Earth observation embeddings are effective sub-grid descriptors for probabilistic weather downscaling

Pedro Sousa[1], Will Tebbutt[2], Sadiq Jaffer[1], Robin Young[1], Anil Madhavapeddy[1], and Richard E. Turner[2]

[1]Department of Computer Science, University of Cambridge
[2]Department of Engineering, University of Cambridge

Abstract

Global weather reanalyses and forecasts resolve the evolving atmospheric state on coarse grids, but site-specific applications require predictions at arbitrary locations where near-surface conditions also depend on unresolved terrain and land-surface properties. Existing probabilistic downscalers address this gap using hand-crafted topographic descriptors. We ask instead whether Earth observation foundation models can provide transferable sub-grid surface representations for probabilistic weather downscaling.

We augment a convolutional conditional neural process that downscales coarse ERA5 reanalysis fields at ∼25 km resolution with a learned local surface descriptor, obtained by compressing a patch of Tessera embeddings at 10 m resolution. Although these embeddings summarise surface conditions over annual timescales, they improve downscaling of instantaneous 2 m temperature and 10 m wind speed by encoding persistent surface properties that capture a location's departure from the coarse-grid atmospheric state. Across five climatically diverse regions, the embedding improves point and probabilistic skill at stations held out in both space and time, overall improving CRPS skill by 11.5% for 2 m temperature and 6.2% for 10 m wind speed. We further analyse how its contribution differs by variable, finding that topography explains more of temperature's sub-grid structure, while Tessera provides additional surface information for wind speed.

These improvements persist when the coarse input is changed from ERA5 reanalysis fields to forecasts from the Aurora AI forecasting model, and when predicting at newly deployed stations with no regional history. To our knowledge, this is the first evidence that long-timescale Earth-observation embeddings can support short-timescale weather downscaling where sub-grid departures are systematically structured by persistent surface properties.

1 Introduction

Weather reanalyses and forecasts are often produced on relatively coarse spatial grids (e.g., 25 km), limiting their ability to represent weather variability at the local scales required by many downstream applications [40, 49]. The near-surface state at a specific location can depart substantially from the cell average captured by a coarse grid because a range of sub-grid processes produces structured variation within individual cells [34].

This sub-grid variation reflects both local surface properties and interactions with the surrounding environment. Land cover affects near-surface temperature within a single coarse cell: the urban–rural contrast associated with the urban heat island is typically 1–3 °C hotter [1] and can be considerably larger on clear, calm nights [37]. Terrain also shapes temperature through processes such as cold-air drainage, whereby dense air pools in valley floors and hollows that can sit 4–8 °C colder than slopes only tens of metres higher at the same coarse-grid elevation [21].

Wind is similarly shaped by local and non-local effects, accelerating where it is channelled along valleys or through terrain gaps and slowing over forest canopy or built surfaces relative to open ground and water. Recent kilometre-scale forecasts over Texas have improved near-surface wind speed predictions by ∼23% over ECMWF's operational global forecasts, with corresponding gains for wind-power forecasting [27]. These effects occur below the resolution of a coarse atmospheric grid and are properties of *where a point sits on the surface* rather than of the large-scale flow alone.

Statistical downscaling is the standard approach for capturing such sub-grid variation by learning relationships between coarse-scale atmospheric predictors and local observations. This includes both a perfect prognosis (PP), where models are trained using large-scale fields analysed, such as reanalysis products, and model output statistics (MOS), where they are trained directly on model outputs forecast [32]. Under either framing, the aim is to recover the local value at a target location from the resolved large-scale state together with information about the unresolved local environment. Generalising to *unseen locations at unseen times* holds the most operational value and is this paper's central task: predicting at a point that contributed no data during training (e.g., a newly instrumented site, or a location that will never host a sensor), and doing so at future dates. This is the hardest test of a downscaling model, but where much of its deployment value lies.

An alternative to off-grid downscaling is grid-to-grid super-resolution modelling that maps a coarse field to a finer one, thereby avoiding an explicit off-grid mapping if the predicted field's resolution is sufficient. Examples include convolutional regression [6, 25] and generative models that sample a fine-scale field: GANs for precipitation [23] and for wind over complex terrain [36], and diffusion models reaching kilometre scale [33]. However, depending on the target resolution, the output grid still carries sub-grid variation that needs resolving to serve downstream decisions at specific locations. Additionally, grid-to-grid methods require dense, high-resolution target fields for training, often generated by regional NWP models, so the learned downscaler may inherit biases and structural assumptions from its model-derived targets.

Closer to our setting, several station-level forecasting and downscaling methods improve local predictions by conditioning on observations from the target station or incorporating station measurements into a high-resolution prediction framework [4, 48]. These, by construction, fit only locations already served with weather stations. In contrast, we deliberately withhold local history, requiring the downscaling model to predict at previously unobserved sites.

Generalising beyond instrumented sites means predicting directly at arbitrary target locations relying only on the coarse atmospheric fields and local surface descriptors, while being trained against ground-truth station observations. Although this replaces dense gridded supervision with sparse, scattered point targets, it defines the downscaling objective directly against station measurements rather than model-derived high-resolution fields, consistent with a recent line of observation-driven forecasting[1].

[1] Allen et al. [3] encoder ingests raw observations as *input* to estimate the global atmospheric state, replacing data assimilation, while a downstream ConvCNP decoder targets station observations to produce local forecasts. It uses reanalysis data for training, but not in deployment. ECMWF's direct-observation-prediction line [2] goes one step

Off-grid, multi-site, probabilistic prediction from a gridded field is precisely the setting the convolutional conditional neural process (ConvCNP) [20], a translation-equivariant member of the conditional neural process family [17], is designed for. Vaughan et al. [45] applied a ConvCNP to map a coarse reanalysis grid to a predictive distribution at arbitrary station locations, trained by maximum likelihood learning on station data. The same mechanism was adopted for the downscaling stage of Allen et al. [3], one of the first end-to-end data-driven global forecasting models.

In Vaughan et al. [45]'s ConvCNP, the sub-grid signal enters through a hand-crafted topographic descriptor specific to the station via three scalars summarising elevation and terrain position. Terrain, however, is only one axis of the near-surface state, as land cover, canopy structure, surface roughness, soil moisture, water and built structure all modulate how a point departs from its coarse-cell mean. Earth-observation (EO) geospatial foundation models complement this hand-crafted enumeration with a single learned representation of the broader surface state. TESSERA [13] is trained self-supervised on a full year of paired Sentinel-1 radar and Sentinel-2 optical imagery and emits a 128-dimensional embedding for every 10 m pixel, jointly encoding these surface characteristics without requiring them to be specified in advance. It is precomputed and globally available at any terrestrial query location, including sites that have never hosted an instrument. EO embeddings of this kind have been applied mainly to land-cover, crop and change-detection mapping; we are not aware of prior use as a sub-grid descriptor for near-surface weather.

This work's main contributions are as follows:

1. **Construct a learned sub-grid descriptor from EO foundation-model embeddings.** A single TESSERA pixel is too local to capture the range of near-surface properties affecting a weather-station observation, so we pre-train a VAE to compress a neighbourhood of per-pixel embeddings into a compact local surface descriptor (§2.2).

2. **Demonstrate consistent gains in off-grid probabilistic downscaling.** Injecting the TESSERA descriptor into the ConvCNP's decoder alongside topography improves point skill (MAE and RMSE) and probabilistic skill (CRPS) across five climatically diverse regions for 2 m temperature and 10 m wind speed (§3.1).

3. **Differentiate qualitatively how the embedding corrects temperature and wind.** For 2 m temperature, most fine-scale structure is expressible through elevation, so the embedding contributes little additional spatial texture and instead acts primarily as a transferable land-surface prior where scarce observations leave relevant surface conditions poorly represented. For 10 m wind speed, the residual structure is better organised by the embedding, indicating that topography alone is an incomplete near-surface descriptor (§3.2, §3.3).

4. **Demonstrate the method's robustness to forecast weather fields.** Replacing ERA5 with Aurora forecast fields [7], the embedding's uplift persists across lead times, demonstrating its value in an operationally realistic weather forecasting deployment (§3.4).

5. **Show that the embedding improves sample efficiency.** In a simulated Norwegian network ramp-up, TESSERA provides competitive wind-speed downscaling before any local observations are available, already outperforming interpolated ERA5 at zero deployment; the *no-TESSERA* baseline fails to reach the same error level even after 6 years of accumulated local data. This shows that the data efficiency reported for TESSERA upstream [13] carries through to weather downscaling (§3.5).

further, learning the forecast dynamics themselves directly from observations with no reanalysis.

2 Data and methods

We first describe the coarse atmospheric inputs, station observations, and per-location surface descriptors used by the model. We then introduce the ConvCNP downscaling framework, including how the TESSERA surface descriptor is constructed at arbitrary target locations, before detailing the predictive distributions, training objective, and evaluation protocol. We recap only the ConvCNP components needed for the present method and refer to Vaughan et al. [45] for the full architecture; additional implementation details are provided in Appendix A.

2.1 Data

Coarse weather grid. We downscale ERA5 reanalysis [24], a global product on a 0.25° ($\sim 25\,\mathrm{km}$) grid, cropped to a bounding box for each region studied. We use 20 dynamic atmospheric variables, including 2 m temperature and the 10 m eastward and northward wind components, and 13 static surface properties, such as geopotential and other coarse land-surface variables. The full predictor set used by the downscaling module is listed in Table 3. In §3.4, we replace the dynamic ERA5 fields with Aurora forecast grids [7] generated at the same 0.25° resolution.

Station observations. Targets are point observations from the Global Historical Climatology Network–Hourly (GHCNh) station network [35]. We predict two instantaneous near-surface variables at 6-hourly UTC snapshots: 2 m temperature (t2m) and 10 m wind speed. Unlike Vaughan et al. [45], who target daily aggregates, we use instantaneous values because they align directly with forecasting-system outputs at individual lead times and are therefore the relevant quantities for the forecast-driven setting of §3.4.

We evaluate across five climatically diverse regions—Europe, the United States, East Asia, Southern Africa, and Australia—chosen to span a wide range of terrain, land cover, and station density. We use GHCNh rather than the curated station sets of the VALUE intercomparison [22] or ECA&D because those contain only European stations and were assembled for daily aggregates, whereas we require sub-daily observations and multi-continental coverage.

Land-surface embeddings. TESSERA [13] is trained via self-supervised learning to provide a 128 dimensional embedding for any 10 m pixel from a year-long time series of Sentinel-1 and Sentinel-2 observations. Each embedding summarises the characteristic annual surface dynamics at a location. To limit storage requirements, we use the 2017 embedding map for all training, validation, and test snapshots, generated using the 1B-parameter model from TESSERA v2 [14]. The downscaler is trained on approximately a decade of historical data and tested on 2022, with the embedding year lying within the training period. This assumes that the surface characteristics and seasonal patterns encoded in 2017 remain sufficiently stable across the study period; changes in land cover, vegetation, snow regime, or the built environment are therefore not represented.

Topography. Following Vaughan et al. [45], we retain an explicit topographic descriptor, $\mathbf{e}$, comprising three features: elevation; Δelevation, defined as the difference between the true elevation at the off-grid location and ERA5 orographic height, obtained from surface geopotential and linearly interpolated to that location; and the multi-scale topographic position index, mTPI [43], which indicates whether a location lies in a valley or on a ridge.

2.2 Patch compression of TESSERA embeddings

A single TESSERA pixel describes a 10 m spot, which is too local to capture all the surface properties affecting a near-surface weather measurement. However, directly supplying all 128-dimensional embeddings from a 64×64 pixel patch introduces a 524,288-dimensional input, which is impractically large and would risk severe overfitting of the downscaler. Instead, we learn a compact, task-agnostic summary of the neighbourhood using an unsupervised variational autoencoder (VAE), rather than relying on the downscaler itself to learn this bottleneck from sparse station observations. Specifically, the VAE compresses a ~ 640 m neighbourhood centred on each location into a 16-dimensional vector $\mathbf{z}_T$ that captures terrain context, land-cover heterogeneity, and coastal–inland transitions.

The patch is deliberately local: it characterises the immediate surroundings that control how much a weather station's observations depart from the cell mean, rather than summarising the full ~ 25 km cell. The compression rate is severe (2^{15}) and deliberately simple, with larger latent representations, multi-scale patches, stronger encoder architectures, and end-to-end fine-tuning of the VAE left as natural extensions.

2.3 Downscaling model

The backbone is the convolutional conditional neural process of Gordon et al. [20], as instantiated for weather downscaling by Vaughan et al. [45]. Its input is a coarse atmospheric grid $Z \in \mathbb{R}^{D_{\text{lat}} \times D_{\text{lon}} \times C}$, where D_{lat} and D_{lon} are the numbers of latitude and longitude grid points in the regional crop, and C is the number of predictor channels at each grid point, as listed in Table 2. The model maps Z to a predictive distribution $p(y_\star \mid \mathbf{x}_\star, Z)$ at an arbitrary query location $\mathbf{x}_\star = (\text{lat}, \text{lon})$, where $y_\star$ denotes the corresponding target weather observation, in three stages.

First, Z is passed through a convolutional encoder, producing a feature vector at every grid point, $H = \text{CNN}(Z)$. We write $\mathbf{h}_{nm}$ for the feature vector at the grid point indexed by n along longitude and m along latitude, with Δx_1 and Δx_2 the corresponding grid spacings, so that this point sits at coordinates $(n\Delta x_1,\, m\Delta x_2)$.

Second, these gridded features are evaluated at the off-grid target $\mathbf{x}_\star = (x^{(1)},\, x^{(2)})$ by convolution with a smooth exponentiated-quadratic kernel φ,

$$\varphi_c(H,\, \mathbf{x}_\star) \;=\; \sum_{n,m} h_{nm}\, \exp\!\left(-\frac{\big(x_\star^{(1)} - n\Delta x_1\big)^2}{2\ell_1^2} - \frac{\big(x_\star^{(2)} - m\Delta x_2\big)^2}{2\ell_2^2} \right), \tag{1}$$

where ℓ_1, ℓ_2 are learned length-scales. The sum runs over the grid points of the regional crop, each weighted by its distance to $\mathbf{x}_\star$, so contributions from nearby cells dominate and the result is defined for any $\mathbf{x}_\star$, on or off the grid, carrying no sub-grid information.

Third, an MLP maps this representation, concatenated with the per-location descriptors, to the parameters $\boldsymbol{\theta}$ of the predictive distribution,

$$\boldsymbol{\theta}(\mathbf{x}_\star,\, Z,\, \mathbf{e},\, \mathbf{z}_T) \;=\; \psi_{\text{MLP}}\big[\, \varphi_c(H,\, \mathbf{x}_\star)\ ,\ \mathbf{e}\ ,\ \mathbf{z}_T\, \big], \tag{2}$$

where $\mathbf{e}$ is the topographic descriptor and $\mathbf{z}_T$ the TESSERA surface descriptor. For brevity, we write $\boldsymbol{\theta}$ hereafter, with its dependence on $\mathbf{x}_\star$, Z, $\mathbf{e}$, and $\mathbf{z}_T$ understood. Since Z contains information only at the coarse-grid scale, no mapping of its encoded representation can recover location-specific structure that is absent from the grid itself. The per-location descriptors $\mathbf{e}$ and $\mathbf{z}_T$ are therefore the only source of local sub-grid information available.

Predictive heads and training objective. Given a context grid Z, let $\mathbf{x}_\star^{(1:N)} \equiv \{\mathbf{x}_\star^{(n)}\}_{n=1}^N$ denote the set of N target locations and $y_\star^{(1:N)} \equiv \{y_\star^{(n)}\}_{n=1}^N$ the corresponding observations. Following the conditional neural process formulation, we model each target prediction through a separate conditional marginal $p_{\boldsymbol{\theta}}\big(y_\star \mid \mathbf{x}_\star, Z\big)$, where $\boldsymbol{\theta}$ is the parameter vector produced at that location by eq. (2). Training minimises mean negative marginal log likelihood across the valid target observations,

$$\ell(\boldsymbol{\theta}; y_\star^{(1:N)}) = -\frac{1}{N}\sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}\big(y_\star^{(n)} \mid \mathbf{x}_\star^{(n)}, Z\big). \tag{3}$$

The normalisation ensures that snapshots contribute equally to training regardless of how many stations report valid observations at that time. This yields predictive marginals at arbitrary locations but does not explicitly model residual dependence between targets; Appendix A discusses the corresponding factorized joint interpretation and its implications for spatial uncertainty. All reported metrics are per-observation marginal metrics.

Both variables use the same two-parameter form $\boldsymbol{\theta} = (\mu, \sigma)$, differing only in the distribution family. For 2 m temperature, $p_{\boldsymbol{\theta}}$ is a Gaussian $\mathcal{N}(\mu, \sigma^2)$, where the per-observation loss is

$$\ell_{\text{Gauss}}(\mu, \sigma; y_\star) \equiv -\log p_{\boldsymbol{\theta}}\big(y_\star \mid \mathbf{x}_\star, Z\big) \;=\; \tfrac{1}{2}\log\big(2\pi\sigma^2\big) + \frac{(y_\star - \mu)^2}{2\sigma^2}. \tag{4}$$

For 10 m wind speed, a Gaussian is not appropriate, as it assigns probability to negative speeds. Instead, we define $p_{\boldsymbol{\theta}}$ as a Truncated Normal distribution over $y_\star \geq 0$:

$$\ell_{\text{TN}}(\mu, \sigma; y_\star) \equiv -\log p_{\boldsymbol{\theta}}\big(y_\star \mid \mathbf{x}_\star, Z\big) \;=\; \tfrac{1}{2}\log\big(2\pi\sigma^2\big) + \frac{(y_\star - \mu)^2}{2\sigma^2} + \log\Phi\Big(\tfrac{\mu}{\sigma}\Big). \tag{5}$$

where Φ is the CDF of the standard Gaussian distribution and $\Phi(\mu/\sigma)$ is the truncation normaliser. It is a well-established predictive family for wind speed introduced by Gneiting et al. [19] and developed as the standard EMOS formulation for non-negative quantities by Thorarinsdottir and Gneiting [44]. This is the only departure from the predictive families used by Vaughan et al. [45], which pair a Gaussian head for daily maximum temperature with a Bernoulli–Gamma head for precipitation. We target wind speed, identified there as a natural extension, and defer precipitation to future work.

2.4 Experimental setup

Splits. We hold out along two independent axes to evaluate temporal and spatial generalisation simultaneously. *Temporally*, the 6-hourly snapshots are split by date: 2010–2020 for training, 2021 for validation, and 2022 as the test year. *Spatially*, within each region, a fixed random 15% of stations is withheld from training entirely. Every reported score is therefore evaluated at a station the model never saw, on a date it never trained on. Figure 1 shows the resulting partition, which we reuse across variables and experiments, except in §3.5, where Norwegian stations are withheld on a schedule and occupy a contiguous sub-domain. All results are 3-seeds means.

Station density. To relate a region's skill gain to how sparsely it is observed, we quantify sparsity as station density ρ: stations per million square kilometres of the region's bounding box, computed per variable over that region's valid-observation station set. Density spans more than an order of magnitude across our regions (Figure 1).

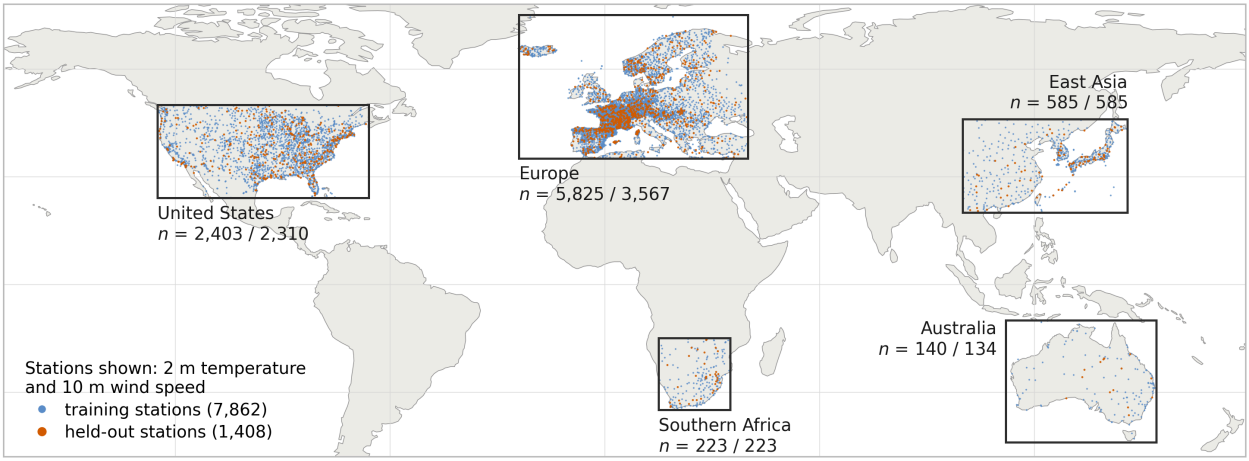


Figure 1: The five evaluation regions and their respective station networks. Black rectangles are the ERA5 crop boxes of Table 3; the United States domain includes some stations in southern Canada, while the Europe domain includes stations in parts of North Africa and western Asia. Dots are the within-region, shortlisted GHCNh stations that pass the selection outlined in Appendix A, for 2 m temperature and 10 m wind speed, colored according to their training membership: blue stations enter training, orange stations are withheld entirely. Region labels give the total station count n as *2 m temperature / 10 m wind speed*. Note the more than an order-of-magnitude spread in network density across regions, from Europe to Australia.

Reference models. We compare against three references of increasing strength. *Persistence* carries forward the most recent prior observation at the station, providing a naive lower bound evaluated only where a valid preceding observation exists.

ERA5 interpolation bilinearly interpolates the coarse field to the station location. For 2 m temperature, we additionally correct for the elevation difference between the station and the interpolated ERA5 orography using $\hat{y}(\mathbf{x}_\star) \equiv y^{\mathrm{ERA5}}(\mathbf{x}_\star) - \Gamma \cdot \Delta\mathrm{elevation}(\mathbf{x}_\star)$ [10, 42], where $\Delta\mathrm{elevation}(\mathbf{x}_\star)$ is the same station-to-grid elevation difference supplied to the ConvCNP and Γ is fitted by least squares on the training split separately for each region. The fitted correction performs at least as well as both uncorrected interpolation and a fixed textbook lapse rate of $6.5\,^{\circ}\mathrm{C\,km^{-1}}$ [10], so we report only the fitted version. No analogous correction is applied to 10 m wind speed since local wind-speed variations depend on terrain exposure, surface roughness, and the incident flow, and cannot generally be represented by a fixed correction based only on station-to-grid elevation difference [46, 47].

The *no-TESSERA ConvCNP* uses the topographic descriptor $\mathbf{e}$, but without the TESSERA surface descriptor $\mathbf{z}_T$. For this baseline, we retain the static ERA5 surface fields in the context grid to provide coarse proxies for surface properties otherwise unavailable to the model. When TESSERA is included, we omit these fields because the embedding supplies a richer station-specific surface descriptor and, empirically, retaining them degrades held-out performance.

Metrics. We report point accuracy with MAE and RMSE, and probabilistic skill with CRPS, evaluated on held-out observations. CRPS is a proper scoring rule that jointly assesses calibration and sharpness, expressed in the target's physical units [18]. We evaluate CRPS in closed form for the Gaussian head, and from $M = 200$ draws of the fitted Truncated Normal using the fair estimator [15]. Its residual Monte Carlo variance has a negligible effect on the aggregated scores. A closed-form solution is also available [44], but $\Phi(\mu/\sigma)$ enters in the denominator and loses numerical precision

in the low-wind regime. CRPS is only defined for the ConvCNP variants, not for the deterministic interpolation and persistence references.

MAE and RMSE are minimised by the median and the mean, respectively. For the Gaussian head, both the predictive median and mean coincide and are given by μ. For the Truncated Normal, the optimal estimators are different and computed in closed form from (μ, σ). Throughout the paper, for wind speed, MAE is evaluated against the predicted median and RMSE against the mean.

3 Experimental Results

This section assesses how TESSERA affects downscaling performance across complementary spatial and operational settings. We first compare predictive skill across regions and weather variables (§3.1), then examine how the embedding changes the fine-scale structure of dense downscaled fields (§3.2) and which descriptor spaces organise the sub-grid correction (§3.3). Finally, we test whether these benefits persist when the input weather field is medium-range Aurora forecasts rather than ERA5 (§3.4), and assess the embedding's sample-efficiency gains during the simulated deployment of a new station network in Norway (§3.5).

3.1 Per-region skill

Table 1 shows that adding TESSERA embeddings improves the ConvCNP in all ten region × variable settings and across all metrics, both deterministic (MAE and RMSE), and probabilistic (CRPS).

The ERA5 interpolation baseline reveals a dependence on station availability for 2 m temperature. In the densely observed European and US regions, the no-TESSERA ConvCNP outperforms lapse-corrected interpolation. However, in East Asia, Southern Africa, and Australia, it is worse than interpolation, suggesting that the model struggles to learn a transferable local correction from sparse station data using topography alone. Adding TESSERA reverses this pattern in East Asia and Australia and nearly closes the gap in Southern Africa, improving RMSE over interpolation but remaining worse on MAE. For 10 m wind speed, both ConvCNP variants outperform interpolation in every region, with TESSERA providing a further consistent gain.

Complementary controls probe the source of these performance gains. Appendix B shows that station-specific descriptor alignment matters for the improvement and compares the VAE encoder with simpler patch statistics. Appendix C controls for TESSERA's richer surface information from Sentinel-1(2) imagery by augmenting the no-TESSERA baseline with the 17 hand-crafted features of Bakketun et al. [5]. These include land-cover fractions, canopy height, and neighbourhood terrain statistics. Although this improves performance in most settings, it recovers only a minority of TESSERA's benefit: averaged across regions, CRPS decreases by 3.2% for 2 m temperature and 2.2% for wind speed, compared with 11.5% and 6.2% using TESSERA. The enriched baseline does not match ConvCNP with TESSERA in any of the 30 region × variable × metric comparisons.

Figure 2 plots the percentage improvement of ConvCNP with TESSERA vs the no-TESSERA baseline. For 2 m temperature, the uplift generally grows as station availability decreases. Together with the comparison against lapse-corrected ERA5 interpolation, this suggests that the embedding acts as a transferable land-surface prior: where observations are abundant, the baseline can learn much of the local correction from topography and station data, whereas in sparsely observed regions, the

Table 1: Per-region results, ordered by station count. Total station counts n (training + held-out) and network densities ρ given as *t2m / wind*. Metrics are reported on the held-out stations at held-out years, with the lowest error per region×metric in **bold**. CRPS is defined only for the probabilistic ConvCNP heads. Each value is the 3-seeds mean, given the small per-cell cross-seed std (≤ 0.02 for most entries, rising to $\sim$0.05–0.07 in Southern Africa and Australia).

		2 m temperature (°C)			*10 m wind speed ($m\,s^{-1}$)*		
Region	Model	MAE	RMSE	CRPS	MAE	RMSE	CRPS
Europe $n = 5{,}825/3{,}567$ $\rho = 322/197$	Persistence	3.53	4.94	—	1.49	2.13	—
	ERA5 interp. (+lapse for t2m)	1.35	1.95	—	1.44	2.15	—
	ConvCNP (no TESSERA)	1.17	1.67	0.84	1.28	1.87	0.92
	ConvCNP with TESSERA	**1.10**	**1.58**	**0.79**	**1.19**	**1.74**	**0.86**
United States $n = 2{,}403/2{,}310$ $\rho = 159/153$	Persistence	3.80	5.29	—	1.68	2.32	—
	ERA5 interp. (+lapse for t2m)	1.52	2.23	—	1.63	2.12	—
	ConvCNP (no TESSERA)	1.41	2.07	1.03	1.45	1.96	1.04
	ConvCNP with TESSERA	**1.30**	**1.95**	**0.95**	**1.36**	**1.84**	**0.97**
East Asia $n = 585/585$ $\rho = 47/47$	Persistence	3.07	4.05	—	1.46	2.04	—
	ERA5 interp. (+lapse for t2m)	1.22	1.70	—	1.44	1.99	—
	ConvCNP (no TESSERA)	1.35	1.87	0.97	1.23	1.71	0.88
	ConvCNP with TESSERA	**1.10**	**1.53**	**0.79**	**1.16**	**1.62**	**0.84**
Southern Africa $n = 223/223$ $\rho = 50/50$	Persistence	5.36	6.90	—	1.48	2.07	—
	ERA5 interp. (+lapse for t2m)	**1.69**	2.56	—	1.38	1.83	—
	ConvCNP (no TESSERA)	2.00	2.75	1.45	1.23	1.65	0.88
	ConvCNP with TESSERA	1.79	**2.55**	**1.31**	**1.16**	**1.57**	**0.84**
Australia $n = 140/134$ $\rho = 8.9/8.5$	Persistence	5.60	6.62	—	2.03	2.61	—
	ERA5 interp. (+lapse for t2m)	1.40	2.06	—	1.46	1.88	—
	ConvCNP (no TESSERA)	1.57	2.16	1.15	1.42	1.81	1.02
	ConvCNP with TESSERA	**1.32**	**1.92**	**0.98**	**1.29**	**1.69**	**0.95**
All regions	Persistence	4.27	5.56	—	1.63	2.23	—
	ERA5 interp. (+lapse for t2m)	1.44	2.10	—	1.47	1.99	—
	ConvCNP (no TESSERA)	1.50	2.10	1.09	1.32	1.80	0.95
	ConvCNP with TESSERA	**1.32**	**1.90**	**0.96**	**1.23**	**1.69**	**0.89**

embedding supplies information that the baseline cannot estimate reliably.

For 10 m wind speed, the uplift is approximately constant across station counts, bar the noisy regions with minimal held-out stations. This suggests that, over the range of sample sizes considered, increasing the number of stations does not remove TESSERA's advantage. Topography alone appears insufficient to locally correct the coarse wind field, while the TESSERA descriptor provides missing land-surface information helpful for wind speed downscaling. The difference in how both variables' performance responds to TESSERA's signal is further probed in the following sections.

3.2 Dense maps

The ConvCNP allows users the flexibility to interpolate weather variables at any off-grid location. However, to better visualise the correction granularity with and without TESSERA, we run both models as dense 0.05° ($\sim$5 km) fields over Iberia and Norway, and ask how much sub-grid structure each resolves below the coarse grid. Both models take terrain elevation from the SRTM 1-arc-second

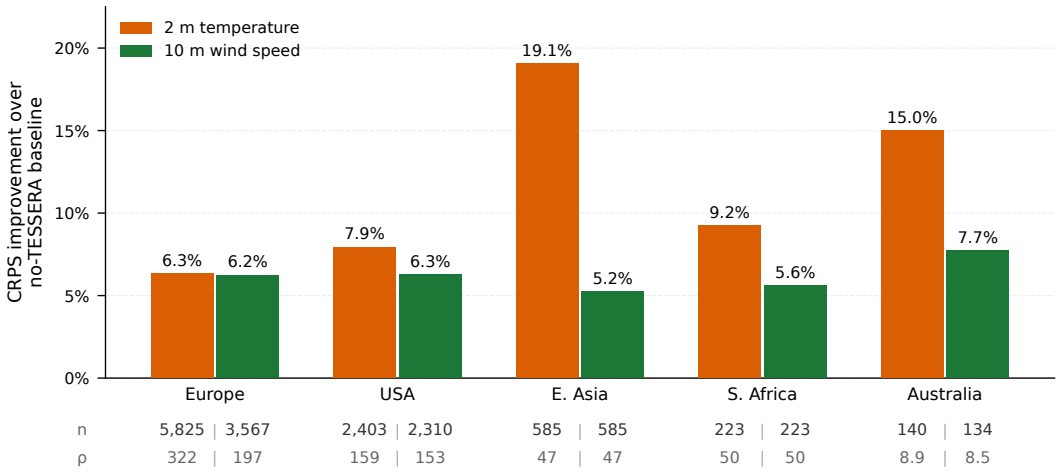


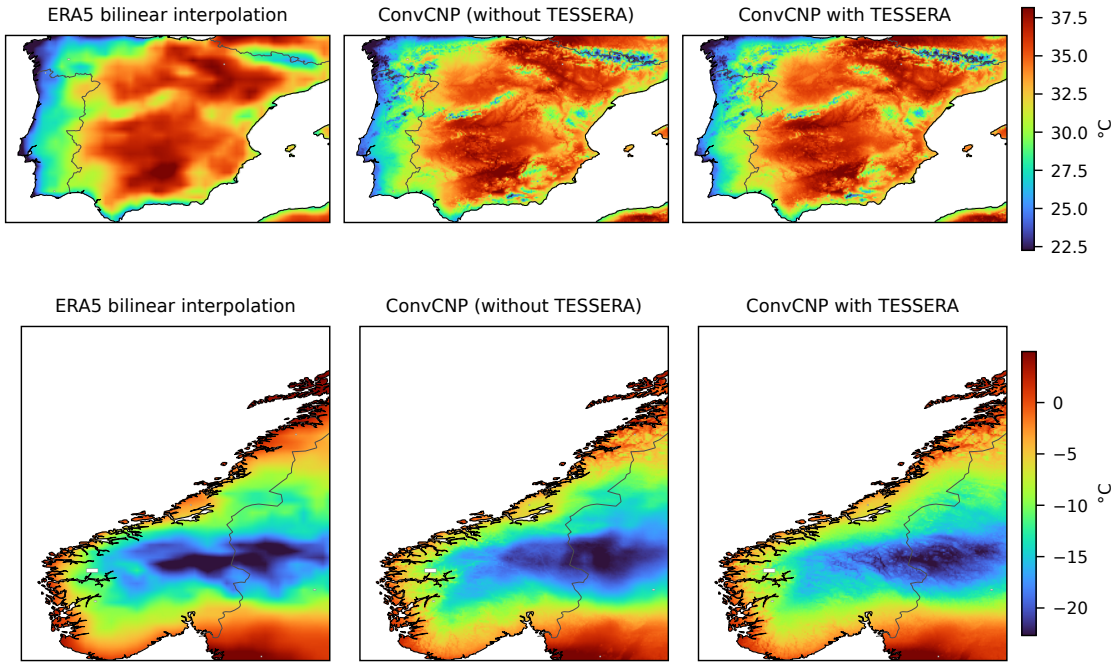
Figure 2: CRPS percentage reduction of ConvCNP with TESSERA vs the no-TESSERA baseline, per region, evaluated on the held-out stations at held-out years. Regions are ordered by station count, high to low; n and ρ are the total network size and density for each variable.

DEM [12], accessed through the Mapzen Terrain Tiles dataset hosted on AWS [31], sampled at each 0.05° cell. The mTPI feature is not defined at map scale and is omitted, so elevation and Δelevation are the sole topographic descriptors supplied to either model.

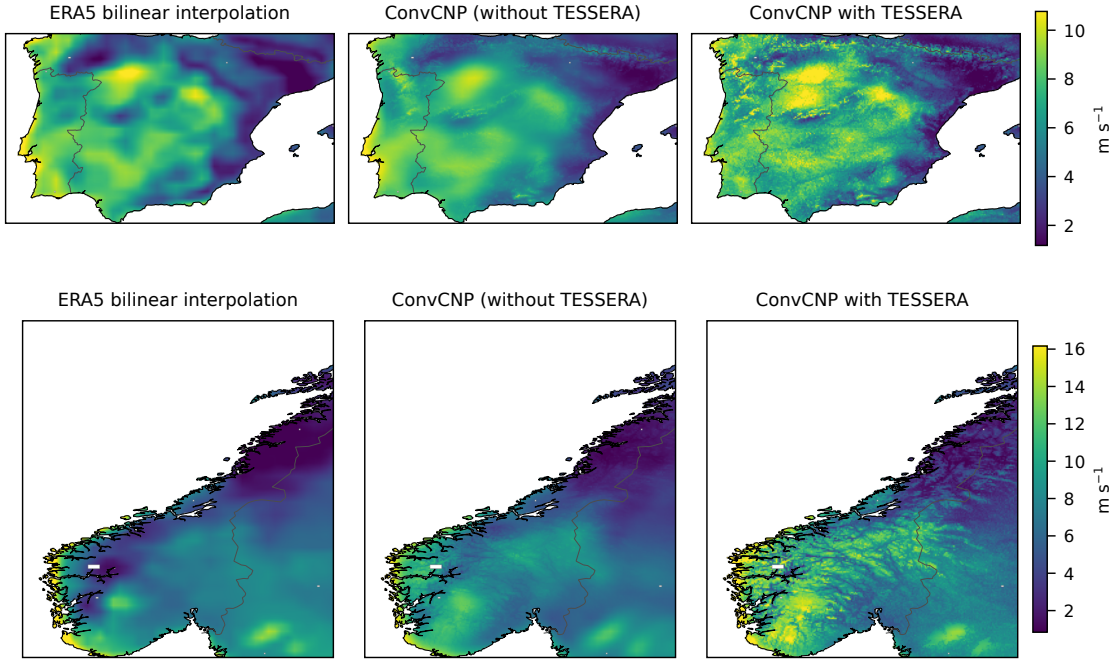
The embedding affects temperature and wind differently. To quantify the fine-scale structure visible in Figure 3, we measure the amplitude of variation below the coarse-grid scale and compare the two model variants (Appendix D). For 2 m temperature, the models resolve almost identical fine-scale amplitudes, with texture ratios of 1.02 in Iberia and 1.08 in Norway. Elevation thus captures most of the amplitude of the resolved temperature structure. The increment maps in Figure 4 nevertheless reveal coherent local warming and cooling patterns, showing that TESSERA can modify *where* and by *how much* this structure is expressed without substantially increasing its total amplitude.

For 10 m wind speed, the effect is qualitatively different: TESSERA increases fine-scale amplitude by 3.28× in Iberia and 3.80× in Norway, with structured increments remaining after removal of the domain-mean shift. Therefore, the embedding contributes location-dependent wind corrections beyond a uniform regional adjustment and beyond the structure represented effectively by elevation alone. Furthermore, independent verification against station networks finds ERA5 near-surface wind to be systematically biased and spatially oversmoothed relative to measurements [48], so adding fine-scale amplitude to the wind field is the expected direction of correction.

Because no dense ground-truth field is available, we validate these map differences at station locations using the error-alignment analysis detailed in Appendix D. The TESSERA increment points in the direction of the baseline error at comparable rates for wind and temperature (69% and 66%, respectively), showing that the local changes are informative for both variables. Their explanatory strength differs substantially, with the increment explaining 56% of the baseline-error variance for wind but only 14% for temperature. TESSERA makes useful local refinements to an already largely elevation-organized temperature field, whereas for wind, it recovers a larger missing spatial correction.

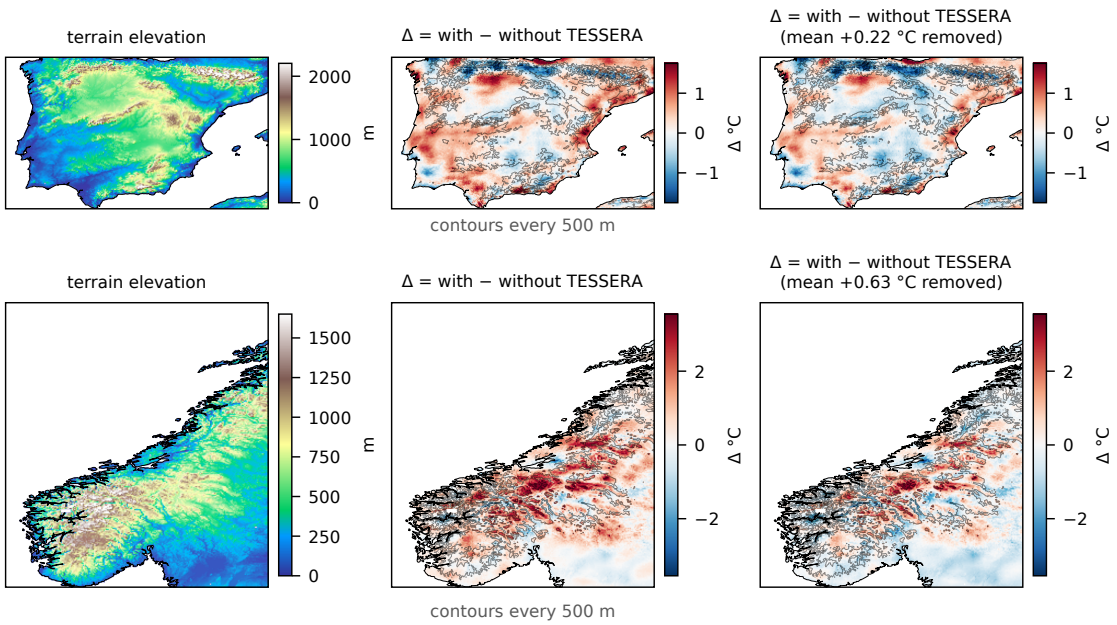


(a) 2 m temperature for Iberia (top; 18 July 2022, 12 UTC) and Norway (bottom; 2 January 2023, 00 UTC). The two ConvCNP variants resolve similar overall fine-scale amplitude.

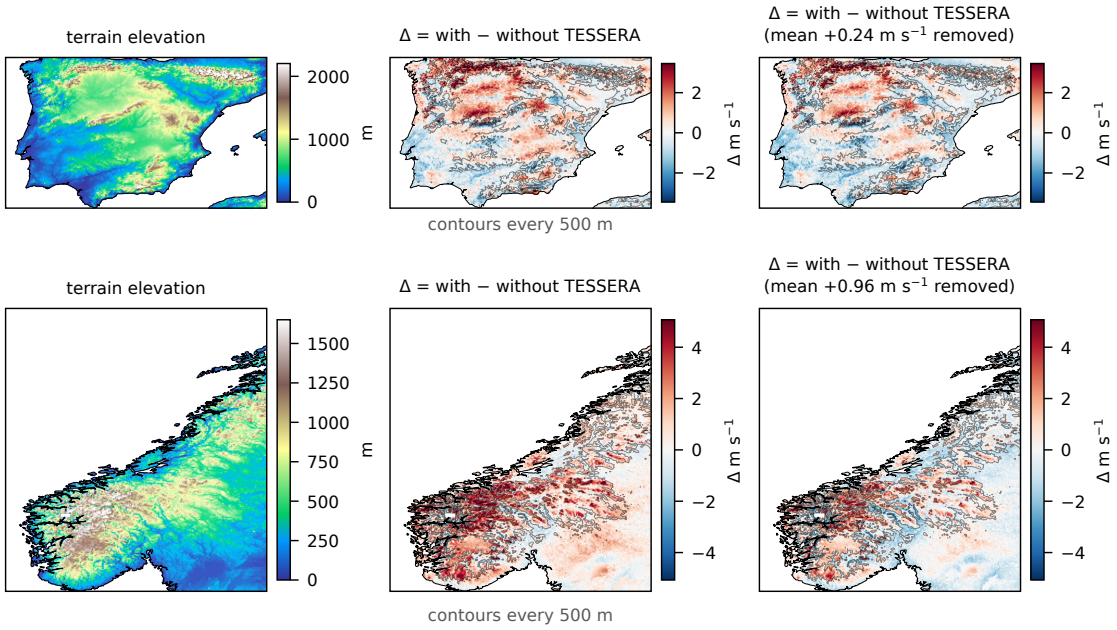


(b) 10 m wind speed for Iberia (top; 12 December 2022, 12 UTC) and Norway (bottom; 30 January 2022, 00 UTC). The TESSERA model resolves substantially stronger fine-scale variation, indicating that the embedding exposes spatial information not captured effectively by elevation only.

Figure 3: Dense 0.05° (∼5 km) downscaling over two climatically distinct regions. Within each map, columns show (left to right) ERA5 bilinear interpolation, no-TESSERA ConvCNP, and ConvCNP with TESSERA, each overlaid on a high-resolution DEM.



(a) TESSERA's increment for 2 m temperature in Iberia (top) and Norway (bottom), shown alongside terrain elevation. Although the total fine-scale amplitude changes little, the embedding introduces coherent spatial adjustments to the predictive mean.



(b) TESSERA's increment for 10 m wind speed in Iberia (top) and Norway (bottom), shown alongside terrain elevation. Strong spatially varying corrections remain in the right-hand panel, after the domain-mean increment has been removed.

Figure 4: Spatial structure of the change in predictive mean induced by TESSERA. Within each map, columns show (left to right) terrain elevation, the raw increment, and the mean-removed increment. Both increment panels are relief-shaded, carry elevation contours at 500 m intervals. The maps show where the two model variants disagree.

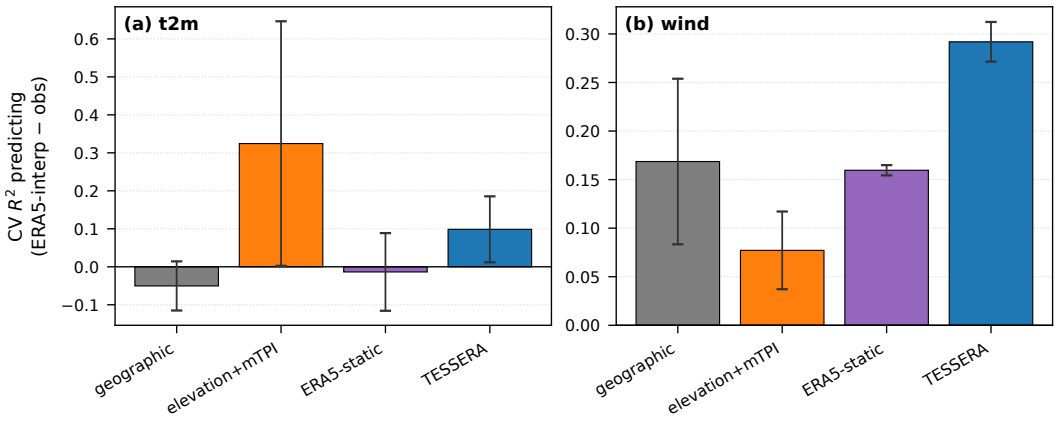


Figure 5: Cross-validated R^2 of random-forest regressors predicting the persistent ERA5-interpolation residual from each descriptor space in Europe and the United States ($n \geq 100$). Left: 2 m temperature. Right: 10 m wind speed. Positive values indicate transferable residual structure at held-out stations; values near or below zero indicate little improvement over predicting the training-fold mean residual.

3.3 Residual structure: what organises the sub-grid correction

To further establish whether the fine-scale structure introduced by the TESSERA in Figure 3 reflects surface signal rather than high-frequency noise, we ask, *independently* of any trained ConvCNP, which static descriptor space makes the ERA5-interpolation residual predictable at unseen stations. Specifically, we fit the same nonlinear regression probe in geographical, topographic, ERA5-static, and TESSERA descriptor spaces, and compare their cross-validated predictive performance. The residual construction, regression procedure, and additional diagnostics are detailed in Appendix E.

Figure 5 shows a clear variable-dependent split. For 2 m temperature, topography is the strongest organiser of the persistent residual in Europe, reaching a cross-validated $R^2 = 0.65$, compared with 0.19 for TESSERA and approximately zero for geography. The sharp separation indicates that the persistent temperature residual is organised much more directly by explicit topography than by the embedding, supporting the map analysis: elevation captures most of the dominant temperature correction, while TESSERA still contains additional, but weaker, transferable information. In the United States, no descriptor attains materially positive performance for temperature, consistent with the weaker topographic mismatch between stations and the ERA5 orography in that network.

For 10 m wind speed, the ordering reverses. The TESSERA descriptor yields the highest cross-validated R^2 in both Europe (0.27) and the United States (0.31), outperforming geography, topography, and the ERA5-static fields. This indicates that the persistent wind correction is organised by land-surface properties not represented effectively by elevation alone. These results provide a model-independent counterpart to Figures 3 and 4: topography largely organises temperature corrections, whereas TESSERA exposes additional surface structure that more effectively organises wind corrections.

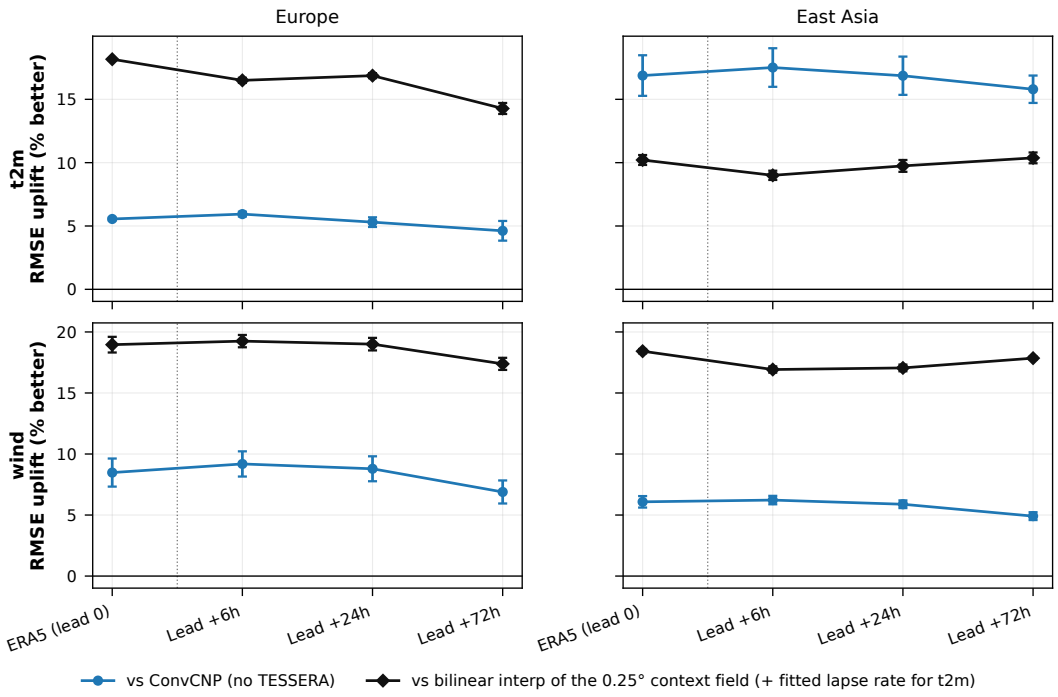


Figure 6: Uplift of ConvCNP with TESSERA in RMSE (% better; positive favours TESSERA) over two references, by variable (rows) and region (columns). *Blue:* the no-TESSERA ConvCNP baseline. *Black:* bilinear interpolation of the 0.25° weather field supplied to the model at that lead — real ERA5 at lead 0, the Aurora forecast at $+6/+24/+72$ h. For 2 m temperature, the interpolation additionally carries the fitted lapse-rate correction $T_{\text{interp}} - \Gamma \cdot \Delta\text{elevation}$ of §3.1. Error bars span the three training seeds.

3.4 Forecast-driven deployment: from reanalysis to Aurora

An important operational use case for downscaling is producing local predictions from medium-range weather forecasts. Here we replace the ERA5 context with forecasts from Aurora [7], a strong deterministic AI weather model, and use these as inputs to the downscaler. The forecast weather fields are generated at 0.25° and initialised from ERA5, storing the +6, +24, and +72 h leads from a single rollout per initialisation. Because Aurora does not forecast 6 h accumulated total precipitation, all input weather grids use 19 dynamic input channels, with total precipitation omitted at every lead, including lead 0.

Since the context weather grid is now a forecast rather than reanalysis, its errors grow and change with lead time, affecting both the expected downscaled value and its uncertainty. A downscaler trained only on ERA5 under a perfect-prognosis setup cannot adapt to these lead-dependent forecast errors and may become increasingly biased or under-dispersed at longer leads. We mitigate this by training jointly on all four context grids (ERA5 at lead 0 and the three Aurora leads — a model output statistics approach), and by passing the normalised horizon (lead/72) as an additional input channel. This allows the downscaler to adjust the full predictive distribution, including both its mean and spread, as a function of lead time. The resulting within-1σ coverage remains near nominal through +72 h.

The Tessera uplift is preserved across leads. We again find that Tessera provides substantial advantages when downscaling medium-range forecasts, with its benefit persisting across the forecast horizon (Figure 6). Measured in RMSE, it retains lower error than the no-Tessera ConvCNP baseline at every lead in both regions: for temperature, the uplift is ∼6% in Europe and ∼17% in East Asia at lead 0, and decays only mildly, to ∼5% and ∼16% by +72 h; for wind it is ∼6–9% at lead 0 and eases to ∼5–7% by +72 h. The operational reading is that the land-surface signal the embedding carries is not eroded by forecast error in the weather grid. The relative degradation of temperature and wind-speed skill with forecast lead is examined separately in Appendix F.

3.5 Sample efficiency under simulated station deployment

Continuing to probe Tessera's benefits in operational settings, we ask whether, as in other downstream tasks where EO embeddings dramatically improve sample efficiency, it also accelerates extrapolation to previously unobserved regions. Specifically, when a new station network is deployed, how long must it collect observations before its downscaling estimates become reliable, and how well are its locations served *before* any local history is available? We investigate this through a simulated deployment of the Norwegian station network, making the time and network-size axes explicit in an operationally realistic scenario.

We simulate the gradual deployment of a new Norwegian station network. At the start of the experiment, no Norwegian probe stations are available for the model to train on, and are then activated over a three-year period, while the remaining European stations stay available throughout. This lets us measure how performance changes as local observational coverage accumulates.

Let $\mathcal{P}$ denote the Norwegian probe stations and $\mathcal{T}_{\overline{\text{NO}}}$ the fixed non-Norwegian European training stations, active throughout. Fix a deployment start T_0 and window $\Delta = 36$ months. A uniform random permutation $\pi : \mathcal{P} \to \{1, \ldots, |\mathcal{P}|\}$ assigns each probe s an activation date on a constant cadence,

$$a_s \;=\; T_0 + \frac{\pi(s)}{|\mathcal{P}|}\,\Delta. \tag{6}$$

At a simulated horizon h after T_0 we retrain on all data up to $T_0 + h$, each probe contributing only its post-activation observations, so the training station set and the observations it exposes are

$$\mathcal{T}(h) \;=\; \mathcal{T}_{\overline{\text{NO}}} \cup \big\{\, s \in \mathcal{P} : a_s \le T_0 + h \,\big\}, \qquad \big\{(s,t) : a_s \le t \le T_0 + h\big\}. \tag{7}$$

Note that the horizon h expands the training window for *every* station. The non-Norwegian stations contribute observations spanning $[\,$2010-01-01, $T_0 + h\,]$, where $T_0 =$ 2015-01-01, carrying five years of history at cold start, eight when the deployment completes at $h = 3$ years, and eleven at $h = 6$ years. Evaluation is fixed throughout, on the held-out year 2022. Consequently, the curves in Figure 7 reflect all station observations available at each training horizon and do not isolate the effect of adding newly deployed probe stations.

The embedding improves wind speed downscaling from cold start. For wind speed, Figure 7 shows that ConvCNP with Tessera holds a clear advantage throughout the deployment across all station partitions. On the permanently held-out Norwegian stations, the gap is widest at and immediately after cold start: Tessera reduces MAE by $0.25\,\mathrm{m\,s^{-1}}$ before any probe stations are deployed and by $0.30\,\mathrm{m\,s^{-1}}$ after the first month, corresponding to ∼15–17% of the no-Tessera baseline. The same holds on the probes not yet deployed, which the figure omits (0.30 and $0.36\,\mathrm{m\,s^{-1}}$,

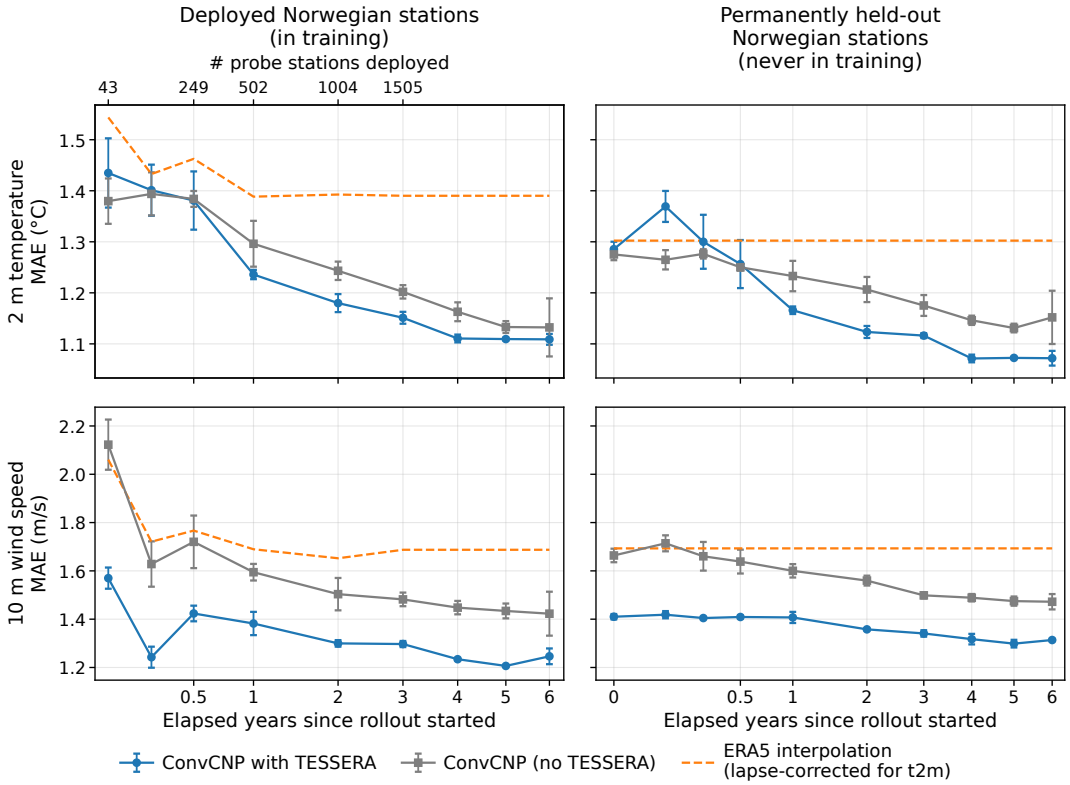


Figure 7: Simulated Norway deployment. MAE on a held-out test year as the Norwegian probe network is progressively deployed into training. Rows: 2 m temperature (top), 10 m wind speed (bottom). Columns split the Norwegian stations by their relation to the training set: *(left)* probes already deployed stations at that horizon; *(right)* the permanently held-out Norwegian test stations. Curves compare ConvCNP with TESSERA (blue) with the no-TESSERA baseline (grey); points are the mean over three seeds and error bars ±1 standard deviation. The dashed line is ERA5 bilinear interpolation — lapse-rate corrected for 2 m temperature — aggregated at each horizon over exactly the stations that panel scores. It is therefore flat on the right, whose station set is fixed, and varies on the left only because the deployed population changes as probes come online. Bottom axis: calendar years elapsed since T_0, on a square-root scale. Top axis (left column): the number of Norwegian probes deployed into training.

18–21%). The advantage subsequently narrows but does not close, remaining 0.16 m s^{-1} (11%) at full deployment. This suggests that increasing station count alone does not fully compensate for the absence of a descriptor exposing the surface structure that governs the local wind-speed correction.

Operationally, against ERA5 interpolation evaluated on the same Norwegian held-out stations, ConvCNP with TESSERA is 16.7% better before a single Norwegian probe is deployed (1.41 against 1.69 m s^{-1}), whereas the no-TESSERA variant merely starts level with it (1.66 m s^{-1}) and needs roughly a year of local deployment to open a clear margin. Six years and 1505 deployed stations later, it reaches 1.47 m s^{-1} — still above the error TESSERA attains with no local stations at all. The same ordering holds under RMSE (2.19 after six years, against 2.15 at cold start).

The temperature benefit is smaller and emerges later. For 2 m temperature, TESSERA brings no advantage at the earliest horizons. On the permanently held-out Norwegian stations, the

two variants are level at cold start, with TESSERA 0.10 °C worse after the first month. A consistent advantage appears only from the first year onwards, reaching an MAE reduction of ∼0.08 °C (7%) after two years and holding between 0.06 and 0.08 °C thereafter. This weaker cold-start effect is consistent with §3.3's conclusion that temperature's downscaling correction is driven primarily by explicit topography: the existing European network already provides substantial coverage of Norwegian elevation conditions, even prior to deploying any probe stations. This does not conflict with §3.1 since TESSERA is most valuable when station scarcity also implies poor coverage of the relevant land-surface conditions, whereas the dense European network contains many stations as close elevation analogues to learn the required correction from.

The same conclusions hold under RMSE and CRPS. Note that, at the earliest horizons, the results for the already deployed Norwegian stations in Figure 7's left column rest on 18 and 12 stations for temperature and wind, respectively. Their early non-monotonicity may therefore reflect high sampling variability and possible overfitting, both of which diminish as more stations are deployed.

To investigate why the wind speed advantage appears before local observations are available, Appendix G compares held-out Norwegian stations with the evolving training set in geography, topography, ERA5-static, and TESSERA descriptor spaces. It shows that TESSERA gives almost all held-out Norwegian stations a nearby training analogue from cold start while retaining surface information that distinguishes Norway from the broader European domain. This supports the interpretation that the embedding enables corrections to transfer from previously observed surface analogues; the complete reachability and separability analyses are deferred to the appendix.

4 Discussion and conclusion

We show that a learned Earth observation representation can serve as a general-purpose sub-grid descriptor for probabilistic weather downscaling. Across five climatically diverse regions, conditioning a ConvCNP's decoder on a TESSERA patch embedding (§2.2, §2.3) improves both deterministic and probabilistic skill for 2 m temperature and 10 m wind speed at held-out weather stations, reducing CRPS by 11.5% and 6.2% across regions, respectively (§3.1). These improvements persist when downscaling medium-range forecasts from the Aurora AI model (§3.4). Controls confirm that the gain derives from station-specific surface information rather than added model capacity (Appendix B), and is not reproduced by a substantially richer hand-crafted descriptor (Appendix C).

For 2 m temperature, elevation + mTPI most strongly organises the ERA5-interpolation residual (§3.3), and adding TESSERA produces little additional texture in densely downscaled fields (§3.2). The embedding's principal benefit for temperature is therefore not to replace topography, but to provide a broader surface prior when the available station observations are insufficient to learn the correction reliably from topographic analogues alone. For 10 m wind speed, the evidence differs consistently across the residual analysis (§3.3), dense maps (§3.2), and deployment experiment (§3.5), showing the correction is organised most effectively in TESSERA space. This suggests that topography provides a less data-efficient representation of the local correction required for near-surface wind speed, as locations of similar elevation can still differ in land cover, exposure, surface roughness, and built environment.

This distinction reinforces the operational value of the embedding. In the simulated Norwegian deployment (§3.5), the wind-speed advantage is largest before any Norwegian observations are available and persists after the local network has been fully deployed. The descriptor-space analysis (Ap-

pendix G) suggests why: TESSERA places almost all unseen Norwegian stations within the support of existing European surface analogues from the outset, while retaining the features that distinguish Norway from the remainder of the training domain. The embedding uplift also survives forecast degradation through +72 h (§3.4), indicating that the local surface correction and the evolving error of the atmospheric context behave independently over the forecast horizons considered here.

Several limitations bound these conclusions. First, the experiments cover only instantaneous 2 m temperature and 10 m wind speed. The relevance of learned surface descriptors is likely to differ for precipitation, humidity, radiation, and temporally aggregated quantities, and should be established separately. The models are also trained independently by variable and region, not exploiting physical dependence between variables or testing whether one globally trained downscaler can transfer across climates without regional retraining. Additionally, GHCNh's station network is not distributed randomly over the surface, so performance at uninstrumented environments that are poorly represented even in the TESSERA descriptor space remains uncertain.

Second, TESSERA is treated as a frozen, static surface representation. Each $\mathbf{z}_T$ compresses a fixed 640 m patch using a separately trained VAE, and neither the patch scale nor the 16-dimensional bottleneck is optimised jointly for the downscaling target. This deliberate decoupling limits overfitting, but it does not establish that the chosen neighbourhood scale or patch encoding objective is optimal. Other EO foundation embeddings may also expose complementary aspects of terrain, land cover, soil properties, and seasonally varying surface state. Time-indexed EO representations could also capture changes in vegetation, snow cover, water extent, or urban development that a static descriptor cannot represent, provided that genuine time-varying surface information can be separated from leakage or mismatched observation dates.

Third, the model's predictive distribution provides calibrated marginal distributions at arbitrary points, but it does not represent residual dependence between nearby stations. This is a property of the predictive construction rather than of the surface descriptor, with the latter being agnostic to how dependence between targets is modelled. Latent-variable neural processes can recover coherent joint structure at the cost of approximate inference [16], while a trained ConvCNP can be deployed autoregressively to produce dependent joint predictions without architectural changes [8]. Alternatively, diffusion-based neural processes offer richer multivariate structure [9]. Applications requiring joint uncertainty (e.g., aggregate wind-power risk across multiple sites) are therefore a natural extension of the TESSERA-conditioned downscaler.

This paper's central contribution is therefore not only an improvement to one downscaling architecture, but evidence for a broader role of Earth-observation foundation models in weather prediction. To our knowledge, this is the first study to demonstrate that a frozen EO foundation representation can serve as an effective point-specific surface descriptor for probabilistic downscaling at previously unseen locations. In the forecasting setting, it also demonstrates how two independently trained foundation models can be composed: Aurora represents the evolving atmospheric state, while TESSERA represents the local surface through which that state is expressed. Their combination yields more skilful local predictions than either the coarse forecast or a topography-conditioned downscaler alone, establishing EO embeddings as a practical interface between global weather models and site-specific forecasting.

References

[1] Hashem Akbari, Melvin Pomerantz, and Haider Taha. Cool surfaces and shade trees to reduce energy use and improve air quality in urban areas. *Solar Energy*, 70(3):295–310, 2001. doi: 10.1016/S0038-092X(00)00089-X.

[2] Mihai Alexe, Eulalie Boucher, Peter Lean, Ewan Pinnington, Patrick Laloyaux, Anthony McNally, Simon Lang, Matthew Chantry, Chris Burrows, Marcin Chrust, Florian Pinault, Ethel Villeneuve, Niels Bormann, and Sean Healy. GraphDOP: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations. *arXiv preprint arXiv:2412.15687*, 2024. doi: 10.48550/arXiv.2412.15687.

[3] Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, 2025. doi: 10.1038/s41586-025-08897-0.

[4] Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023. doi: 10.48550/arXiv.2306.06079.

[5] Åsmund Bakketun, Håvard Homleid Haugen, Jostein Blyverket, Thomas Nils Nipen, and Malte Müller. Enhancing a high resolution data-driven weather prediction model with surface descriptors. *arXiv preprint arXiv:2607.02824*, 2026. doi: 10.48550/arXiv.2607.02824.

[6] Jorge Baño-Medina, Rodrigo Manzanas, and José Manuel Gutiérrez. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124, 2020. doi: 10.5194/gmd-13-2109-2020.

[7] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

[8] Wessel P. Bruinsma, Stratis Markou, James Requeima, Andrew Y. K. Foong, Tom R. Andersson, Anna Vaughan, Anthony Buonomo, J. Scott Hosking, and Richard E. Turner. Autoregressive conditional neural processes. In *11th International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2303.14468.

[9] Vincent Dutordoir, Alan Saul, Zoubin Ghahramani, and Fergus Simpson. Neural diffusion processes. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 8990–9012. PMLR, 2023. URL `https://proceedings.mlr.press/v202/dutordoir23a.html`.

[10] Emanuel Dutra, Joaquín Muñoz-Sabater, Souhail Boussetta, Takuya Komori, Shoji Hirahara, and Gianpaolo Balsamo. Environmental lapse rate for high-resolution land surface downscaling: An application to ERA5. *Earth and Space Science*, 7(5):e2019EA000984, 2020. doi: 10.1029/2019EA000984.

[11] European Space Agency and European Union. Copernicus Digital Elevation Model: GLO-30, 2021. URL https://doi.org/10.5270/ESA-c5d3d65. GLO-30 global digital surface model.

[12] Tom G. Farr, Paul A. Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, David Seal, Scott Shaffer, Joanne Shimada, Jeffrey Umland, Marian Werner, Michael Oskin, Douglas Burbank, and Douglas Alsdorf. The shuttle radar topography mission. *Reviews of Geophysics*, 45(2):RG2004, 2007. doi: 10.1029/2005RG000183.

[13] Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaius, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. TESSERA: Temporal embeddings of surface spectra for earth representation and analysis. pages 34818–34831, 2026.

[14] Zhengpeng Feng, Sadiq Jaffer, Ira Shokar, Jovana Knezevic, Mark Elvers, Clement Atzberger, Robin Young, Aneesh Naik, Niall Robinson, Andrew Blake, David Coomes, Anil Madhavapeddy, and Srinivasan Keshav. TESSERA v2: Scaling pixel-wise earth foundation models. *arXiv preprint arXiv:2607.03949*, 2026. doi: 10.48550/arXiv.2607.03949.

[15] Christopher A. T. Ferro, David S. Richardson, and Andreas P. Weigel. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15 (1):19–24, 2008. doi: 10.1002/met.45.

[16] Andrew Y. K. Foong, Wessel P. Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard E. Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 8284–8295, 2020.

[17] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional neural processes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 2018. URL https://proceedings.mlr.press/v80/garnelo18a.html.

[18] 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.

[19] Tilmann Gneiting, Kristin Larson, Kenneth Westrick, Marc G. Genton, and Eric Aldrich. Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space–time method. *Journal of the American Statistical Association*, 101(475):968–979, 2006. doi: 10.1198/016214506000000456.

[20] Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner. Convolutional conditional neural processes. In *8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020. URL https://openreview.net/forum?id=Skey4eBYPS.

[21] Torbjörn Gustavsson, Maria Karlsson, Jörgen Bogren, and Sven Lindqvist. Development of temperature patterns during clear nights. *Journal of Applied Meteorology*, 37(6):559–571, 1998.

[22] J. M. Gutiérrez, D. Maraun, M. Widmann, et al. An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *International Journal of Climatology*, 39(9):3750–3785, 2019. doi: 10.1002/joc.5462.

[23] Lucy Harris, Andrew T. T. McRae, Matthew Chantry, Peter D. Dueben, and Tim N. Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022. doi: 10.1029/2022MS003120.

[24] Hans Hersbach, Bill Bell, Paul Berrisford, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

[25] Kevin Höhlein, Michael Kern, Timothy Hewson, and Rüdiger Westermann. A comparative study of convolutional neural network models for wind field downscaling. *Meteorological Applications*, 27(6):e1961, 2020. doi: 10.1002/met.1961.

[26] Yuma Ichikawa and Koji Hukushima. Learning dynamics in linear VAE: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with KL annealing. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pages 1936–1944. PMLR, 2024. URL `https://proceedings.mlr.press/v238/ichikawa24a.html`.

[27] Yubin Jin, Zhenzhong Zeng, Yuntian Chen, Yingzuo Qin, Hao Xu, and Dongxiao Zhang. Innovative short-term weather forecasting system combining data-driven and dynamic downscaling approaches. *Artificial Intelligence for the Earth Systems*, 4(3), 2025. doi: 10.1175/AIES-D-24-0125.1.

[28] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. doi: 10.48550/arXiv.1312.6114.

[29] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the Earth. *Nature Ecology & Evolution*, 7(11):1778–1789, 2023. doi: 10.1038/s41559-023-02206-6.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

[31] Mapzen. Terrain tiles. Registry of Open Data on AWS, 2017. URL `https://registry.opendata.aws/terrain-tiles/`. Accessed 5 August 2026.

[32] Douglas Maraun and Martin Widmann. *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, Cambridge, 2018. doi: 10.1017/9781107588783.

[33] Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6:124, 2025. doi: 10.1038/s43247-025-02042-5.

[34] Peter J. Marinescu, Susan C. van den Heever, Leah D. Grant, Jennie Bukowski, and Itinderjot Singh. How much convective environment subgrid spatial variability is missing within atmospheric reanalysis data sets? *Geophysical Research Letters*, 51(24):e2024GL111856, 2024. doi: 10.1029/2024GL111856.

[35] Matthew J. Menne, Simon Noone, Nancy W. Casey, Robert H. Dunn, Shelley McNeill, Diana Kantor, Peter W. Thorne, Karen Orcutt, Sam Cunningham, and Nicholas Risavi. Global historical climatology network–hourly (GHCNh). NOAA National Centers for Environmental Information, Version 1, 2023. Accessed 28 July 2026.

[36] Ophelia Miralles, Daniel Steinfeld, Olivia Martius, and Anthony C. Davison. Downscaling of historical wind fields over Switzerland using generative adversarial networks. *Artificial Intelligence for the Earth Systems*, 1(4):e220018, 2022. doi: 10.1175/AIES-D-22-0018.1.

[37] T. R. Oke. The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, 108(455):1–24, 1982. doi: 10.1002/qj.49710845502.

[38] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318. PMLR, 2013. URL `https://proceedings.mlr.press/v28/pascanu13.html`.

[39] 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.

[40] Neelesh Rampal, Sanaa Hobeichi, Peter B. Gibson, Jorge Baño-Medina, Gab Abramowitz, Tom Beucler, Jose González-Abad, William Chapman, Paula Harder, and José Manuel Gutiérrez. Enhancing regional climate downscaling through advances in machine learning. *Artificial Intelligence for the Earth Systems*, 3(2):e230066, 2024. doi: 10.1175/AIES-D-23-0066.1.

[41] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

[42] Peter Sheridan, Samantha Smith, Andrew Brown, and Simon Vosper. A simple height-based correction for temperature downscaling in complex terrain. *Meteorological Applications*, 17(3): 329–339, 2010. doi: 10.1002/met.177.

[43] David M. Theobald, Dylan Harrison-Atlas, William B. Monahan, and Christine M. Albano. Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. *PLOS ONE*, 10(12):e0143619, 2015. doi: 10.1371/journal.pone.0143619.

[44] Thordis L. Thorarinsdottir and Tilmann Gneiting. Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):371–388, 2010. doi: 10.1111/j.1467-985X.2009.00616.x.

[45] Anna Vaughan, Will Tebbutt, J. Scott Hosking, and Richard E. Turner. Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, 15(1): 251–268, 2022. doi: 10.5194/gmd-15-251-2022.

[46] Wensong Weng, Peter A. Taylor, and John L. Walmsley. Guidelines for airflow over complex terrain: Model developments. *Journal of Wind Engineering and Industrial Aerodynamics*, 86 (2–3):169–186, 2000. doi: 10.1016/S0167-6105(00)00009-X.

[47] Adam Winstral, Tobias Jonas, and Nora Helbig. Statistical downscaling of gridded wind speed data using local topography. *Journal of Hydrometeorology*, 18(2):335–348, 2017. doi: 10.1175/JHM-D-16-0054.1.

[48] Qidong Yang, Jonathan Giezendanner, Daniel Salles Civitarese, Johannes Jakubik, Eric Schmitt, Anirban Chandra, Jeremy Vila, Detlef Hohl, Chris Hill, Campbell Watson, and Sherrie Wang. Multi-modal graph neural networks for localized off-grid weather forecasting. *arXiv preprint arXiv:2410.12938*, 2024. doi: 10.48550/arXiv.2410.12938.

[49] Kei Yoshimura and Masao Kanamitsu. Dynamical global downscaling of global reanalysis. *Monthly Weather Review*, 136(8):2983–2998, 2008. doi: 10.1175/2008MWR2281.1.

[50] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Myroslava Lesiv, Martin Herold, Nandin-Erdene Tsendbazar, Panpan Xu, Fabrizio Ramoino, and Olivier Arino. ESA WorldCover 10 m 2021 v200, 2022. URL `https://doi.org/10.5281/zenodo.7254221`.

A Implementation details

This appendix specifies the coarse predictor set, the construction of the station set, the patch encoder, the downscaler architecture, and the training configuration. Where not otherwise stated, the ConvCNP construction follows Vaughan et al. [45].

A.1 Coarse predictor fields and context grid

We take ERA5 [24] from the analysis-ready archive distributed with WeatherBench2 [41], on the 0.25° grid, at the four synoptic hours 00, 06, 12 and 18 UTC. Table 2 lists the fields. The predictor set follows Vaughan et al. [45] in its choice of upper-air variables and pressure levels (500, 700 and 850 hPa), and differs in using instantaneous 2 m temperature rather than daily maximum, and in adding 6-hourly accumulated precipitation. The static surface fields are the full set of ERA5 time-invariant fields, all 13 of which are supplied to the no-Tessera baseline, as per §2.1.

In addition to the fields of Table 2, the context tensor Z carries two coordinate channels holding the latitude and longitude of each grid cell, and four temporal channels, cos and sin of day-of-year and of hour-of-day. The CNN encoder sees 39 channels for the no-Tessera baseline and 26 for ConvCNP with Tessera, with the forecast-driven models of §3.4 carrying one further channel holding the normalised lead time. Dynamic, static and coordinate channels are z-scored per region using training-split statistics, and the temporal channels are left unscaled since bounded in $[-1, 1]$.

Table 2: Coarse predictor fields supplied to the convolutional encoder. Dynamic fields vary per snapshot; static fields are time-invariant and are used by the no-TESSERA baseline only. ERA5 short names are given in parentheses.

Group	Field	Levels
Surface (dynamic)	2 m temperature (`t2m`)	surface
	10 m eastward wind (`u10`)	surface
	10 m northward wind (`v10`)	surface
	Mean sea-level pressure (`msl`)	surface
	Total precipitation, 6 h accumulation (`tp`)	surface
Upper air (dynamic)	Temperature (`t`)	500, 700, 850 hPa
	Eastward wind (`u`)	500, 700, 850 hPa
	Northward wind (`v`)	500, 700, 850 hPa
	Specific humidity (`q`)	500, 700, 850 hPa
	Geopotential (`z`)	500, 700, 850 hPa
Static surface	Surface geopotential (`z`)	surface
	Land–sea mask (`lsm`)	surface
	Soil type (`slt`)	surface
	High-vegetation cover (`cvh`)	surface
	Low-vegetation cover (`cvl`)	surface
	Type of high vegetation (`tvh`)	surface
	Type of low vegetation (`tvl`)	surface
	Angle of sub-grid orography (`anor`)	surface
	Anisotropy of sub-grid orography (`isor`)	surface
	Slope of sub-grid orography (`slor`)	surface
	Std. dev. of filtered sub-grid orography (`sdfor`)	surface
	Lake cover (`cl`)	surface
	Lake total depth (`dl`)	surface

Each region is a fixed latitude–longitude box (Table 3) to which the ERA5 grid is cropped, so the encoder sees a regional rather than a global field, as in Vaughan et al. [45]. The boxes do not overlap, and stations outside all of them are discarded, so every station belongs to exactly one region. These box areas are the denominators of the station densities ρ in Table 1.

A.2 Station set and descriptor sources

Because every model in the paper must be scored on an identical set of locations for the rows of Table 1 to be comparable, the station set is fixed once by the intersection of three requirements and then reused everywhere.

A station is retained if (i) its GHCNh metadata elevation is finite and lies in $(-900, 8848]$ m, which rejects the sentinel values GHCNh uses for missing elevation and a small number of physically impossible entries; (ii) its TESSERA patch can be filled to at least 50 % coverage from the released tiles; and (iii) that patch is not all-zero, non-finite, or a gross-magnitude outlier.

Station elevation is taken from the GHCNh station metadata rather than from a DEM, and Δelevation is formed against ERA5 orography derived from the static surface geopotential as z/g with $g = 9.81\,\mathrm{m\,s^{-2}}$ and linearly interpolated to the station. mTPI is sampled from the ALOS Global mTPI product [43](`CSP/ERGo/1_0/Global/ALOS_mTPI`, band `AVE`, $\approx$270 m resolution, metres). All three features are expressed in metres and divided by 1,000 before entering the decoder,

Table 3: Region bounding boxes. Grid shape is the number of 0.25° ERA5 cells (latitude × longitude) after cropping.

Region	Latitude	Longitude	Grid shape
Europe	35°–75°N	24°W–40°E	161 × 257
United States	24°–50°N	125°–66°W	105 × 237
East Asia	20°–46°N	100°–146°E	105 × 185
Southern Africa	35°–15°S	15°–35°E	81 × 81
Australia	44°–10°S	112°–154°E	137 × 169

putting them on a comparable scale to the normalised weather features.

A.3 Patch compression VAE

The per-location TESSERA surface descriptor $z(x_\star)$ is produced by a convolutional variational autoencoder (VAE) [28] trained once, unsupervised, on the station patch corpus and thereafter frozen.

Architecture. The encoder is a stack of four stride-2 blocks, each a 3×3 convolution followed by batch normalisation, ReLU and 2-D dropout at rate 0.1, with channel widths $128 \to 256 \to 256 \to 512$. The four blocks take the patch from 64×64 down to 4×4, and the resulting $512 \times 4 \times 4$ feature map is flattened and projected by two parallel linear layers to the mean and log-variance of a diagonal Gaussian posterior over a 16-dimensional vector. Flattening rather than global-average-pooling the final feature map is deliberate: it lets the compressed surface descriptor retain where in the patch a feature occurs, rather than only which channels are active on average.

The decoder mirrors the encoder. A linear layer maps the lower-dimension surface descriptor back to $512 \times 4 \times 4$, followed by three transposed-convolution blocks ($512 \to 256 \to 256 \to 128$, each with batch normalisation and ReLU) and a final transposed convolution to 128 output channels with no normalisation or activation, since the reconstruction target is a z-scored embedding field and is not sign- or range-constrained.

Training corpus. Patches are z-scored per embedding channel using statistics estimated from a random sample of valid patches. The VAE is fitted globally, rather than separately within each region, so that a single encoder serves all regions. Its training corpus includes station-centred patches from all downstream splits, but the encoder sees neither weather observations nor split labels: it is trained only to reconstruct the globally available TESSERA surface field and is then frozen. This is consistent with the intended deployment setting, in which a surface descriptor can be computed at any query location, regardless of whether that location hosts a weather station.

Objective. Training minimises

$$\mathcal{L} \;=\; \mathcal{L}_{\text{recon}} \;+\; \lambda_\nabla \mathcal{L}_\nabla \;+\; \beta(t)\, D_{\text{KL}}(q(z \mid x) \,\|\, \mathcal{N}(0, I))\,, \tag{8}$$

where $\mathcal{L}_{\text{recon}}$ is the mean squared error between the reconstructed and input patch, and $\mathcal{L}_\nabla$ is the same error evaluated on first-order spatial finite differences of the patch in each direction, which discourages the decoder from settling on a spatially flat reconstruction that matches only the patch mean. We use $\lambda_\nabla = 0.5$.

The KL weight β is annealed linearly from zero to $\beta = 5 \times 10^{-4}$ over the first 5000 optimisation

steps (constant thereafter) to reduce the risk of posterior collapse [26]. We found the small terminal β necessary since, at KL weights of order one, the approximate posterior converged toward the prior and the latent ceased to carry useful station-specific information. The posterior log-variance is clamped to $[-10, 10]$ for numerical stability.

Optimisation. We optimise the VAE for 200 epochs using AdamW [30] with learning rate 10^{-3}, weight decay 10^{-5}, batch size 64, and gradient-norm clipping at 1.0 [38]. The learning rate follows a cosine schedule, and we retain the checkpoint with the lowest validation loss.

Inference and normalisation. The frozen encoder is applied once per station and the posterior mean (deterministic) is taken as the descriptor $z(x_\star)$. Latents are z-scored per dimension using statistics computed over all valid stations; the same statistics are reused for every region, split and experiment.

A.4 Downscaler architecture

Convolutional encoder. The context weather grid is processed by a residual CNN that preserves spatial resolution throughout: no striding, no pooling, and zero padding on every convolution. It consists of a 3×3 convolution projecting the input channels to 128 features, followed by three residual blocks — each two 3×3 convolutions at 128 channels with ReLU and an identity skip — and a final 1×1 convolution to 128 output channels. The output h therefore has 128 features at every grid point.

Grid-to-station evaluation. The kernel of eq. (1) is computed as a product of one-dimensional Gaussian weightings along latitude and longitude, then divided by the summed kernel density so the interpolant is insensitive to how many grid points fall within its support. The length-scale is parameterised as $\log \ell$ so it stays positive, and initialised at $\ell = 0.5°$ — roughly 55 km at mid-latitudes, or two ERA5 grid spacings. The initialisation is deliberately toward locality rather than region-wide smoothing, and leaves the optimiser free to widen the kernel if the data support it.

Decoder. The interpolated features are concatenated with the topographic descriptor and, for the TESSERA variants, the learned 16-dimensional surface descriptor, and passed through three Linear $\rightarrow$ ReLU layers of width 128. The decoder input width is therefore 147 for the TESSERA variants $(128 + 3 + 16)$ and 131 for the baseline. A single linear layer per target variable maps the final hidden state to that variable's distribution parameters.

Both variants have approximately 1.0×10^6 trainable parameters: 9.97×10^5 for the no-TESSERA baseline and 9.84×10^5 for ConvCNP with TESSERA. The baseline is marginally the larger of the two, because the 13 static channels it receives widen the first convolution. Capacity, therefore, does not favour the TESSERA model and the shuffled-surface control of Appendix B addresses the same question directly.

Marginal prediction and joint uncertainty The objective in eq. (3) trains the conditional predictive distribution at each target location $\mathbf{x}_\star$, which is equivalent to maximum-likelihood estimation under the standard conditional neural process factorisation [17],

$$p\big(y_\star^{(1:N)} \mid \mathbf{x}_\star{}^{(1:N)}, Z\big) \;=\; \prod_{n=1}^{N} p_{\boldsymbol{\theta}}\big(y_\star^{(n)} \mid \mathbf{x}_\star{}^{(n)}, Z\big). \tag{9}$$

However, this factorisation would be exact only if the context Z and the descriptors $\mathbf{e}$ and $\mathbf{z}_T$ rendered the targets $y_\star^{(1:N)}$ conditionally independent, which they do not (e.g., a synoptic feature displaced in the coarse field biases neighbouring stations coherently). It buys an exact, tractable likelihood, compatible with maximum likelihood training, at the cost of expressing residual spatial dependence with incoherent joint draws; constructions that recover coherent joint structure are discussed in §4.

Although residual stochastic dependence is not modelled, the same convolutional network and decoder, with weights shared across all target locations, produce $\boldsymbol{\theta} = (\mu\,, \sigma)$. Consequently, nearby locations are governed by the same learned spatial mapping and can have smoothly varying, physically similar marginal predictions.

A.5 Training configuration

One episode matches a single snapshot for a single region, with every valid station in that region as a target. Each episode forms one optimisation step, so the effective target-set size per step is the region's station count rather than a fixed number. Stations with a missing observation at that snapshot are masked out of the loss.

Optimisation uses Adam with a learning rate of 2.5×10^{-5} and weight decay 10^{-4}. The learning rate is warmed up linearly from 10^{-3} of its target value over the first 5 % of the nominal training steps and held constant thereafter; no decay schedule is applied, because runs typically early-stop well before a long decay would take effect. Gradients are clipped to a global norm of 1.0, and any step whose gradient contains a non-finite entry is skipped rather than applied.

Training runs for at most 100 epochs with early stopping on the validation negative log-likelihood, patience 10 epochs, and the checkpoint with the lowest validation loss is the one evaluated. Every configuration is trained from scratch with seeds 42, 123 and 456.

All models were trained on a single NVIDIA GH200 GPU per run, within a 24 h wall-clock limit that no run reached; in total, the experiments reported here — including patch-encoder pre-training, ablations, and Aurora forecast generation — consumed approximately 9,600 GPU-hours.

A.6 Forecast-driven context

For §3.4 we use the *pretrained* 0.25° Aurora checkpoint rather than the fine-tuned one: the latter is matched to IFS HRES analysis inputs whereas we initialise from ERA5, so keeping the pretrained model makes the experiment a single distribution shift, from reanalysis to forecast. Only the dynamic fields of Table 2 are replaced, except total precipitation, which Aurora does not produce and which is dropped from all four context settings, including lead 0. The static, coordinate, and temporal channels are unchanged.

Forecast frames are cropped to the boxes of Table 3 on the same grid indices as the ERA5 fields, and the crops are verified to align cell-for-cell with the reanalysis grid, so that a lead-0 and a lead-72 h context differ only in their field values.

B Shuffled-descriptor control and simpler patch summaries

The comparisons of §3.1 leave open two questions about *why* the TESSERA surface descriptor helps. First, the gain could in principle arise from the mechanics of adding a 16-dimensional input — extra parameters in the decoder's first layer, or a regularising effect of a high-entropy input — rather than from the surface information $z(x_\star)$ carries about each station. Second, assuming it carries real signal, the pre-trained VAE encoder of §2.2 is only one way to compress the patch, and a far simpler summary might suffice. We address the first question with a shuffled-descriptor control and the second with a hand-crafted summary-statistics descriptor computed over the same patches. Table 4 compares both with the correctly paired TESSERA surface descriptor and the no-TESSERA baseline across all five regions.

Table 4: Descriptor controls, evaluated on the same held-out stations and years as Table 1, in its identical configuration — the *none* and *VAE descriptor* rows reproduce that table's ConvCNP rows. Within each region the four models share architecture, parameter count, splits, and seeds; the three descriptor rows differ only in the content of the 16-dimensional per-station vector: the VAE-learned TESSERA surface descriptor, the same descriptors under a fixed random station permutation, and summary statistics computed over the same patch of embeddings. Values are means over 3 seeds; the *All regions* block is the unweighted mean over regions; lowest error per region × metric in **bold**.

		2 m temperature (°C)			10 m wind speed ($m\,s^{-1}$)		
Region	Additional surface descriptor	MAE	RMSE	CRPS	MAE	RMSE	CRPS
Europe	none (no TESSERA)	1.169	1.673	0.843	1.283	1.874	0.918
	VAE descriptor, shuffled	1.128	1.616	0.811	1.236	1.796	0.892
	patch summary statistics	1.102	1.583	0.792	1.212	1.767	0.867
	VAE descriptor	**1.098**	**1.577**	**0.789**	**1.191**	**1.739**	**0.861**
United States	none (no TESSERA)	1.413	2.070	1.029	1.446	1.960	1.038
	VAE descriptor, shuffled	1.339	2.000	0.976	1.406	1.905	1.011
	patch summary statistics	1.315	1.975	0.960	1.365	1.847	0.977
	VAE descriptor	**1.299**	**1.946**	**0.947**	**1.362**	**1.841**	**0.973**
East Asia	none (no TESSERA)	1.347	1.867	0.974	1.231	1.705	0.882
	VAE descriptor, shuffled	1.124	1.563	0.808	1.173	1.628	0.840
	patch summary statistics	**1.097**	1.533	**0.788**	**1.164**	**1.620**	**0.833**
	VAE descriptor	**1.097**	**1.530**	**0.788**	**1.164**	1.622	0.836
Southern Africa	none (no TESSERA)	1.998	2.749	1.447	1.231	1.646	0.885
	VAE descriptor, shuffled	1.879	2.637	1.367	1.239	1.659	0.889
	patch summary statistics	1.928	2.664	1.394	1.175	1.588	0.845
	VAE descriptor	**1.791**	**2.546**	**1.313**	**1.164**	**1.566**	**0.835**
Australia	none (no TESSERA)	1.573	2.164	1.149	1.422	1.812	1.025
	VAE descriptor, shuffled	1.358	1.980	1.006	1.327	1.727	0.964
	patch summary statistics	**1.313**	1.933	0.982	**1.265**	**1.639**	**0.906**
	VAE descriptor	1.323	**1.922**	**0.977**	1.292	1.690	0.945
All regions	none (no TESSERA)	1.500	2.105	1.088	1.323	1.799	0.949
	VAE descriptor, shuffled	1.365	1.959	0.994	1.276	1.743	0.919
	patch summary statistics	1.351	1.937	0.984	1.236	**1.692**	**0.886**
	VAE descriptor	**1.321**	**1.904**	**0.963**	**1.234**	**1.692**	0.890

Shuffled-descriptor control. We control for the station–descriptor pairing by reassigning each station the TESSERA surface descriptor of another station via a random permutation (fixed across variables and seeds). The architecture, parameter count, optimisation, and splits are all unchanged, and the empirical distribution of surface descriptors is preserved exactly; only the per-station content is now uninformative. Any benefit that survives shuffling cannot derive from station-specific surface information.

Table 4 shows that shuffling degrades CRPS in all ten region × variable settings, and, averaging over regions, it gives back a quarter of the CRPS gap to the no-TESSERA baseline for temperature and half of it for wind speed. We interpret the paired live-versus-shuffled contrast as the controlled estimate of the station-specific contribution. Moreover, because weather stations occupy a non-random subset of surface environments, shuffled surfaces are still drawn from the empirical distribution of station surfaces rather than from arbitrary noise. They may therefore retain a broad station-surface prior even after their location-specific alignment is removed.

Summary statistics instead of the learned patch encoder. The second control replaces the VAE with the simplest defensible summary of the same patch: the identical 64×64-pixel ($\sim$640 m) window of TESSERA embeddings the VAE encodes. We select the four embedding channels whose per-station spatial means vary most across *training* stations (held-out stations are excluded from the selection), and describe each selected channel by four statistics over the patch: mean, standard deviation, and the 10th and 90th percentiles. This produces a 16-dimensional vector, matched in dimensionality to the VAE's compressed latent. The two descriptors differ only in how the patch is summarised: a pre-trained non-linear encoder vs sixteen fixed-order statistics.

The summary statistics recover most of the gain. Averaged over regions, they retain 84% of the CRPS improvement provided by the VAE surface descriptor for temperature and effectively all of it for wind speed, and both descriptors outperform the no-TESSERA baseline in every region. The learned encoder retains a small but fairly consistent edge (lower CRPS in seven of ten region × variable settings, with the statistics ahead in East Asia and Australia for wind speed and tied in East Asia for temperature). We therefore retain it as the canonical descriptor, although the choice of patch-compression method is evidently not the load-bearing element of the approach.

Read together with Appendix C, this locates the signal precisely. An extended *hand-crafted physical* descriptor — seventeen land-cover, soil, canopy, and neighbourhood-terrain features — recovers roughly a quarter to a third of the embedding's benefit, whereas sixteen naive order statistics *of the embedding itself* recover 84% of it for temperature and effectively all of it for wind speed. What matters is access to the TESSERA representation, not the sophistication of either the feature list or the patch encoder. Combined with the shuffled control, these results support the reading that the embedding supplies transferable, station-specific surface information whose benefit is not explained by added model capacity and is not fully reproduced by an enumeration of physical surface properties, while leaving open the design of stronger patch encodings noted in §4.

C An extended hand-crafted surface descriptor

The comparison in §3.1 sets a 16-dimensional learned TESSERA surface descriptor against the three-feature topographic descriptor $\mathbf{e}$ of Vaughan et al. [45]. A natural objection is that the no-TESSERA baseline is disadvantaged by the absence of land-surface information: given a richer hand-crafted

Table 5: The 17 features of the extended hand-crafted descriptor, after Bakketun et al. [5]. The terrain group supplies neighbourhood topographic structure absent from **e**.

Group	Feature	Units	Source (radius)
Surface 320 m radius	`forest_frac`	fraction	ESA WorldCover v200 [50]
	`lowveg_frac`	fraction	
	`crop_frac`	fraction	
	`built_frac`	fraction	
	`bare_frac`	fraction	
	`snowice_frac`	fraction	
	`water_frac`	fraction	
	`tree_height`	m	ETH canopy height (2020) [29]
	`clay_frac`	$\mathrm{g\,kg^{-1}}$	SoilGrids 0–5 cm [39]
	`sand_frac`	$\mathrm{g\,kg^{-1}}$	
Terrain 6.25 km radius	`elev_mean`	m	Copernicus GLO-30 at 250 m [11]
	`elev_std`	m	
	`elev_min`	m	
	`elev_max`	m	
	`slope`	tan(slope)	
	`dz_dn`	$\mathrm{m\,m^{-1}}$	
	`dz_de`	$\mathrm{m\,m^{-1}}$	

vector, it might close the gap without any learned embedding. This appendix tests that directly.

Descriptor. We adopt the station-descriptor set of Bakketun et al. [5], which augments topography with explicit land-cover, vegetation, soil and neighbourhood-terrain features (Table 5). The 17 features divide into two groups. A *surface* group of ten features is aggregated over a 320 m radius from ESA WorldCover v200 [50], the ETH global canopy-height product [29], and SoilGrids at 0–5 cm depth [39]. A *terrain* group of seven features is aggregated over a 6.25 km radius from Copernicus GLO-30 [11], resampled to 250 m. The terrain group is worth emphasising: it supplies neighbourhood elevation statistics, slope and directional gradients that **e** does not contain, so the enriched baseline receives more topographic information as well as more land-surface information.

Protocol. The descriptor is concatenated onto the per-location input of eq. (2), so the enriched baseline uses $[\,\varphi_c(H, \mathbf{x}_\star)\,,\,\mathbf{e}\,,\,\mathbf{d}\,]$ with $\mathbf{d}$ the 17-feature vector. Splits, seeds, station sets, architecture and optimisation are identical to §3.1; only the decoder input width changes. We train the enriched variant of both the no-Tessera baseline and the Tessera model in all five regions, for both variables, over the same three seeds. Table 6 reports the results at three decimal places, since several of the differences of interest are smaller than the two-decimal precision of Table 1.

The enriched descriptor never matches the embedding. Across the ten region × variable settings and the three reported metrics, the enriched no-Tessera baseline does not reach ConvCNP with Tessera in any of the thirty comparisons. Averaged over regions, it reduces CRPS by 3.2% for 2 m temperature and 2.2% for 10 m wind speed, against Tessera's 11.5% and 6.2%, recovering roughly 28% and 36% of the embedding's gain, respectively. The recovered fraction varies widely by region, but is never complete, suggesting that the embedding supplies information that a substantially richer hand-crafted vector, including additional terrain structure, does not.

Enriching the descriptor is not uniformly beneficial. The extended descriptor improves the baseline in most settings but degrades it in 4 of the 30 region × variable × metric comparisons, con-

Table 6: Effect of the extended hand-crafted descriptor, evaluated on the same held-out stations and years as Table 1. Lowest error per region×metric in **bold**. Values are means over 3 seeds.

		2 m temperature (°C)			*10 m wind speed ($m\,s^{-1}$)*		
Region	Model	MAE	RMSE	CRPS	MAE	RMSE	CRPS
Europe	ConvCNP (no TESSERA)	1.169	1.673	0.843	1.283	1.874	0.918
	+ extended descriptor	1.152	1.647	0.829	1.244	1.813	0.893
	ConvCNP with TESSERA	1.098	1.577	0.789	1.191	1.739	0.861
	+ extended descriptor	**1.066**	**1.522**	**0.766**	**1.182**	**1.709**	**0.845**
United States	ConvCNP (no TESSERA)	1.413	2.070	1.029	1.446	1.960	1.038
	+ extended descriptor	1.381	2.023	1.007	1.416	1.910	1.013
	ConvCNP with TESSERA	1.299	1.946	0.947	1.362	1.841	0.973
	+ extended descriptor	**1.280**	**1.918**	**0.935**	**1.353**	**1.818**	**0.965**
East Asia	ConvCNP (no TESSERA)	1.347	1.867	0.974	1.231	1.705	0.882
	+ extended descriptor	1.207	1.669	0.870	1.233	1.697	0.879
	ConvCNP with TESSERA	**1.097**	**1.530**	**0.788**	1.164	1.622	0.836
	+ extended descriptor	1.101	**1.530**	0.790	**1.149**	**1.611**	**0.825**
Southern Africa	ConvCNP (no TESSERA)	1.998	2.749	1.447	1.231	1.646	0.885
	+ extended descriptor	1.980	2.725	1.438	1.250	1.671	0.896
	ConvCNP with TESSERA	1.791	2.546	1.313	**1.164**	**1.566**	**0.835**
	+ extended descriptor	**1.776**	**2.540**	**1.302**	1.173	1.570	0.842
Australia	ConvCNP (no TESSERA)	1.573	2.164	1.149	1.422	1.812	1.025
	+ extended descriptor	1.498	2.159	1.123	1.331	1.713	0.961
	ConvCNP with TESSERA	**1.323**	**1.922**	**0.977**	1.292	1.690	0.945
	+ extended descriptor	1.369	1.959	1.004	**1.279**	**1.660**	**0.922**
All regions	ConvCNP (no TESSERA)	1.500	2.105	1.088	1.323	1.799	0.949
	+ extended descriptor	1.444	2.044	1.053	1.295	1.761	0.928
	ConvCNP with TESSERA	1.321	1.904	0.963	1.234	1.692	0.890
	+ extended descriptor	**1.319**	**1.894**	**0.959**	**1.227**	**1.674**	**0.880**

centrated on Southern Africa wind speed. As noted in §3.1, results in the sparsest regions are noisier because they are estimated from relatively few held-out stations, so these isolated degradations should be interpreted cautiously. We used the three-feature **e** as the principal no-TESSERA baseline because it isolates explicit topography, allowing the contrasting roles of topographic and learned surface information for temperature and wind speed to be assessed cleanly in §3.2 and §3.3.

The two descriptors are only partly complementary. Adding the extended descriptor *on top of* TESSERA helps in the two data-rich regions but is worse than TESSERA alone in 8 of 30 comparisons, concentrated in East Asia, Southern Africa and Australia. Aggregated over all regions, the combination is essentially indistinguishable from TESSERA alone for temperature (1.319 vs 1.321 MAE) and marginally better for wind (1.227 vs 1.234). We therefore do not claim the two descriptors are complementary in general: where stations are plentiful, the additional features carry some independent signal, and where they are scarce, TESSERA's generalisation capabilities are preferred.

Relation to the residual analysis. Re-running the random-forest probe of §3.3 over these descriptor spaces shows that the extended vector is not information-poor, as per Figure 8. For wind speed, the extended vector rivals the embedding (CV R^2 0.29 for both), but only as a whole: its terrain and land-cover halves score 0.06 and 0.23 in isolation; for temperature, **e** alone dominates every other space (0.32, against 0.10 for TESSERA).

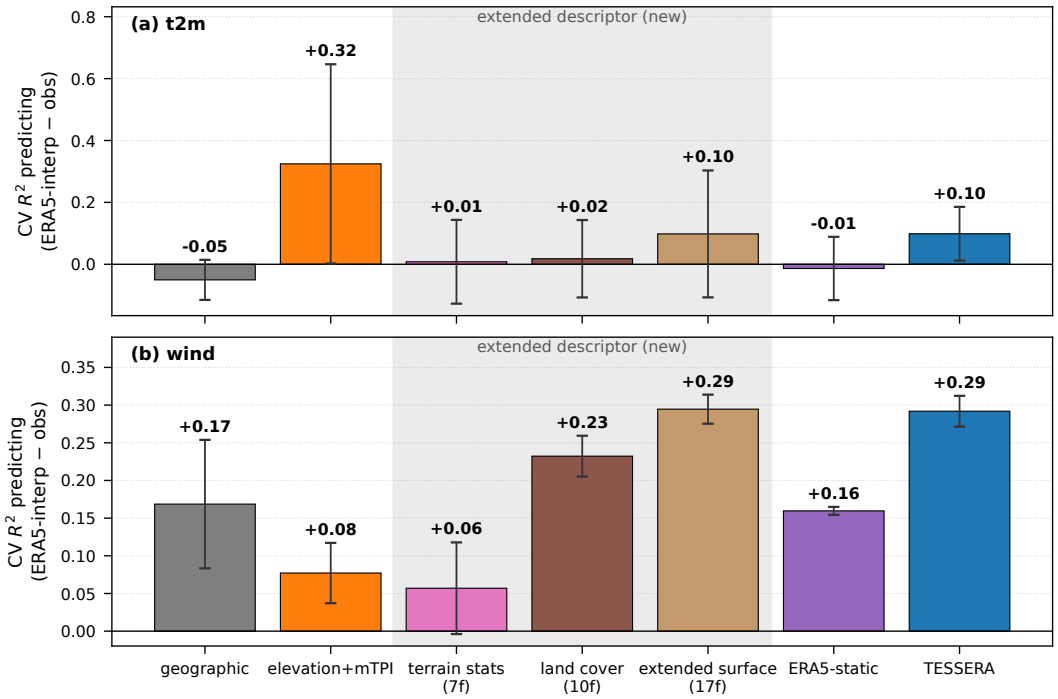


Figure 8: The probe of Figure 5, widened with the extended hand-crafted descriptor. Cross-validated R^2 of random-forest regressors fitted independently in each descriptor space on the ERA5-interpolation residual. The three shaded spaces come from the 17-feature descriptor of Table 5, entered both whole and as its two halves: *terrain stats* and *land cover*. Bars are the mean over the well-sampled regions (Europe and the United States, $n \geq 100$ stations). Left: 2 m temperature. Right: 10 m wind speed.

The probe and the skill table therefore disagree in both directions, and that is informative rather than contradictory: the probe scores a static, time-averaged residual, whereas Table 6 scores per-snapshot predictive distributions. The embedding's advantage is not that it encodes more persistent surface signal than an explicit feature list, but that the downscaler extracts more usable conditioning from it. The dissociation of §3.3 is unaffected — temperature residuals are organised by topography, wind residuals by land-surface character — and what Table 6 adds is that it is the learned representation of that character, not its enumeration, that converts into probabilistic skill.

D Quantifying and validating fine-scale map structure

This appendix provides the quantitative analysis underlying the dense-map interpretation in §3.2. We first measure how much fine-scale spatial variation each model resolves, and then test at observed station locations whether the changes induced by TESSERA are aligned with errors in the baseline prediction.

Texture statistic. We measure the structure each field carries below the coarse grid scale. Throughout, $\hat{y}(x)$ is the predictive mean at cell x (median for wind), averaged over the three training seeds. Therefore, the mapped predictive 0.05° field is **not** a draw from the joint predictive

$p\big(y_\star^{(1:N)} \mid x_\star^{(1:N)}, Z\big)$. The texture reported below is the deterministic model structure rather than sampling noise since draws from the factorised predictive of eq. (9) are spatially incoherent by construction.

Writing $\langle \hat{y} \rangle_\sigma(x)$ for a Gaussian low-pass of $\hat{y}$ with standard deviation $\sigma = 3$ map cells (0.15°), renormalised over valid cells so that the sea mask does not corrupt the average near coastlines, the fine-scale component is the high-pass residual

$$r(x) \;=\; \hat{y}(x) \;-\; \langle \hat{y} \rangle_\sigma(x), \tag{10}$$

and the field's texture is its spatial standard deviation over the $|X|$ mapped cells,

$$\mathcal{T} \;=\; \sqrt{\tfrac{1}{|X|} \sum_{x \in X} \big(r(x) - \bar{r}\big)^2}, \qquad \bar{r} = \tfrac{1}{|X|} \sum_{x \in X} r(x), \tag{11}$$

i.e. the amplitude of variation below the 0.15° smoothing scale that the downscaler resolves. We report the ratio $\mathcal{T}_{\textsc{Tessera}}/\mathcal{T}_{\text{base}}$ of the two models' textures, and confirm the maps are skilful by scoring both models at in-region stations on the mapped snapshots.

For 2 m temperature, the two models resolve nearly identical fine-scale amplitudes, with the ratio ≈ 1 in both regions (1.02 in Iberia and 1.08 in Norway; Figure 3a). This indicates that elevation already captures most of the *amplitude* of the resolved temperature sub-grid structure. However, this scalar comparison does not imply that the two models place that structure identically.

For 10 m wind speed, the baseline field is comparatively smooth, whereas Tessera increases the fine-scale amplitude by 3.28× in Iberia and 3.80× in Norway (Figure 3b), tracking land-cover and surface-roughness boundaries rather than elevation contours.

Station-level error alignment. On the mapped snapshots, Tessera lowers aggregate in-region station CRPS for both variables in both regions (Iberia wind $1.48 \to 1.27\,\mathrm{m\,s^{-1}}$, Norway temperature $1.55 \to 1.42\,^\circ\mathrm{C}$), with more even per-station win rates (∼57–65%). These scores show that the embedding-induced changes are beneficial on average, but not whether they constitute spatially directed corrections rather than arbitrary local perturbations.

Let $\hat{y}_b(x_s)$ and $\hat{y}_{\textsc{Tessera}}(x_s)$ be predictions at station s for the baseline and Tessera, respectively, and y_s the observation. We decompose the Tessera prediction into the baseline plus an increment, $\Delta_s \equiv \hat{y}_{\textsc{Tessera}}(x_s) - \hat{y}_b(x_s)$, so that

$$\hat{y}_{\textsc{Tessera}}(x_s) = \hat{y}_b(x_s) + \Delta_s, \tag{12}$$

where Δ_s is the local change in predictive mean induced by the embedding. With the signed baseline error $e_s = y_s - \hat{y}_b(x_s)$, the per-station skill gain is

$$g_s \;=\; |e_s| - |e_s - \Delta_s|, \qquad |g_s| \le |\Delta_s|, \tag{13}$$

with the bound following from the reverse triangle inequality. A larger increment thus does not necessarily imply a larger gain: it reduces the baseline error only when Δ_s has the same sign as e_s and does not overshoot the observation, that is, $g_s > 0 \iff 0 < \Delta_s/e_s < 2$. Whether the embedding-induced changes (i.e. Δ at each downscaled point) are skilful is therefore a question of their alignment with the baseline errors e_s.

Regressing the baseline error on the increment across in-region stations, $e_s = \beta\,\Delta_s + \eta_s$, Δ_s explains a fraction $R^2 = \mathrm{corr}(\Delta, e)^2$ of the baseline-error variance and is correctly signed at a hit rate

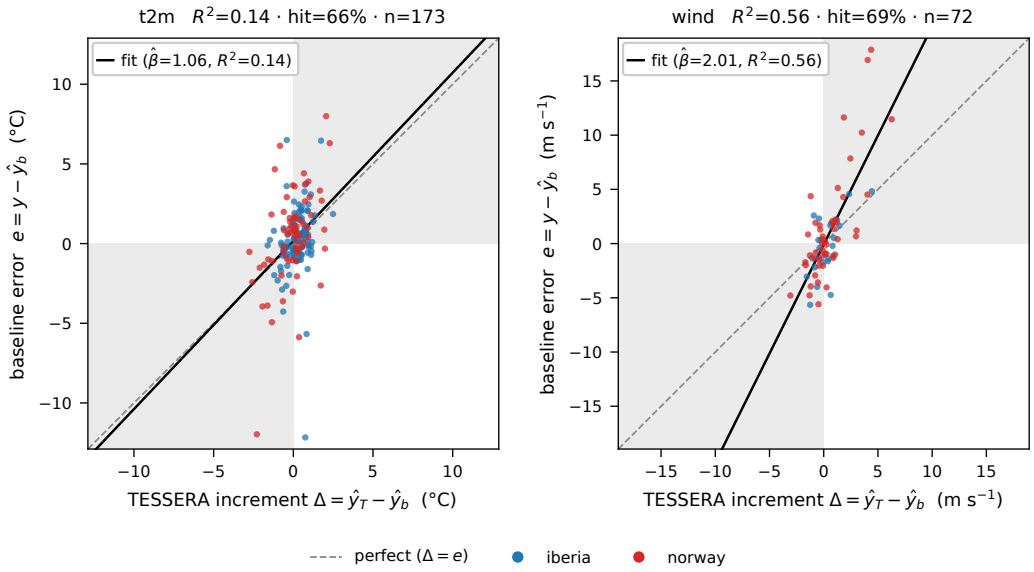


Figure 9: Per-station baseline error $e = y - \hat{y}_b$ against the TESSERA increment $\Delta = \hat{y}_{\text{TESSERA}} - \hat{y}_b$ at in-region test stations for the mapped snapshots (t2m left, wind right). Shaded quadrants denote corrections with the same sign as the baseline error, giving directional hit rates $H \approx 66$–69%. The dashed line denotes perfect correction, $\Delta = e$, and the solid line is the fitted regression $e = \hat{\beta}\Delta$. The slope measures average magnitude agreement, while R^2 measures how consistently the increment explains variation in the baseline error.

$H = \frac{1}{N}\sum_s \mathbb{1}[\operatorname{sign}\Delta_s = \operatorname{sign} e_s]$. The comparable hit rates ($H = 69\%$ wind, 66% temperature) show that the TESSERA increments are directionally informative for both variables. However, their effects differ: for wind, the increment explains $R^2 = 0.56$ ($n = 72$) of the baseline-error variance, a substantial missing spatial correction; for temperature, the lower $R^2 = 0.14$ ($n = 173$) indicates that elevation already captures most of the dominant sub-grid structure, while TESSERA provides smaller, more localised refinements that are useful but explain only a modest fraction of the remaining error.

Alignment strengthens with increment magnitude. The alignment strengthens monotonically with how much structure TESSERA adds. Stratifying every test observation across all regions by the increment magnitude $|\Delta_s|$ — the station-level analogue of the map texture — the directional hit rate rises from $\approx$51% in the lowest $|\Delta|$ decile to $\approx$76% in the highest, and the increment explains $R^2 \approx 0.3$ of the baseline-error variance within the top decile versus ≈ 0 within the bottom ($n > 10^6$, temperature and wind alike). Where the embedding barely moves the prediction, it is essentially undirected but also negligible; where it adds substantial structure, exemplified by the dense maps, it is reliably aimed at the baseline's error. The modest full-test R^2 is therefore a dilution by near-zero increments, not evidence against the mapped structure being skillful.

Random-perturbation control. We replace the observed increments $|\Delta_s|$ with independent Gaussian perturbations matched in amplitude, thereby removing their systematic sign alignment with the baseline errors. For the observed wind increment, the rank correlation is $\rho(|\Delta|, g) = +0.14$, i.e. larger adjustments yield larger error reductions; under the random control, it reverses to -0.27. Independent perturbations increase the error most where their magnitude is largest, whereas the observed increments reduce it because they are aligned in sign with the baseline errors. Therefore,

the added wind texture reflects skilful correction rather than merely larger local adjustments.

E Model-independent analysis of residual structure

This appendix provides the construction and validation details for the descriptor-space analysis summarised in §3.3. The goal is to compare how effectively different static representations organise the persistent ERA5-interpolation residual, independently of any trained downscaling model.

Residual target. We predict residuals rather than station values directly. Conceptually, a station observation can be written as

$$y_{s,t} = y_{s,t}^{\text{interp}} - r_s + \varepsilon_{s,t}, \tag{14}$$

where $y_{s,t}^{\text{interp}}$ represents the resolved, time-varying atmospheric state, r_s is a persistent location-specific interpolation bias, and $\varepsilon_{s,t}$ collects transient local effects and observation noise. Since the candidate descriptors are static properties of the station location, asking them to predict $y_{s,t}$ directly would primarily test their ability to explain station climatology while confounding the analysis with the much larger temporal variability of the weather. Subtracting the station observation from ERA5 interpolation removes much of this resolved temporal variability, while averaging over snapshots suppresses transient noise. The resulting target therefore isolates the persistent spatial component that a surface descriptor should organise.

We define the ERA5-interpolation residual at station s as the mean signed interpolation error over its set T_s of valid test snapshots:

$$r_s = \frac{1}{|T_s|} \sum_{t \in T_s} \left(y_{s,t}^{\text{interp}} - y_{s,t} \right), \tag{15}$$

where $y_{s,t}^{\text{interp}}$ is the bilinearly interpolated ERA5 value and $y_{s,t}$ is the corresponding station observation. The negative residual, $-r_s$, is the mean additive correction that a downscaler would need to apply to ERA5 at station s. Since reversing the target sign does not affect the relative predictive performance of the regressors, analysing r_s is equivalent to analysing the required correction for the purpose of comparing descriptor spaces.

Unlike the residual of a trained ConvCNP, r_s is constructed without using any station descriptor. Geographical coordinates, topography, ERA5-static features[2], and the Tessera embedding can therefore all be evaluated independently against the same target. Appendix C extends the same probe to a richer hand-crafted descriptor.

Descriptor-space regression probe. For each candidate descriptor space $\mathcal{D}$, let $d_s^{\mathcal{D}} \in \mathbb{R}^{p_{\mathcal{D}}}$ denote its $p_{\mathcal{D}}$-dimensional feature vector at station s. We fit a separate random-forest regressor

$$\widehat{r}_s^{\mathcal{D}} = f_{\mathcal{D}}\big(d_s^{\mathcal{D}}\big) \tag{16}$$

[2]We use the ERA5 static fields bilinearly interpolated to each station, not the model's CNN-encoded grid Z. The encoded grid Z is dominated by the time-dependent dynamic weather processed by the same encoder, which is the field being *corrected* at the station, not a persistent driver of the local correction. Including it would flag Norway as out-of-distribution for reasons orthogonal to whether a surface analogue exists. The static interpolant isolates the persistent surface character.

for each descriptor space. The random forest is not intended as the final downscaling model, but as a common nonlinear probe. It tests whether the information and geometry of descriptor space $\mathcal{D}$ make the persistent ERA5-interpolation residual predictable at stations excluded from fitting.

Predictive performance is measured using 5-fold cross-validation over stations, and we report the mean of the per-fold $R^2_{\mathrm{CV}}(\mathcal{D})$. A positive coefficient means that descriptor d predicts residuals at held-out stations more accurately than the constant predictor $\widehat{r} = \bar{r}$, where $\bar{r}$ is the mean residual estimated from the corresponding training folds. Values near zero indicate that the descriptor exposes little transferable residual structure, while negative values indicate worse generalisation than the mean predictor. Since the residual target, folds, and regression procedure are held fixed across descriptors, the ranking of $R^2_{\mathrm{CV}}(\mathcal{D})$ identifies which representation most effectively *organises* the persistent sub-grid residual.

Figure 5 ranks descriptors only in the well-sampled regions (Europe and United States). The impartial target is a per-station average over a finite number of valid snapshots, so each $\hat{r}_s$ carries sampling noise, and the CV R^2 is computed over the S stations in a region. With S of order tens (Southern Africa 37, Australia 13–14), a handful of noisy stations can reorder the descriptors.

Temperature-network diagnostics. In the United States, no descriptor attains materially positive performance for temperature. This is consistent with the station network containing substantially less unresolved topographic variation than the European network: 24% of European stations are displaced by more than 200 m from the linearly interpolated ERA5 orography, and the residual correlates with that displacement at $\rho = 0.83$, compared with 5% and $\rho = 0.32$ in the United States. A topographic descriptor can organise the interpolation residual only to the extent that unresolved topographic mismatch is present in the evaluated station network.

F Variable-specific degradation with forecast lead

As the Aurora context degrades, Figure 10 shows t2m CRPS growing steeply, rising by ∼20–27% from lead 0 to +72 h, while wind degrades far less (∼3–8%). This behaviour holds across both regions, and for MAE and RMSE. It reads that t2m's downscaling correction is closely tied to the coarse forecast grid, which loses accuracy with lead, degrading the downscaling correction with it. Wind's correction is governed more by the local surface, which the embedding supplies and is lead-invariant.

G Descriptor-space coverage of Norwegian probe stations

The simulated deployment in §3.5 shows that TESSERA improves wind-speed downscaling before any Norwegian observations are available. Here, we investigate a possible mechanism for this cold-start advantage by asking whether held-out Norwegian stations have nearby analogues in the existing European training set under different location descriptors. We compare the *four* descriptor spaces $\mathcal{D}$ used in §3.3: geography, elevation + mTPI, the interpolated ERA5 static fields, and TESSERA.

Within the training set, the nearest-neighbour distance of each station is

$$\nu_s^{\mathcal{D}} \;=\; \min_{s' \in \mathcal{T}(h),\, s' \neq s} \left\| d_s^{\mathcal{D}} - d_{s'}^{\mathcal{D}} \right\|_2, \qquad s \in \mathcal{T}(h), \tag{17}$$

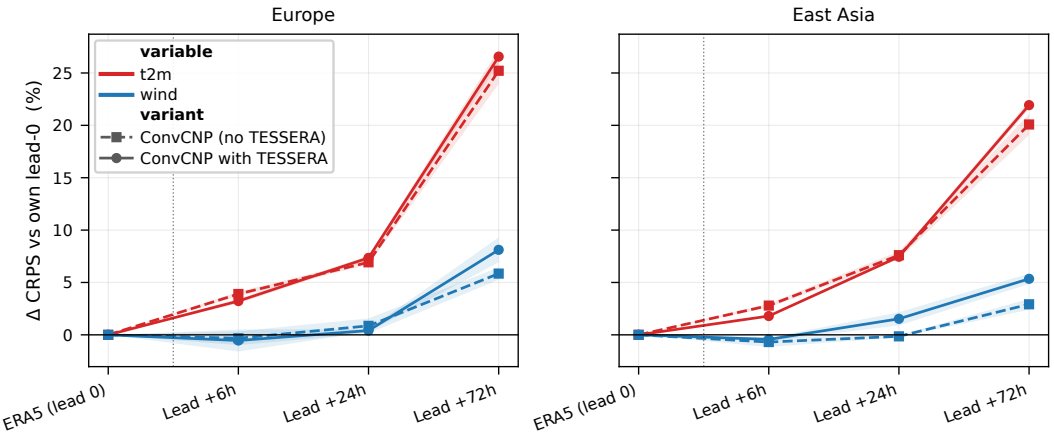


Figure 10: Relative CRPS skill decay with forecast lead, normalised separately for each model variant to its own lead-0 value, in Europe (left) and East Asia (right). Curves show 2 m temperature and 10 m wind speed for ConvCNP with (solid) and without (dashed) TESSERA. Because each curve is normalised to its own lead-0 error, the figure compares rates of degradation across leads and variables, not absolute error between model variants.

where $d_s^{\mathcal{D}} \in \mathbb{R}^{p_{\mathcal{D}}}$ denotes the feature vector in space $\mathcal{D}$ at station s. We take the reachability radius to be their 95th percentile, $\tau^{\mathcal{D}}(h) = Q_{0.95}\big(\{\nu_s^{\mathcal{D}} : s \in \mathcal{T}(h)\}\big)$. A held-out station is reachable if it has a training neighbour within $\tau^{\mathcal{D}}$, and *reachability* is the fraction of the out-of-training Norwegian set that has an in-distribution analogue in the training set,

$$\mathrm{Reach}^{\mathcal{D}}(h) \;=\; \frac{1}{|\mathcal{H}_{\mathrm{NO}}^{\mathrm{unseen}}(h)|} \sum_{s \in \mathcal{H}_{\mathrm{NO}}^{\mathrm{unseen}}(h)} \mathbb{1}\Big[\min_{s' \in \mathcal{T}(h)} \big\|d_s^{\mathcal{D}} - d_{s'}^{\mathcal{D}}\big\|_2 \;\le\; \tau^{\mathcal{D}}(h)\Big]. \tag{18}$$

High reachability alone does not establish that the resulting neighbours are informative analogues. A descriptor may place many stations close together simply because it omits characteristics that distinguish their surface environments, hence our use of *separability* AUC as a control metric. For each descriptor, we fit an L_2-regularised logistic classifier under 4-fold cross-validation to the membership label $z_s = \mathbb{1}[\,s \in \mathcal{H}_{\mathrm{NO}}^{\mathrm{unseen}}(h)\,]$, giving each station an out-of-fold score $\hat{p}_s$. The separability AUC is the probability that the classifier ranks a random held-out Norwegian station above a random training station. AUC $= \frac{1}{2}$ means the two are statistically indistinguishable (Norway in-distribution), and AUC $\to 1$ means the descriptor resolves it as distinct.

TESSERA provides both local support and descriptor fidelity at cold start. Figure 11 visualises each descriptor space at the cold-start extreme. For each non-geographic descriptor, the horizontal axis is the normalised Norway-vs-rest discriminant: the leading direction $u^{\mathcal{D}}$ of a linear discriminant analysis separating the Norwegian stations from the non-Norwegian training stations in that descriptor space, normalised to unit length. A station's coordinate is the projection of its standardised descriptor onto this direction. The vertical axis carries the first principal component of what remains after this Norway-vs-rest direction is removed, $d_s^{\mathcal{D}} - \langle u^{\mathcal{D}}, d_s^{\mathcal{D}} \rangle\, u^{\mathcal{D}}$.

Geographically, Norway begins as an almost pure extrapolation problem, with only ∼2% of Norwegian stations having a nearby European training station. ERA5-static provides intermediate coverage of ∼50%. The TESSERA and elevation + mTPI spaces, by contrast, provide near-complete reachability at 96% and 95%, respectively.

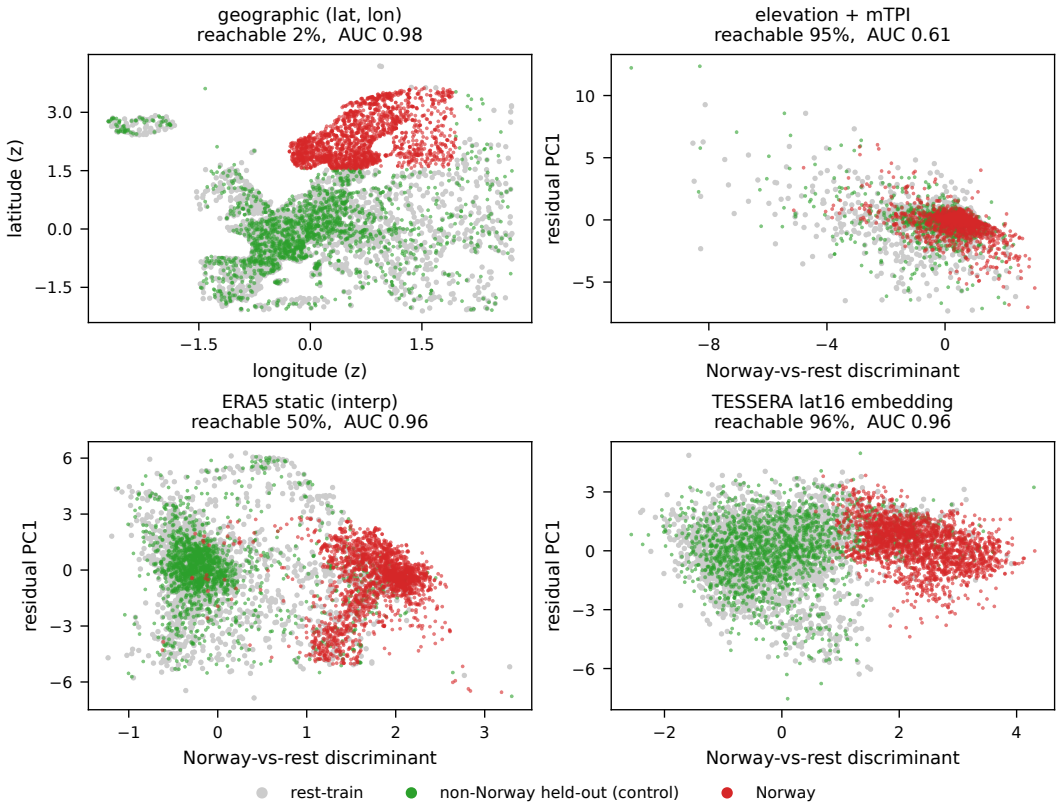


Figure 11: Where each descriptor places Norway relative to the European training set prior to deploying any stations. Each panel is a 2-D view of one descriptor space. For non-geographic descriptor spaces, the horizontal axis is the *Norway-vs-rest discriminant* — each station's score along the linear combination of that descriptor's features that best separates Norway from the training set — and the vertical axis is the leading principal component of the residual descriptor variation. The geographic panel shows standardised longitude and latitude. Grey: a subsample of the non-Norwegian European training stations; green: a non-Norwegian held-out control; red: Norway. In TESSERA space, the Norwegian stations overlap locally with a subset of the European training stations, providing candidate surface analogues, while remaining globally distinguishable from the broader non-Norwegian distribution.

The separability results distinguish these latter two cases. Elevation + mTPI cannot tell Norway apart from the rest of Europe (AUC ≈ 0.6), with its high reachability coming at the cost of defining similarity through a non-discriminative feature: many European stations share similar elevation and terrain position despite potentially differing in surface characteristics, potentially transferring incorrect corrections. By contrast, TESSERA combines near-complete local coverage with strong global separability with 0.96 AUC, indicating that its reachable neighbours are faithful surface analogues. This strong combination is illustrated by the red cluster overlapping locally with a subset of the grey European training stations while remaining strongly linearly separable from the non-Norwegian held-out stations in the green cluster.

The cold-start geometry persists as stations are added. Figure 12 shows that the relative ordering of the descriptor spaces changes little during deployment. Geographical reachability rises from approximately 2% at cold start to approximately 90% at full deployment, while ERA5-static

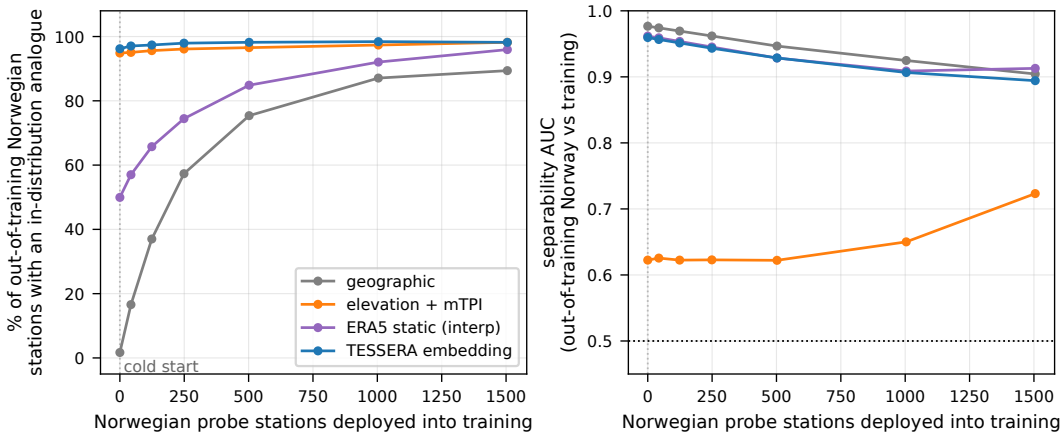


Figure 12: At each point along the phased deployment, the out-of-training Norwegian stations $\mathcal{H}_{\mathrm{NO}}^{\mathrm{unseen}}(h)$ (not-yet-deployed probes plus the permanently held-out Norwegian test set) are compared against the current training set $\mathcal{T}(h)$ (non-Norway Europe plus the probes deployed by horizon h) in four per-station descriptor spaces, all in standardised coordinates. *(a)* Reachability (eq. (18)): the fraction of the Norwegian stations with a training neighbour within the reachability radius $\tau^{\mathcal{D}}(h)$, the 95th-percentile nearest-neighbour distance among the training stations themselves — i.e. the fraction with an in-distribution analogue in the training set. *(b)* Separability AUC: a cross-validated L_2-logistic classifier's probability of ranking a random Norwegian station above a random training one, measuring how distinguishable the two are (0.5 = statistically indistinguishable / in-distribution; $\to 1$ = distributionally distinct).

coverage rises from an intermediate starting point. In contrast, TESSERA and elevation + mTPI are already near saturation before any Norwegian probes are deployed and remain at $\sim 96\%$. The corresponding separability curves also remain qualitatively stable, indicating that the cold-start comparison captures the main descriptor-space geometry relevant to the experiment.

Connection to the wind speed cold-start advantage. The analysis indicates that TESSERA provides the strongest combination of local coverage and descriptor fidelity. This distinction is especially relevant for wind speed, given §3.3's findings that its residual is organised more effectively in TESSERA space than by explicit topography or ERA5-static fields. The nearby, *reachable* European training examples identified by TESSERA are therefore more likely relevant to the correction being transferred.

This provides an input-space explanation for the result in §3.5: the per-location embedding allows the model to reuse wind corrections learned at previously observed surface analogues before Norwegian observations are available. It also clarifies why increasing station count alone does not remove TESSERA's advantage since, for wind speed, additional observations alone do not recover the missing surface structure information governing the local correction.

Nevertheless, this is a *preliminary, input-space* geometric reading: reachability and separability describe only how each feature clusters the stations, computed independently of the downscaling target. Whether the TESSERA descriptor improves the downscaling depends on how much a location's correction, per-variable, is governed by fine surface structure rather than by the smooth, large-scale fields and the less discriminative elevation features both models share. These input-space diagnostics simply provide a plausible mechanism for the gains measured directly in Figure 7.