokyo retrieves Osaka, Los Angeles and Seattle. Read across spaces, the decoded indicator panel reproduces relationships nobody supplied as a target, among them the Preston curve between income and life expectancy at $\rho = 0.74$, income against child mortality at -0.81, life expectancy against fertility at -0.78 and income against sanitation access at 0.73. The task embeddings of each pair are only weakly aligned, at cosine 0.13 to 0.44, so the relationship is carried by the country representation rather than by two task rows that happen to sit near one another.

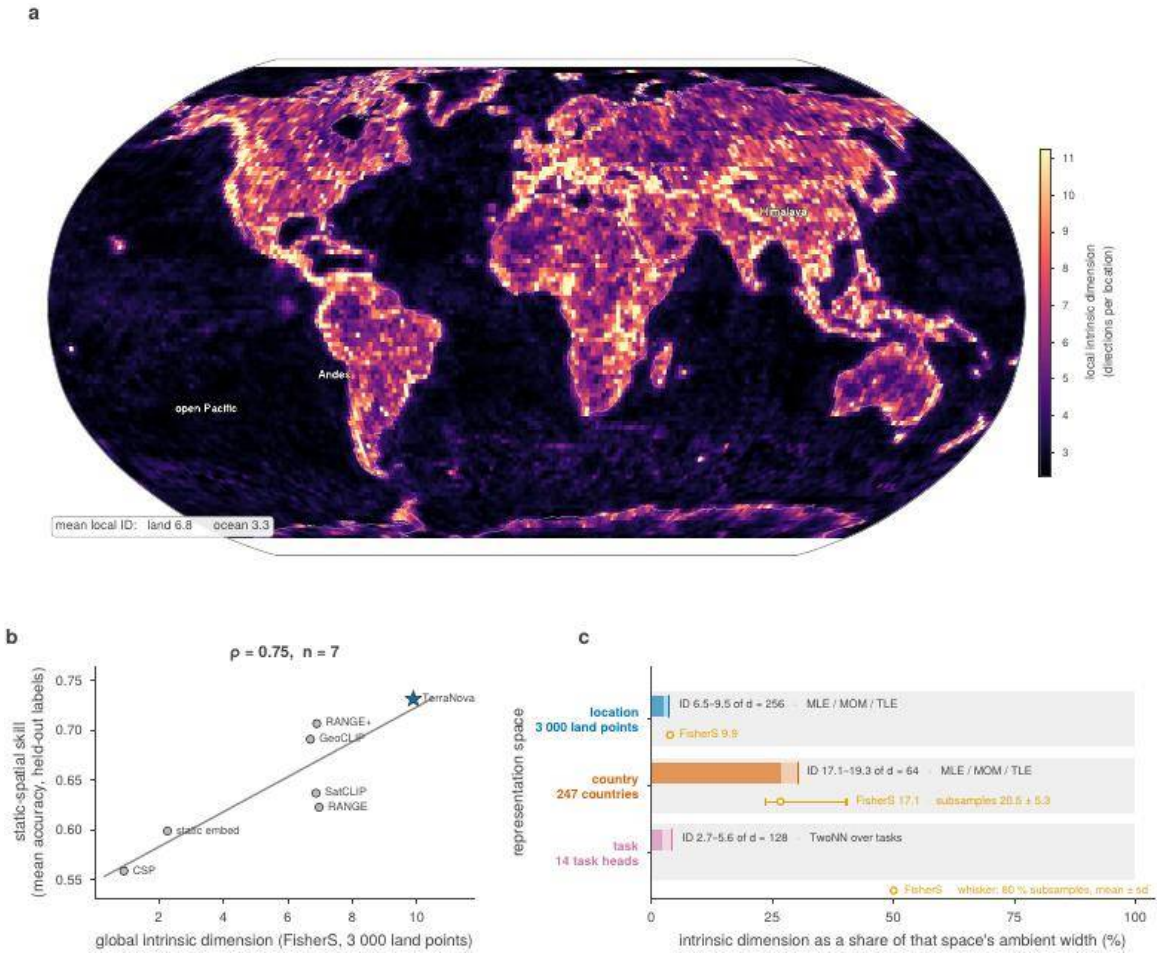

Figure 82: **Intrinsic dimension of the learned spaces.** **(a)** Local intrinsic dimension per location, high over structurally complex land and low over the open ocean. **(b)** Whole-set dimension against static-spatial skill for the benchmark location encoders. **(c)** Dimension as a share of the ambient width, for the location and country embedding spaces and for the penultimate layer of the per-task probe heads, with several estimators.

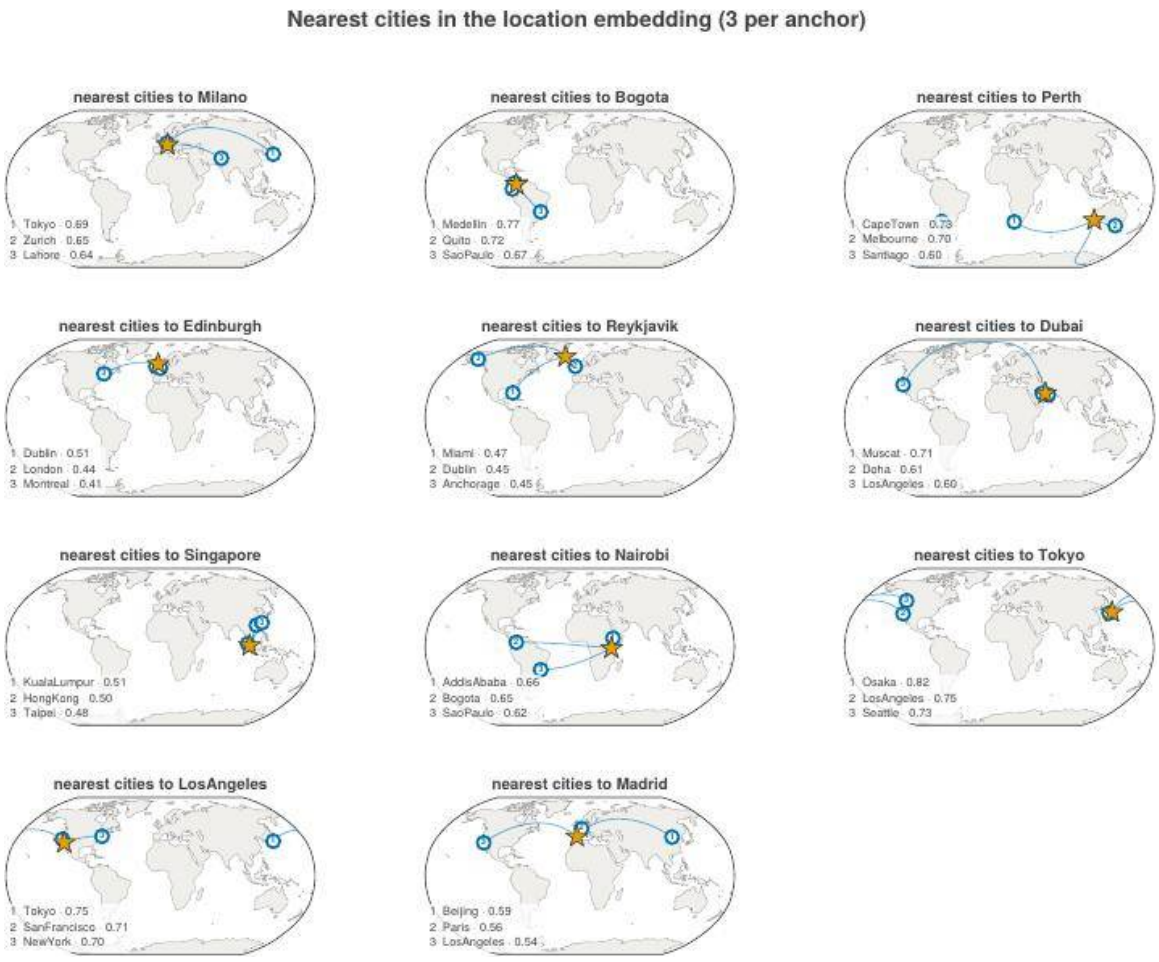

Figure 83: **Nearest cities in the location space.** For each anchor, the three nearest cities by cosine similarity of their spatiotemporal representation, with the similarity listed. Neighbours are organised by what a place is like rather than by how close it is.

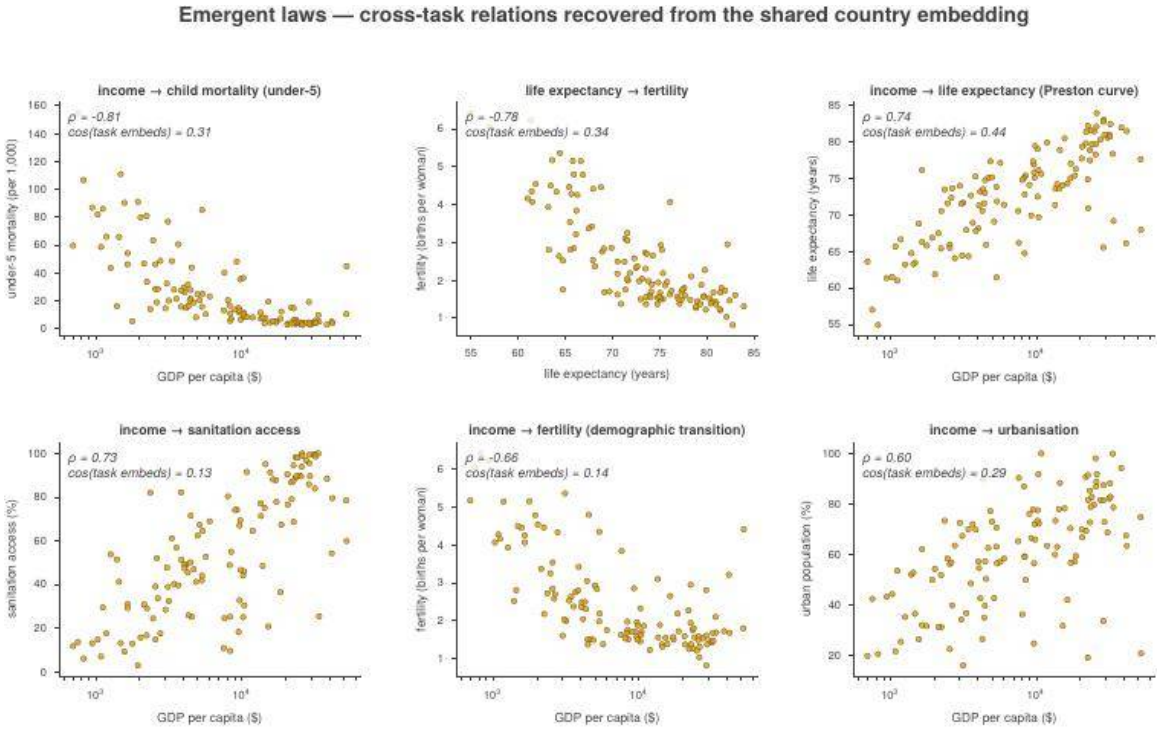

Figure 84: **Relationships recovered from the decoded country representation.** Each panel decodes two indicators for every country from the same embedding and plots one against the other, with the Spearman correlation and the cosine similarity of the two task embeddings. None of these relationships was supplied as a training target.

Appendix D. Computational Cost

D.1 Measurement protocol

Everything in §D.2 to §D.5 was measured on one machine and one checkpoint. The device is a laptop with an NVIDIA GeForce RTX 5070 Laptop GPU, an 8 GB card that the driver reports as 7,682.8 MiB of usable memory, next to a 32-core CPU on which torch was given 24 threads; the software is Python 3.14.2, torch 2.11.0+cu130 and CUDA 13.0, and the model is the production checkpoint `ckpt/last-ema.ckpt` built from `terranova_v2/configs/yaml/training/evidential_full_512.yaml`. The repository, the checkpoint and the label caches were all on a local NVMe working copy, so no network filesystem and no cold cache enters any timing. Adaptation uses rank-4 MiSS unless a row names another operator, and every adaptation cell was run for three seeds on the GPU and one seed on the CPU. Setup, fit and evaluation are timed separately in the record (`setup_seconds`, `fit_seconds`, `eval_seconds`), and the tables report fit time as the cost of adapting, with evaluation time given separately where it dominates. The CPU cells tabulated in §D.3 are therefore that single seed-0 run, whereas Table 1 of the main text reports three-seed medians on both devices, so the CPU numbers here and there are not expected to agree: the country-route fit at $K = 300$ is 604.6 s at seed 0 below against a three-seed median of 538.1 s in the main text.

Energy is integrated from `nvidia-smi power.draw` sampled at 100 ms across each measured cell. The raw integral is `joules`; the tables report `joules_net`, which subtracts the idle baseline of 21.65 W probed on the same device immediately before the run. The reading is whole-device, so any other process resident on the GPU is included in it, and this card reports no power limit, so no utilisation-against-TDP figure is derived anywhere in this section. Cells shorter than about a second collect only two or three samples, and the closed-form probe on the country route collects one, so their energy is indicative rather than a measurement; the sample count is written to every row as `power_samples`, so any cell can be checked against its own sampling density. The sampling window brackets the whole cell rather than the fit alone, which is why the energy columns cover setup, fit and evaluation together while the time columns separate them.

D.2 The adaptation ladder on a consumer GPU

The ladder was measured on one representative task per geometry, the national indicator *SDG1 poverty headcount (nowcast)* on the country route and *species richness (vertebrates)* on the gridded route, at label budgets $K \in \{30, 100, 300, 1000, 3000\}$ for each of the six operators of §A.23.2 and three seeds. Every gradient operator takes the same 800 optimiser steps, recorded as `steps_taken`, and the linear probe is closed form and takes none, so within a geometry fit time is set by the batch, which is the whole label set up to a cap of 1,024. That cap and the size of the country task are why the two largest budgets cost the same: the country indicator supplies only 1,113 to 1,214 labelled points depending on the seed's split, so its $K = 3{,}000$ row is the same fit as $K = 1{,}000$ on a batch of 1,024 rather than a larger one, and on the gridded route the $K = 3{,}000$ fit runs 1,024-point minibatches rather than one 3,000-point batch. The skill column is the record's `skill` field, which on the gridded

route equals the recorded Pearson correlation in every row and on the country route is the anomaly correlation; `corr` and `r2` are stored alongside it for every cell.

Table 26: **Adaptation ladder on the country route, RTX 5070 Laptop GPU.** Task: SDG1 poverty headcount (nowcast). Fit time is the median over three seeds with the seed range beside it; skill is the three-seed mean and standard deviation of the recorded anomaly correlation; net kJ is the median over seeds of `joules_net` and covers setup, fit and evaluation together. At $K = 3{,}000$ the task supplies only 1,113–1,214 labels, so that row is not a larger fit than $K = 1{,}000$.

operator	trainable	K	fit (s)	fit, seed range (s)	skill (mean ± s.d.)	net kJ
linear probe	256	30	0.012	0.002–0.047	-0.026 ± 0.093	0.003
		100	0.002	0.002–0.199	0.318 ± 0.080	0.003
		300	0.002	0.002–0.005	0.456 ± 0.145	0.002
		1,000	0.004	0.003–0.014	0.498 ± 0.136	0.003
		3,000	0.006	0.004–0.094	0.498 ± 0.133	0.002
task embedding	256	30	11.6	11.5–12.3	0.073 ± 0.032	0.544
		100	20.5	20.1–21.9	0.081 ± 0.035	0.869
		300	46.2	45.4–49.5	0.148 ± 0.142	1.858
		1,000	134.9	133.0–146.8	0.190 ± 0.209	5.466
		3,000	135.6	133.3–142.5	0.247 ± 0.164	5.519
VeRA (r=4)	101,324	30	14.6	14.4–15.4	-0.006 ± 0.172	0.676
		100	22.2	21.0–24.5	0.116 ± 0.064	0.927
		300	49.6	46.9–54.4	0.456 ± 0.295	1.894
		1,000	147.0	139.6–154.5	0.612 ± 0.255	6.036
		3,000	148.2	140.1–167.1	0.363 ± 0.176	5.877
MiSS (r=4)	416,000	30	14.0	13.9–14.4	0.214 ± 0.231	0.629
		100	22.7	21.7–23.6	0.568 ± 0.115	0.923
		300	49.5	41.7–51.9	0.602 ± 0.149	1.968
		1,000	144.9	138.9–160.6	0.713 ± 0.157	5.754
		3,000	145.8	139.2–155.9	0.694 ± 0.165	5.909
LoRA (r=4)	803,072	30	14.1	14.0–14.4	0.261 ± 0.203	0.623
		100	22.2	21.9–23.3	0.512 ± 0.173	0.935
		300	49.2	47.8–51.5	0.624 ± 0.172	1.998
		1,000	144.4	140.4–156.0	0.655 ± 0.165	5.886
		3,000	145.2	140.4–152.8	0.718 ± 0.134	5.931
finetune-last	25,717,508	30	16.2	15.8–17.5	0.234 ± 0.250	0.774
		100	25.1	24.0–28.1	0.585 ± 0.090	1.045
		300	52.0	49.4–55.9	0.658 ± 0.144	2.083
		1,000	144.9	136.8–163.1	0.557 ± 0.142	5.664
		3,000	145.1	137.0–159.6	0.629 ± 0.187	5.784

Two things are visible in the two tables. First, fit cost is governed by the label budget and not by the operator: at $K = 300$ on the country route the five gradient operators span 46.2 to 52.0 s, a factor of 1.1, and at $K = 3{,}000$ on the gridded route they span 148.3 to 174.3 s,

Table 27: **Adaptation ladder on the gridded route, RTX 5070 Laptop GPU.** Task: species richness (vertebrates). Columns as in Table 26; skill here is the recorded Pearson correlation. Evaluation on this route costs 23 to 27 s per cell against 0.17 to 0.19 s on the country route, which is why the energy column does not fall as steeply as the fit column at the smallest budgets.

operator	trainable	K	fit (s)	fit, seed range (s)	skill (mean ± s.d.)	net kJ
linear probe	256	30	0.305	0.281–0.314	0.250 ± 0.027	0.531
		100	0.316	0.305–0.336	0.433 ± 0.087	0.528
		300	0.308	0.292–0.354	0.677 ± 0.032	0.479
		1,000	0.317	0.298–0.350	0.705 ± 0.022	0.457
		3,000	0.358	0.318–0.369	0.719 ± 0.022	0.448
task embedding	256	30	14.2	13.6–14.2	0.529 ± 0.091	1.678
		100	22.7	20.8–22.7	0.707 ± 0.063	2.022
		300	50.2	47.1–51.0	0.825 ± 0.024	3.123
		1,000	147.3	140.6–149.7	0.824 ± 0.016	7.240
		3,000	148.3	143.3–151.1	0.847 ± 0.010	7.278
VeRA (r=4)	101,324	30	17.9	16.9–17.9	0.525 ± 0.131	1.794
		100	25.4	24.0–25.8	0.793 ± 0.040	2.105
		300	55.6	52.7–56.5	0.895 ± 0.057	3.222
		1,000	163.8	156.6–167.4	0.922 ± 0.034	7.613
		3,000	174.3	156.9–176.4	0.959 ± 0.004	7.525
MiSS (r=4)	416,000	30	16.9	16.5–17.0	0.726 ± 0.128	1.804
		100	24.7	24.0–25.5	0.897 ± 0.052	2.085
		300	53.3	52.3–55.2	0.952 ± 0.004	3.298
		1,000	161.9	154.2–162.3	0.959 ± 0.006	7.565
		3,000	163.6	154.8–165.3	0.965 ± 0.006	7.577
LoRA (r=4)	803,072	30	16.4	16.2–17.1	0.725 ± 0.138	1.813
		100	24.4	23.6–25.3	0.947 ± 0.006	2.116
		300	53.4	51.9–55.3	0.940 ± 0.020	3.304
		1,000	157.3	154.9–162.0	0.961 ± 0.003	7.716
		3,000	158.2	156.7–163.5	0.961 ± 0.002	7.753
finetune-last	25,717,508	30	20.2	18.8–21.4	0.888 ± 0.019	1.843
		100	28.2	27.0–29.6	0.938 ± 0.024	2.157
		300	56.6	54.9–59.8	0.957 ± 0.002	3.281
		1,000	158.1	153.1–165.9	0.959 ± 0.003	7.433
		3,000	157.3	154.2–169.3	0.964 ± 0.005	7.371

a factor of 1.2, across a trainable-parameter range of 256 to 25,717,508, because the work is the forward and backward pass through the frozen backbone rather than the optimiser state. The linear probe is the exception and is two to four orders of magnitude cheaper, with a median fit of 0.002 to 0.012 s on the country route and 0.305 to 0.358 s on the gridded route, since it is solved in closed form on cached embeddings. Second, the whole ladder is small in absolute terms: the longest cell measured, VeRA at $K = 3{,}000$ on the gridded route, is 174.3 s and 7.5 kJ of net GPU energy, no cell in either table exceeds 7.8 kJ, and rank-4 MiSS at $K = 300$ reaches 0.952 ± 0.004 on the gridded route for 53.3 s and 3.3 kJ. Seed spread on the country route is wide at every budget, for instance 0.602 ± 0.149 for MiSS at $K = 300$, so the ordering of operators there is not resolved by three seeds; on the gridded route the spread narrows to a few thousandths at the largest budgets, where VeRA, MiSS, LoRA and finetune-last converge between 0.959 and 0.965 while the task-embedding baseline stops at 0.847 and the linear probe at 0.719.

D.3 The same fit without a GPU

Table 28 repeats the rank-4 MiSS fit at seed 0 on the CPU of the same laptop, against the matching GPU cell. Against the seed-0 GPU cell the CPU is 6.3 to 14.5 times slower on the fit, and against the three-seed GPU medians the main text quotes it is 6.2 to 11.2 times slower, which puts the largest country fit at 26.8 minutes and the largest gridded fit at 24.4 minutes against 2.3 and 2.6 minutes on the GPU. Evaluation moves the same way but not by the same factor: at $K = 30$ on the gridded route it is 24.3 s on the GPU against 265.1 s on the CPU.

Table 28: **The same MiSS fit on the CPU of the same machine.** Rank-4 MiSS, seed 0, one run per cell, against the seed-0 GPU cell of Tables 26 and 27. The ratio is CPU fit time over GPU fit time. CPU rows carry no energy and no peak-memory figure: the power sampler is GPU-only and is recorded as disabled on them, and `torch_peak_mib` is not written for a CPU run.

		fit (s)			evaluation (s)	
route	K	CPU	GPU	ratio	CPU	GPU
country	30	116.2	14.0	8.3	1.5	0.168
	100	214.8	21.7	9.9	1.5	0.170
	300	604.6	41.7	14.5	1.6	0.178
	1,000	1,357.9	138.9	9.8	2.0	0.173
	3,000	1,607.9	139.2	11.5	1.5	0.168
gridded	30	104.4	16.5	6.3	265.1	24.3
	100	237.8	24.0	9.9	258.5	24.2
	300	595.9	52.3	11.4	261.4	24.0
	1,000	1,428.8	154.2	9.3	240.6	24.0
	3,000	1,465.0	154.8	9.5	235.5	23.9

D.4 Inference throughput

Table 29 times a forward pass through the full model on both routes and both devices, five timed repeats after three warm-ups, reporting the median. The coordinate route is fed uniform random (lon, lat) and the country route the 247 registered countries sampled with replacement, both at year 2015. Throughput saturates by a batch of 512 on the GPU, at 9,770 samples/s on the coordinate route and 10,708 on the country route, and by a batch of 4,096 on the CPU at 778 and 868 samples/s; at batch 1 the model is latency-bound, 9.6 ms per coordinate call and 6.6 ms per country call on the GPU. Peak torch memory at the largest batch is 3,151.0 MiB on the coordinate route and 3,154.8 MiB on the country route, roughly 3.1 GiB of the card's 7,682.8 MiB, and the record marks batch 16,384 `largest_fitting` on every route and device with no cell reporting an out-of-memory status, so no memory ceiling was reached in this sweep. The whole-device `nvidia-smi` peak at that batch, 6,577 and 6,581 MiB, includes everything else resident on the display GPU and is not the model's footprint.

Table 29: **Forward-pass throughput and memory.** Median of five timed repeats after three warm-up passes, both routes, both devices, at year 2015. Peak memory is `torch.cuda.max_memory_allocated` for the cell; it is not recorded on CPU rows (*n.r.*). Every cell returned status `ok`, and the record marks batch 16,384 `largest_fitting` on all four route/device combinations.

	coordinate route			country route		
batch	samples/s	μs/sample	peak MiB	samples/s	μs/sample	peak MiB
RTX 5070 Laptop GPU						
1	104.4	9,575.4	1,516.8	152.4	6,563.1	1,516.8
8	855.7	1,168.6	1,517.5	1,198.5	834.4	1,517.5
64	5,033.7	198.7	1,523.1	6,152.7	162.5	1,523.1
512	9,770.4	102.3	1,469.7	10,708.0	93.4	1,469.8
4,096	9,999.8	100.0	1,925.3	10,228.9	97.8	1,926.3
16,384	9,092.1	110.0	3,151.0	10,151.7	98.5	3,154.8
CPU, 24 threads						
1	22.9	43,665.5	*n.r.*	21.8	45,983.2	*n.r.*
8	149.7	6,680.8	*n.r.*	139.2	7,181.1	*n.r.*
64	484.3	2,064.8	*n.r.*	395.5	2,528.7	*n.r.*
512	732.4	1,365.4	*n.r.*	625.4	1,599.1	*n.r.*
4,096	777.6	1,286.0	*n.r.*	868.3	1,151.7	*n.r.*
16,384	764.6	1,307.9	*n.r.*	874.6	1,143.4	*n.r.*

D.5 FLOPs and parameters per component

Table 30 breaks the forward pass down by component. The components account for the model exactly: their parameters sum to 366,304,274, the model total, with `params_unaccounted` = 0, and their FLOPs sum exactly to the recorded route totals of 729,823,838,208 on the coordinate route and 640,189,464,576 on the country route at a batch of 1,024, that is 712.72

and 625.19 MFLOP per sample. The two fusion stacks do almost all of the arithmetic, 85.2% of the coordinate route and 97.2% of the country route, and the difference between the routes is the location encoder, 91.96 MFLOP per sample and 12.9% of the coordinate route, which is idle on the country route where the country encoder and the geometry routing together cost 4.42 MFLOP per sample instead. The task encoder registers no matmul-class work although it holds 1,048,576 parameters, being an embedding lookup, and neither the temporal transport operator nor the alignment heads are on the inference path. The decoder line is the hypernetwork decoder as a whole; within its 9.10 MFLOP per sample the record attributes 8.39 to the first-layer weight generator, 0.52 to the hypernetwork trunk and 0.03 to the bias generator, identically on both routes, and this nested breakdown is not an additional component.

Table 30: **Parameters and forward FLOPs per component.** Measured with `torch.utils.flop_counter.FlopCounterMode` at a batch of 1,024 and year 2015, counting matmul, convolution and scaled-dot-product-attention class operations at two FLOPs per multiply–accumulate; elementwise, activation and normalisation work is not counted. Shares are stored rounded to five decimals and therefore sum to 99.999 rather than 100.

		coordinate route		country route	
component	parameters	MFLOP/sample	share (%)	MFLOP/sample	share (%)
location encoder	50,202,506	91.96	12.90	0.00	0.00
country encoder	2,170,176	0.00	0.00	4.26	0.68
time encoder	2,129,185	4.24	0.60	4.24	0.68
task encoder	1,048,576	0.00	0.00	0.00	0.00
geometry routing	83,456	0.00	0.00	0.16	0.03
temporal fusion	101,394,945	202.60	28.43	202.60	32.41
task·ST fusion	202,590,752	404.82	56.80	404.82	64.75
decoder (hypernetwork)	4,549,572	9.10	1.28	9.10	1.46
temporal transport	2,135,106	0.00	0.00	0.00	0.00
alignment heads	0	0.00	0.00	0.00	0.00
total	366,304,274	712.72	99.999	625.19	99.999

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. ICLR 2017 Workshop Track.

Luca Albergante, Jonathan Bac, and Andrei Zinovyev. Estimating the effective dimension of large biological datasets using Fisher separability analysis. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. doi: 10.1109/IJCNN.2019.8852450.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14927–14937, 2020.

Laurent Amsaleg, Oussama Chelly, Michael E. Houle, Ken-ichi Kawarabayashi, Miloš Radovanović, and Weeris Treeratanajaru. Intrinsic dimensionality estimation within tight localities. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2019. doi: 10.1137/1.9781611975673.21.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Noel A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, New York, NY, revised edition, 1993. ISBN 978-0-471-00255-0. doi: 10.1002/9781119115151.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1): 12140, 2017. doi: 10.1038/s41598-017-11873-y.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016a. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016b. doi: 10.1007/978-3-319-46493-0_38.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020. doi: 10.1002/qj.3803.

Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 2024.

Jiale Kang and Qingyu Yin. MiSS: Revisiting the trade-off in LoRA with an efficient shard-sharing structure. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=gohmWoUSoS. Formerly titled "Bone: Block Affine Transformation as Parameter Efficient Fine-tuning Methods for Large Language Models".

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4347–4355, 2025. doi: 10.1609/aaai.v39i4.32457.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11454.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR, 2018.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, pages 777–784, 2004.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3–12. Springer, 1994. doi: 10.1007/978-1-4471-2099-5_1.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1711.05101.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9134–9142, 2023. doi: 10.1609/aaai.v37i8.26096.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2022. doi: 10.1145/3503250.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):102:1–102:15, 2022. doi: 10.1145/3528223.3530127.

Natural Earth. Natural Earth: Free vector and raster map data. https://www.naturalearthdata.com, 2022. Version 5.1.2, released 13 May 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Laura Poggio, Luis M. de Sousa, Niels H. Batjes, Gerard B. M. Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240, 2021. doi: 10.5194/soil-7-217-2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

Arjun Rao, Ruth Crasto, Tessa Ooms, David Rolnick, Konstantin Klemmer, and Marc Rußwurm. Localized, high-resolution geographic representations with Slepian functions. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.00392.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 0-262-18253-X.

Carlos Rodriguez-Pardo. Geospatial Embeddings Wrapper, 2026. URL https://github.com/crp94/geospatial_embeddings_wrapper.

Carlos Rodriguez-Pardo and Massimo Tavoni. A harmonised dataset for Earth system foundation models. *Scientific Data*, 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-07913-w. URL https://doi.org/10.1038/s41597-026-07913-w.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G. Baraniuk. WIRE: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18507–18516, 2023.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, pages 517–524. ACM, 1968. doi: 10.1145/800186.810616.

J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3D neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20875–20886, 2023.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 7462–7473, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Jan Teorell, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon, and Cem Mert Dalli. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute, 2021. doi: 10.18157/qogstdjan21.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 8690–8701, 2023. doi: 10.52202/075280-0379.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005. ISBN 0-387-00152-2. doi: 10.1007/b106715.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified Copernicus foundation model for Earth vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9888–9899, 2025. Copernicus-FM.