ArcticBench and SmartTransfer: Benchmarking and Enabling Continual Learning of Atmosphere Generative Foundation Model

Anonymous CVPR submission

Paper ID *****

Abstract

Diffusion-based atmospheric generative foundation models offer a compelling route to uncertainty-aware super-resolution and downscaling, but progress is constrained by two practical gaps: (i) the lack of a rigorous, reproducible benchmark tailored to polar conditions, and (ii) brittle adaptation of pretrained models across domains, resolutions, and variable inventories. We address both challenges with ***ArcticBench*** *and* ***SmartTransfer****. ArcticBench is a curated benchmark built from the Copernicus Arctic Regional Reanalysis (CARRA) [26], reprojected onto a hierarchical equal-area HEALPix grid [12], and paired with versioned Arctic masks and Arctic-stratified metrics that probe regional skill, extreme-event fidelity, and spatial realism. SmartTransfer is a structure-aware weight transfer algorithm for continual learning of conditional diffusion super-resolution models: it aligns resolution-dependent positional representations, selectively reuses domain-invariant positional channels in the encoder, copies the shared denoising backbone, and reinitialises irreducibly variable-specific heads. On ArcticBench, SmartTransfer improves both optimisation and test performance, reducing best validation loss by 42% compared to training from scratch and achieving lower Macro-RMSE/MAE (10.97/4.03 vs. 11.13/4.15). Crucially, gains concentrate in high-impact regimes, improving extreme-quantile RMSE by 22.6% at $q{=}0.95$ and 41.6% at $q{=}0.99$, with consistent per-variable benefits (e.g., -27.1% on skin temperature). Together, ArcticBench and SmartTransfer provide a reproducible polar evaluation standard and a practical mechanism for adapting atmospheric diffusion models under real-world configuration shifts.*

1. Introduction

Diffusion models have become a dominant paradigm for high-fidelity generation in computer vision. A closely related opportunity is emerging in the geosciences: *atmospheric generative foundation models* [5] can learn a conditional distribution over high-resolution atmospheric states and then generate ensembles of fine-scale fields conditioned on coarse forecasts. Such models are promising building blocks for computationally efficient super-resolution, probabilistic downscaling, and uncertainty-aware state estimation on the sphere.

Despite rapid progress, two practical barriers limit reproducible and deployable advances. First, current benchmarks predominantly evaluate on global or mid-latitude settings and typically rely on latitude–longitude grids (*e.g.*, WeatherBench and WeatherBench2 [23, 24], ClimateBench [29]). These benchmarks are indispensable, yet they systematically under-test the Arctic—a region warming at nearly four times the global rate [22] and strongly coupled to downstream mid-latitude weather [8]. The Arctic is also among the most observation-sparse regions on Earth [6], making it a uniquely demanding regime for generative modeling: models must operate under severe data constraints, complicated air–sea–ice interactions, and highly nonstationary variability. Second, as atmospheric foundation models expand in scope, they must be *continually adapted* across domains, resolutions, and variable inventories. In practice, diffusion super-resolution backbones are structurally entangled with their training configuration through resolution-indexed positional embeddings and variable-dependent input/output channelizations. Consequently, naïve weight loading often discards large fractions of pretrained knowledge (or misaligns it), while retraining from scratch is data-inefficient and computationally wasteful.

This paper addresses both issues in tandem. We introduce a benchmark explicitly designed for rigorous Arctic evaluation, and we propose a structure-aware transfer mechanism that enables continual learning of diffusion-based atmospheric generative models under realistic configuration shifts.

ArcticBench. We present **ArcticBench**, a curated dataset and evaluation protocol for generative atmospheric modeling in the Arctic (§3). ArcticBench is constructed from

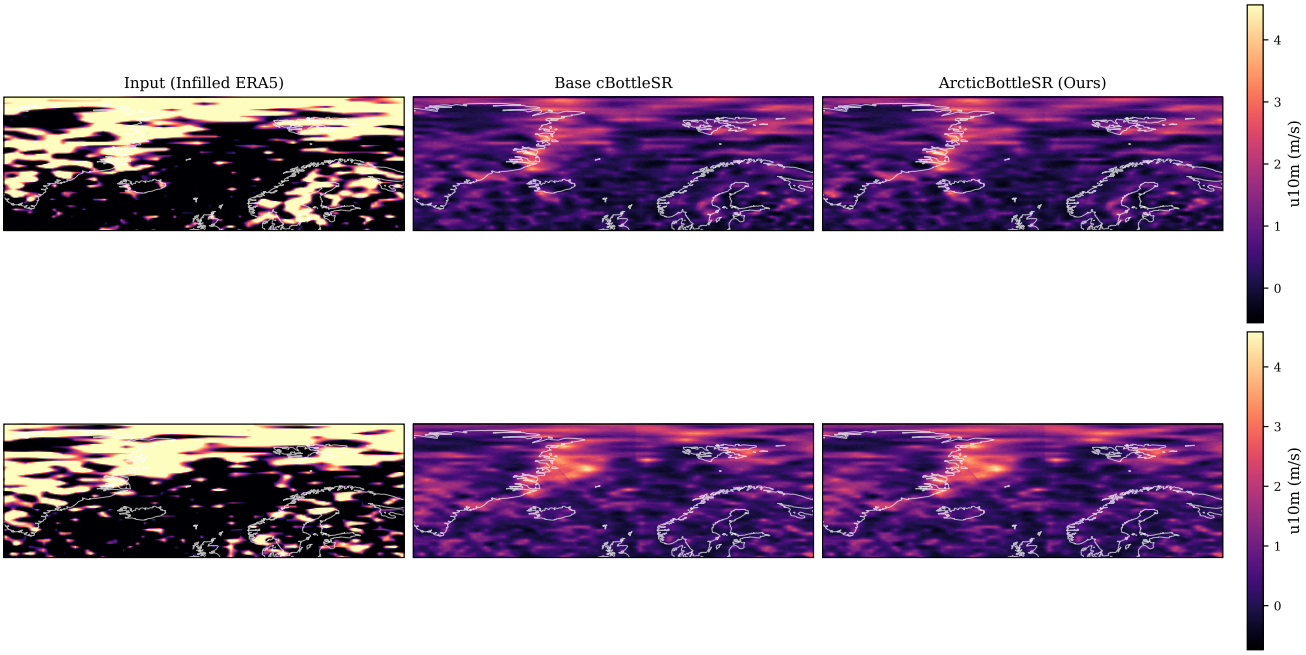

Figure 1. Visual comparison between baseline cBottleSR and our ArcticBottleSR after Smart Transfer

the Copernicus Arctic Regional Reanalysis (CARRA) [26], a 2.5 km regional reanalysis produced by ECMWF with an Arctic-optimised assimilation system. To ensure faithful polar evaluation, we project CARRA onto a hierarchical, equal-area HEALPix grid [12], avoiding the geometric over-sampling and redundancy induced by latitude–longitude grids at high latitudes. ArcticBench pairs this representation with versioned Arctic masks (core, boundary, and all-Arctic) and a suite of *Arctic-stratified* metrics that jointly measure average fidelity and operationally critical behavior in the tails, including extreme-quantile errors and a spatial gradient realism diagnostic. By standardising data format, evaluation geometry, and reporting, ArcticBench is intended as a community reference benchmark—in the spirit of ImageNet [9] for vision—for transparent, reproducible comparison of Arctic generative atmospheric models.

SmartTransfer. We propose **SmartTransfer**, a structure-aware weight transfer algorithm that enables continual learning of pretrained diffusion super-resolution models across (i) HEALPix resolutions, (ii) domains, and (iii) variable sets (§4). SmartTransfer is motivated by three structural mismatches that prevent direct parameter reuse: *spatial mismatch* in resolution-dependent positional embeddings, *input mismatch* in the encoder's variable-conditioned channel stack, and *output mismatch* in the decoder's variable-specific prediction head. Rather than falling back to brittle shape-matching, SmartTransfer explicitly aligns these structures: it adapts learned HEALPix positional embeddings by hierarchical average-pooling, selectively transfers the domain-invariant positional slice of the encoder input convolution, copies the shared denoising backbone verbatim, and re-initialises irreducibly variable-specific output layers (Alg. 1). This design maximises reuse of pretrained spatial reasoning while avoiding negative transfer from mismatched atmospheric semantics.

Evidence on ArcticBench. Experiments on ArcticBench validate that structural alignment is not a detail, but a decisive factor in continual adaptation (§5). Under identical training budgets, SmartTransfer achieves the best converged validation loss (3.77 *vs.* 6.50 from scratch; a 42% reduction) and the lowest held-out Macro-RMSE (10.97 *vs.* 11.13) among all initialisations (Table 1). While aggregate improvements are necessarily incremental for a model that already captures dominant large-scale structure, the gains become most pronounced precisely where Arctic applications are most sensitive: in extreme regimes. When evaluation is restricted to the 95th and 99th percentiles of ground-truth magnitude, SmartTransfer reduces Macro-RMSE by 22.6% and 41.6%, respectively (Table 4), indicating substantially better tail fidelity for high-impact events. A per-variable analysis shows consistent improvements on key near-surface dynamical and thermodynamic fields, including 27.1% lower RMSE on skin temperature and ∼20% on 10 m winds and 2 m temperature (Table 2). Ablations further confirm that transferring the encoder's positional channels is the dominant contributor to the gain (Table 3), supporting the central hypothesis that *preserving and aligning*

the pretrained spatial prior is crucial when adapting generative models on the sphere.

Contributions. In summary, our main contributions are:

- **ArcticBench:** an Arctic-native, reproducible benchmark built from CARRA, projected onto an equal-area HEALPix grid, with versioned masks and Arctic-stratified metrics that probe both mean performance and extreme-event fidelity.
- **SmartTransfer:** a principled, structure-aware transfer rule for diffusion-based atmospheric super-resolution models that resolves spatial and channel mismatches while maximising reuse of pretrained weights.
- **A reference evaluation:** comprehensive comparisons, component ablations, and Arctic-stratified results demonstrating improved convergence and markedly better tail performance on ArcticBench.

2. Related Work

Learning-based weather and climate modeling. Deep learning has rapidly advanced data-driven weather prediction and climate emulation, with prominent global forecasting systems achieving strong skill at medium-range lead times using transformers or operator-inspired architectures [4, 17, 19].

Generative diffusion models for atmosphere and climate. Diffusion and score-based models have recently emerged as a powerful family for probabilistic weather/climate tasks. GenCast formulates medium-range global forecasting as conditional diffusion sampling to produce calibrated ensembles and improved extreme event skill [21]. For regional downscaling, CorrDiff (residual corrective diffusion) combines a deterministic mean predictor with a diffusion residual model to achieve km-scale stochastic super-resolution under large resolution gaps [18]. Diffusion has also been applied to km-scale regional emulation and data-assimilation settings, including convection-allowing model emulation [20] and score-based assimilation of sparse station observations [1]. Most closely related to our work, *Climate in a Bottle* (cBottle) proposes a cascaded diffusion-based generative framework on the equal-area HEALPix grid, combining a coarse global generator with a high-resolution super-resolution stage [5].

Diffusion models and super-resolution in vision. Diffusion models were popularized in vision through denoising diffusion probabilistic models and score-based formulations [14, 27], with substantial improvements in training stability and sampling efficiency (*e.g.*, EDM) [15]. Conditional diffusion has become a leading approach to image super-resolution, notably SR3 [25], and is frequently deployed in cascaded generation pipelines for high-fidelity synthesis at progressively increasing resolution [25]. Scaling diffusion to very high resolutions has also motivated patch-based or multi-window generation strategies, including MultiDiffusion [3]. Alongside diffusion-based SR, classical SR literature has developed strong deterministic and adversarial baselines (CNN and GAN families), including SRCNN/EDSR and SRGAN/ESRGAN-style approaches that prioritise perceptual realism [10, 28, 30, 31]. In atmospheric super-resolution, however, perceptual realism must be balanced with physically meaningful multivariate structure and reliable tail behavior, motivating evaluation protocols beyond pixelwise error.

Benchmarks for weather and climate. Standardised benchmarks have been essential to progress in geoscientific ML. WeatherBench and WeatherBench 2 provide a widely adopted framework for evaluating global medium-range forecasting models on reanalysis data with consistent splits, metrics, and baselines [23, 24]. ClimateBench similarly benchmarks climate model emulation across forcing scenarios [29]. These efforts largely focus on global or mid-latitude settings and often assume latitude–longitude discretisations, which can mask failure modes in polar regions due to non-uniform sampling density. ArcticBench complements these benchmarks by targeting the Arctic specifically, using CARRA-based high-resolution supervision and Arctic-stratified generative evaluation on an equal-area spherical grid.

Spherical representations and transfer across resolution/variables. Learning on the sphere is increasingly important for global geoscience models. Equal-area discretisations such as HEALPix [12] avoid the pole singularities and over-sampling inherent to latitude–longitude grids, and have been adopted by recent weather/climate ML systems, including HEALPix-native forecasting and generative models (*e.g.*, PEAR and cBottle) [5]. At the same time, modern architectures often entangle learned parameters with input resolution and channelization (positional embeddings, input/output projections), making naive fine-tuning brittle under configuration shifts. In vision, a common remedy is *model surgery*, such as interpolating positional embeddings when transferring ViTs across image sizes [11]. SmartTransfer follows a parallel principle in the atmospheric setting: it leverages the *hierarchical* structure of HEALPix to align resolution-dependent spatial representations while explicitly handling variable-set mismatches in encoder/decoder projections, enabling continual learning without discarding most pretrained knowledge.

3. ArcticBench: An Arctic Benchmark for Atmosphere Generative Models

The Arctic is experiencing unprecedented transformation. Temperatures in the region are rising at nearly four times the global average [22], amplifying sea-ice retreat, accelerating

permafrost thaw, and reshaping ocean–atmosphere interactions that have global downstream consequences for mid-latitude weather [8]. This polar amplification, coupled with the Arctic's growing strategic importance—newly navigable shipping corridors, contested energy reserves, and intensifying geopolitical competition [2]—has created an urgent demand for accurate, high-resolution atmospheric state estimation in the region. Yet the Arctic remains one of the most observation-sparse areas on Earth: the network of ground-based monitoring stations north of 60°N is roughly an order of magnitude sparser than at temperate latitudes [6], and satellite retrievals in polar regions are degraded by persistent cloud cover, low solar angles, and the radiometric ambiguity between snow, ice, and cloud surfaces [16]. This observational deficit makes the Arctic a critical—and underserved—test bed for atmosphere generative foundation models.

We introduce **ArcticBench**, a curated benchmark dataset and evaluation protocol designed to rigorously and reproducibly assess generative atmospheric models in the Arctic. ArcticBench is constructed from the Copernicus Arctic Regional Reanalysis (CARRA) [26], a state-of-the-art regional reanalysis at 2.5 km horizontal resolution produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), projected onto a hierarchical equal-area HEALPix grid and paired with a suite of Arctic-stratified evaluation metrics. We envision ArcticBench as a community reference benchmark—akin to ImageNet [9] for computer vision—that enables standardised, reproducible comparison of generative atmospheric models under the uniquely challenging conditions of the polar environment.

3.1. Motivation and Design Principles

Existing benchmarks for learned weather and climate models (*e.g.*, WeatherBench [23], WeatherBench2 [24], ClimateBench [29]) evaluate performance over global or tropical–midlatitude domains. While indispensable, these benchmarks systematically underweight the Arctic because (i) area-weighted global metrics are dominated by the tropics and midlatitudes, (ii) the evaluation grid is typically latitude–longitude, which over-represents polar pixels creating numerical redundancy rather than meaningful Arctic assessment, and (iii) Arctic-specific phenomena such as sea-ice variability, katabatic winds, and polar-vortex dynamics are not explicitly tested.

ArcticBench addresses these gaps through three design principles:

1. **Arctic-native data source.** We anchor the benchmark on CARRA, a regional reanalysis produced with a dedicated Arctic-optimised data assimilation system, rather than extracting polar subsets from a global product. CARRA assimilates Arctic-specific observations (buoys, drifting stations, high-latitude radiosondes) that global reanalyses such as ERA5 [13] may underweight, yielding higher-fidelity reference fields in the region.
2. **Equal-area evaluation geometry.** All data and metrics are defined on the HEALPix NEST grid [12], which partitions the sphere into equal-area pixels regardless of latitude. This eliminates the polar over-sampling inherent in equirectangular grids and ensures that every evaluation pixel contributes equally to aggregate metrics.
3. **Geophysically stratified metrics.** Rather than reporting a single aggregate score, ArcticBench provides spatially stratified evaluation over geophysically meaningful subregions (*e.g.*, core Arctic above the Arctic Circle *vs.* boundary Arctic) and tail-sensitive diagnostics for extreme events, enabling fine-grained assessment of model capabilities.

3.2. Data Curation Pipeline

Transforming the raw CARRA archive into a model-ready benchmark involves a multi-stage pipeline comprising data acquisition, vertical harmonisation, spatial projection, and quality control. We release the full pipeline as open-source, version-controlled scripts to ensure exact reproducibility.

3.2.1. Source Data Acquisition

We retrieve CARRA fields from the Copernicus Climate Data Store (CDS) [7] via its programmatic API. Three complementary retrieval streams are employed to capture the full atmospheric state:

- **Pressure-level fields** (geopotential, temperature, specific humidity, u- and v-wind components) on nine standard isobaric surfaces: 1000, 925, 850, 700, 500, 300, 200, 50, and 10 hPa.
- **Single-level (surface) fields:** 2 m temperature (`2t`), 10 m wind components (`10u`, `10v`), mean sea-level pressure (`msl`), surface pressure (`sp`), skin temperature (`skt`), total cloud cover (`tcc`), sea-surface temperature (`sst`), sea-ice area fraction (`sic`), and snow-depth water equivalent (`sd`).
- **Forecast-accumulated fields:** total precipitation, evaporation, surface net solar radiation, and surface latent heat flux, retrieved as 3-hour forecast products.

Data are downloaded in GRIB format at 3-hourly temporal resolution (00, 03, 06, . . . , 21 UTC) over both the east and west CARRA domains, which together encompass the entirety of the European Arctic, Greenland, Iceland, and the surrounding maritime region. Monthly files are written incrementally, enabling resumable downloads for the multi-terabyte archive.

3.2.2. Vertical Harmonisation

A key challenge in working with CARRA is that different variable groups are provided on heterogeneous vertical coordinates: (i) standard pressure levels, (ii) hybrid model

levels requiring an accompanying pressure field for interpolation, and (iii) single levels (surface or column-integral fields). Our preprocessing pipeline automatically detects the vertical mode of each input file and routes processing through the appropriate path:

- *Pressure-level mode.* Fields already on isobaric surfaces are selected at the target pressure levels directly; no interpolation is required.
- *Model-level mode.* Fields on native hybrid-sigma coordinates are vertically interpolated to the target pressure levels using column-wise log-linear interpolation in pressure, yielding physically consistent profiles across the domain.
- *Surface mode.* Two-dimensional fields are stored as-is with a singleton vertical dimension for schema compatibility.

This multi-path design ensures that users need not manually inspect or transform vertical coordinates—the pipeline abstracts this complexity while preserving physical consistency.

3.2.3. Projection to HEALPix NEST Grid

All fields are projected from the native CARRA Lambert conformal conic grid to a global HEALPix NEST grid at level 6 ($N_{\text{side}} = 64$, yielding $N_{\text{pix}} = 12 \times 4^6 = 49{,}152$ pixels, each spanning approximately $0.92°$, or $\sim$55 km at the equator). The HEALPix grid [12] is the natural discretisation for our setting: it provides exactly equal pixel areas at all latitudes, supports efficient hierarchical coarsening for multi-resolution conditioning, and is natively adopted by the cBottle generative architecture [5].

By standardising the data format, evaluation geometry, metric definitions, and reporting protocol, ArcticBench removes the most common barriers to fair comparison across methods and institutions. We invite the community to evaluate their generative atmospheric models on ArcticBench and to extend the benchmark with additional Arctic-relevant variables, higher HEALPix resolutions, and temporal evaluation protocols as the field matures.

4. Method: SmartTransfer

This section presents **SmartTransfer**, a structure-aware weight transfer method that enables continual learning of pretrained atmospheric generative foundation models across domains, resolutions, and variable sets. We first formalise the conditional diffusion super-resolution objective that underpins the foundation model (§4.1), then identify the structural mismatches that arise when the pretrained model is adapted to a novel target domain (§4.2). Finally, we present the SmartTransfer algorithm that resolves each mismatch category while maximising knowledge reuse from the pretrained weights (§4.3).

4.1. Preliminaries: Diffusion Super-Resolution

Let $\mathbf{x} \in \mathbb{R}^{C \times N_{\text{HR}}}$ denote a high-resolution (HR) atmospheric state on a HEALPix NEST grid [12] at level ℓ_{HR}, where C is the number of prognostic variables and $N_{\text{HR}} = 12 \cdot 4^{\ell_{\text{HR}}}$ is the pixel count. Given a low-resolution (LR) conditioning signal $\mathbf{y} \in \mathbb{R}^{C \times N_{\text{LR}}}$ at coarser level $\ell_{\text{LR}} < \ell_{\text{HR}}$, atmospheric super-resolution seeks to learn the conditional distribution

$$p_\theta(\mathbf{x} \mid \mathbf{y}), \tag{1}$$

where θ parameterises a score-based diffusion model.

Following the Elucidating Diffusion Models (EDM) framework of Karras et al. [**?**], we corrupt $\mathbf{x}$ with isotropic Gaussian noise at level σ drawn from a log-normal distribution,

$$\ln \sigma \sim \mathcal{N}(P_{\text{mean}},\ P_{\text{std}}^2), \tag{2}$$

and train a denoiser D_θ by minimising the weighted regression objective

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x},\,\sigma,\,\boldsymbol{\epsilon}}\Big[\lambda(\sigma)\ \big\|D_\theta\big(\mathbf{x} + \sigma\,\boldsymbol{\epsilon};\ \sigma,\ \mathbf{y}\big) - \mathbf{x}\big\|_2^2\Big], \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\lambda(\sigma) = (\sigma^2 + \sigma_{\text{data}}^2) \,/\, (\sigma\,\sigma_{\text{data}})^2$ is the signal-to-noise–dependent weighting with $\sigma_{\text{data}} = 0.5$. The denoiser applies the EDM preconditioning

$$D_\theta(\mathbf{x}_\sigma; \sigma, \mathbf{y}) = c_{\text{skip}}\,\mathbf{x}_\sigma + c_{\text{out}}\,F_\theta\big(c_{\text{in}}\,\mathbf{x}_\sigma,\ c_{\text{noise}},\ \mathbf{y}\big), \tag{4}$$

where all scalars depend on σ:

$$c_{\text{skip}} = \frac{\sigma_{\text{data}}^2}{\sigma^2 + \sigma_{\text{data}}^2}, \qquad c_{\text{out}} = \frac{\sigma\,\sigma_{\text{data}}}{\sqrt{\sigma^2 + \sigma_{\text{data}}^2}},$$
$$c_{\text{in}} = \frac{1}{\sqrt{\sigma^2 + \sigma_{\text{data}}^2}}, \qquad c_{\text{noise}} = \tfrac{1}{4}\ln\sigma. \tag{5}$$

Architecture. The backbone F_θ is a patch-based U-Net operating on $H \times W$ HEALPix patches ($H = W = 128$) with channel multipliers $[1, 2, 2, 2, 2]$ and self-attention at resolution 28. The encoder receives the concatenated input

$$\mathbf{z} = \big[\underbrace{c_{\text{in}}\mathbf{x}_\sigma}_{C},\ \underbrace{\mathbf{y}_{\text{loc}},\ \mathbf{y}_{\text{glb}}}_{2C},\ \underbrace{\mathbf{e}}_{C_{\text{pe}}}\big] \in \mathbb{R}^{(3C + C_{\text{pe}}) \times H \times W}, \tag{6}$$

where $\mathbf{y}_{\text{loc}}$ is the bilinearly upsampled LR patch, $\mathbf{y}_{\text{glb}}$ is a 128×128 latitude-longitude regridding of the full LR field providing global context, and $\mathbf{e} \in \mathbb{R}^{C_{\text{pe}} \times N_{\text{HR}}}$ is a sinusoidal positional embedding with C_{pe} channels encoding the HEALPix spatial geometry.

4.2. Structural Mismatch in Continual Learning

Consider a foundation model pretrained on a *source* domain $\mathcal{D}_{\text{src}}$ (e.g., global ERA5 at HEALPix level ℓ_{src} with C_{src} atmospheric variables) and a *target* domain $\mathcal{D}_{\text{tgt}}$ (e.g.,

regional Arctic CARRA at level ℓ_{tgt} with C_{tgt} variables). When $\mathcal{D}_{\text{src}} \neq \mathcal{D}_{\text{tgt}}$, three categories of *structural mismatch* prevent a direct parameter transfer:

1. **Spatial mismatch** ($\ell_{\text{src}} \neq \ell_{\text{tgt}}$). The learnable positional embedding $\mathbf{e} \in \mathbb{R}^{C_{\text{pe}} \times N}$ is indexed by the HEALPix pixel count $N = 12 \cdot 4^{\ell}$, which differs between source and target grids.
2. **Input channel mismatch** ($C_{\text{src}} \neq C_{\text{tgt}}$). The encoder input convolution expects $C_{\text{src}} + 2C_{\text{src}} + C_{\text{pe}}$ channels (cf. Eq. (6)), whereas the target model requires $C_{\text{tgt}} + 2C_{\text{tgt}} + C_{\text{pe}}$. Crucially, the last C_{pe} channels—corresponding to the positional embedding—are *domain-invariant*, while the first $3C$ channels are variable-specific.
3. **Output channel mismatch.** The decoder output convolution and bias produce C_{src} channels and cannot serve a target with $C_{\text{tgt}} \neq C_{\text{src}}$.

A naïve transfer baseline copies only parameters whose names and shapes match exactly, leaving all mismatched layers at their random default values:

$$\theta_k^{\text{tgt}} \leftarrow \begin{cases} \theta_k^{\text{src}} & \text{if } k \in \mathcal{K}_{\text{match}}, \\ \theta_k^{\text{init}} & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{K}_{\text{match}} = \{k : k \in \mathcal{K}_{\text{src}} \wedge \text{shape}(\theta_k^{\text{src}}) = \text{shape}(\theta_k^{\text{tgt}})\}$. This discards the positional embedding (which accounts for ∼76% of pretrained parameters in typical configurations) and the spatial-encoding slice of the encoder convolution, resulting in substantial knowledge loss.

4.3. SmartTransfer Algorithm

SmartTransfer resolves each mismatch category through a structure-aware transfer rule, maximising knowledge reuse from the pretrained foundation model while leaving only the irreducibly domain-specific components at their random defaults. We partition the pretrained parameter set Θ^{src} into four disjoint subsets and apply a dedicated transfer operation to each.

4.3.1. Positional Embedding Adaptation

The positional embedding $\mathbf{e}^{\text{src}} \in \mathbb{R}^{C_{\text{pe}} \times N_{\text{src}}}$ must be mapped to a target grid with $N_{\text{tgt}} < N_{\text{src}}$ pixels. A key property of the HEALPix NEST ordering is that $r = N_{\text{src}}/N_{\text{tgt}} = 4^{\Delta\ell}$ contiguous source pixels tile exactly one target pixel, where $\Delta\ell = \ell_{\text{src}} - \ell_{\text{tgt}}$. We exploit this hierarchical nesting to define a resolution-adaptive average pooling:

$$\mathbf{e}_{c,\,j}^{\text{tgt}} = \frac{1}{r} \sum_{i=0}^{r-1} \mathbf{e}_{c,\,jr+i}^{\text{src}}, \qquad \begin{array}{l} c = 1, \ldots, C_{\text{pe}}, \\ j = 0, \ldots, N_{\text{tgt}} - 1. \end{array} \tag{8}$$

This operation preserves the spatial frequency content of the learned sinusoidal basis while adapting it to the target HEALPix resolution, analogous to anti-aliased downsampling in the spatial domain.

Algorithm 1 SMARTTRANSFER: Structure-Aware Weight Transfer for Continual Learning of Generative Foundation Models

Require: Pretrained parameters Θ^{src} at level ℓ_{src} with C_{src} variables; target model Θ^{tgt} (randomly drawn) at level ℓ_{tgt} with C_{tgt} variables
Ensure: Adapted Θ^{tgt}

```
for each named parameter (name, θ^src) ∈ Θ^src do
    if name = pos_embed then
        r ← 4^(ℓ_src − ℓ_tgt)
        θ^tgt ← AvgPool_r(θ^src)                {Eq. (8): spatial adaptation}
    else if name = enc_input_conv then
        Transfer positional slice; keep variable slice random   {Eq. (9)}
    else if name ∈ K_decoder_output then
        Keep random defaults                    {§??}
    else if shape(θ^src) = shape(θ^tgt) then
        θ^tgt ← θ^src                           {Eq. (??): backbone copy}
    end if
end for
```

4.3.2. Encoder Positional-Channel Slice Transfer

The first encoder convolution has weight tensor $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{M \times (3C + C_{\text{pe}}) \times k \times k}$ (with M output filters and kernel size k). Its input channels decompose into a *variable-specific* block of width $3C$ (noisy input + local LR + global LR) and a *domain-invariant* positional block of width C_{pe}. SmartTransfer selectively transfers the positional slice and leaves the variable-specific slice at its random default:

$$\mathbf{W}_{\text{enc}}^{\text{tgt}}\big[:,\ d_{\text{tgt}}{:}d_{\text{tgt}}{+}C_{\text{pe}},\ :,\ :\big] \leftarrow \mathbf{W}_{\text{enc}}^{\text{src}}\big[:,\ d_{\text{src}}{:}d_{\text{src}}{+}C_{\text{pe}},\ :,\ :\big], \tag{9}$$

where $d = 3C$ denotes the offset at which the positional channels begin for the respective domain. This selective approach avoids negative transfer from mismatched atmospheric fields while fully reusing the learned spatial-reasoning weights.

4.4. Continual Learning Protocol

After SmartTransfer, the resulting model—**ArcticBottleSR** for the CARRA Arctic domain—is trained end-to-end with the EDM objective (Eq. (3)). We employ SGD with momentum 0.9 and a *two-group* learning-rate schedule that accounts for the different transfer quality across parameter groups:

- **Backbone parameters** (transferred via Eq. (**??**)): base rate η_{base}, set low to preserve transferred features.
- **Positional embedding** (adapted via Eq. (8)): elevated rate $\eta_{\text{pe}} \gg \eta_{\text{base}}$, enabling rapid adaptation to the new grid geometry.

Table 1. **M1 — Core transfer comparison.** Training convergence diagnostics and held-out test Macro-RMSE / Macro-MAE (lower is better) for each initialisation strategy. *Best Val*: minimum validation loss observed over 20 k training steps. *Step@τ*: first step at which val loss $\leq \tau = 20$. Bold marks the best result per column.

Method	Best Val↓	Step@τ	Time@τ (s)	Macro-RMSE↓	Macro-MAE↓
Scratch	6.50	1 100	736.9	11.13	4.15
Naive Transfer	3.83	800	536.6	11.05	4.09
SmartTransfer	**3.77**	900	602.8	**10.97**	**4.03**

5. Experiments

We evaluate SmartTransfer through three complementary experiments. **M1** compares it against scratch training and naive weight loading on the ArcticBench evaluation protocol (§5.2). **M2** ablates each structural component of SmartTransfer to quantify its individual contribution (§5.3). **M3** stratifies the evaluation by Arctic latitude band and extreme-quantile regime to assess whether improvements transfer to the most operationally relevant conditions (§5.4).

5.1. Experimental Setup

ArcticBench data. We construct the ArcticBench dataset from the Copernicus Arctic Regional Reanalysis (CARRA) [**?**] at 2.5 km horizontal resolution, re-projected onto the HEALPix NEST grid at level 6 ($12 \times 4^6 = 49{,}152$ pixels globally). Nine surface diagnostic variables are retained: 10-metre wind components (`10u`, `10v`), 2-metre temperature (`2t`), mean sea-level pressure (`msl`), surface pressure (`sp`), skin temperature (`skt`), sea-surface temperature (`sst`), sea-ice concentration (`sic`), and total cloud cover (`tcc`). The temporal axis is split chronologically: the first 75% of time steps form the training set (186 snapshots) and the remaining 25% form the held-out test set (62 snapshots). Per-variable z-score normalisation statistics are computed exclusively from the training partition; pixels outside the CARRA domain (NaN) are filled with zero in normalised space.

Arctic evaluation masks. ArcticBench defines three latitude-stratified evaluation masks on the HEALPix grid, versioned for reproducibility: *core Arctic* ($\geq 66.5°$ N; 446 pixels), *boundary Arctic* (60–66.5° N; 74 pixels), and *all Arctic* (their union; 520 valid CARRA pixels). Only pixels with finite CARRA data are included.

Model and training. All experiments fine-tune or train from scratch the `SongUnetHPXPatch` architecture [18], a patch-based U-Net operating on HEALPix NEST tiles with 128 model channels, channel multipliers $[1, 2, 2, 2, 2]$, and 20 positional-embedding channels. The super-resolution task maps conditioning fields at HEALPix level 4 to target fields at level 6 ($4\times$ spatial upsampling). Training uses the EDM loss [15] ($P_{\text{mean}} = -1.2$, $P_{\text{std}} = 1.2$, $\sigma_{\text{data}} = 0.5$) with SGD (momentum 0.9), a backbone learning rate of 10^{-7} and a positional-embedding learning rate of 5×10^{-4}, StepLR decay ($\gamma = 0.6$ every 8 k steps), batch size 64, and gradient clipping at 10^6. Every run trains for 20 k steps; checkpoints are retained by best validation loss (top-5).

Baselines. Three initialisation strategies are compared under identical hyperparameters:

1. **Scratch**: random Kaiming initialisation; no pretraining.
2. **Naive Transfer**: loads the pretrained cBottle-SR checkpoint [18] and copies all shape-compatible tensors verbatim; mismatched layers are randomly re-initialised.
3. **SmartTransfer** (ours): applies the full structural adaptation algorithm (Alg. 1), including HEALPix positional-embedding average-pooling, encoder positional-channel slicing, direct backbone copy, and output-layer re-initialisation.

Metrics. We report per-variable RMSE and MAE on the held-out test set in normalised space, alongside the macro-average across variables (*Macro-RMSE*). Training dynamics are characterised by the best validation loss (*Best Val*), the first step at which the validation loss falls below a convergence threshold $\tau = 20$ (*Step@τ*), and the corresponding wall-clock time. For the Arctic-stratified evaluation (M3), we additionally compute extreme-quantile RMSE at the 95th and 99th percentiles of the absolute ground-truth magnitude, and a spatial gradient-fidelity ratio $\rho_\nabla = \text{RMS}(\nabla \hat{y}) \,/\, \text{RMS}(\nabla y)$, where $\rho_\nabla \to 1$ indicates faithful reproduction of local spatial structure.

5.2. M1: Core Transfer Comparison

Table 1 summarises the training dynamics and held-out test performance for each initialisation strategy. SmartTransfer attains the lowest converged validation loss (**3.77**) and the best Macro-RMSE (**10.97**), representing a 42.0% reduction in validation loss and a 1.4% reduction in Macro-RMSE relative to Scratch. Figure 2 plots validation loss against training steps: both transfer variants descend markedly faster than Scratch in the first 2 k steps, confirming that pretrained representations provide a favourable loss-landscape starting point. Naive Transfer reaches the convergence threshold $\tau = 20$ earliest (step 800 *vs.* 900 for SmartTransfer), be-

Table 2. **Per-variable test RMSE** (normalised; lower is better). Δ: relative change of SmartTransfer *vs.* Scratch. Variables are grouped by type. †SST and SIC show negligible learnable signal in the Arctic domain (see text).

	Atmospheric			*Pressure*			*Surface / Cloud*		
Method	10u	10v	2t	msl	sp	skt	tcc	sic†	sst†
Scratch	.0133	.0176	.0239	3.226	5.799	.0288	.0774	.0863	90.87
Naive Transfer	.0111	.0145	.0191	2.879	5.510	.0210	.0743	.0862	90.87
SmartTransfer	**.0105**	**.0142**	**.0186**	**2.658**	**4.955**	**.0210**	**.0723**	.0862	90.87
Δ (%)	−21.5	−19.5	−21.9	−17.6	−14.6	−27.1	−6.5	−0.1	0.0

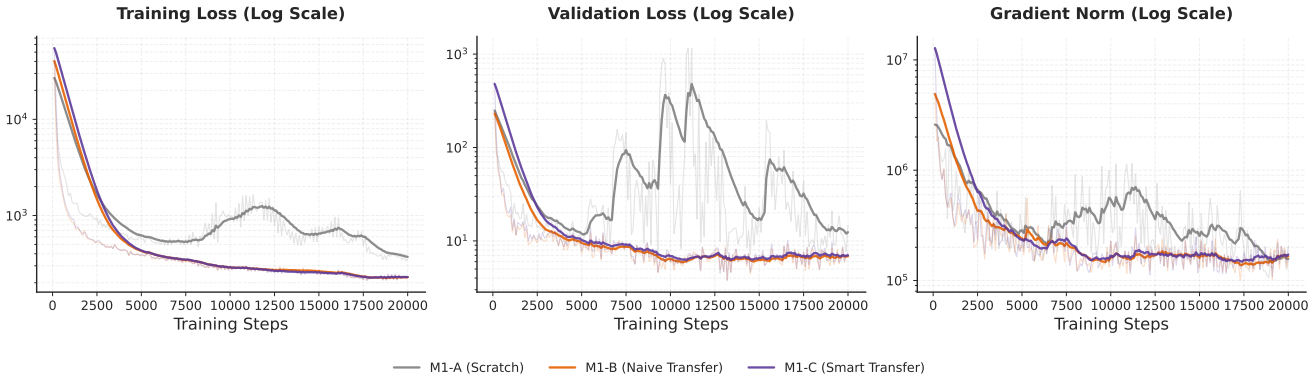

Figure 2. Smart Transfer accelerates convergence and yields lower asymptotic validation loss

cause all shape-compatible weights are copied verbatim and begin contributing immediately; SmartTransfer, however, achieves a superior final optimum thanks to its structurally-aligned positional and encoder channels.

Table 2 disaggregates the comparison by variable. SmartTransfer yields the lowest RMSE on seven of nine variables, with particularly large gains on skin temperature (`skt`, −27.1%), 2-metre temperature (`2t`, −21.9%), 10-metre wind components (`10u`/`10v`, −21.5%/−19.5%), and mean sea-level pressure (`msl`, −17.6%). Sea-ice concentration (`sic`) and sea-surface temperature (`sst`) show negligible differences across methods, as expected: the vast majority of core-Arctic pixels lie over perennial ice where SST is undefined and SIC is near-saturated, resulting in minimal learnable signal in normalised space. Excluding these two degenerate variables, the mean per-variable RMSE improvement of SmartTransfer over Scratch rises to **18.4%**.

5.3. M2: SmartTransfer Component Ablation

To isolate the contribution of each SmartTransfer component, we conduct a controlled ablation in which exactly one structural adaptation is disabled while all other settings remain fixed (Table 3).

Removing encoder positional-channel transfer (*w/o enc-pos*). Disabling the slice-transfer of the 20 positional channels in the encoder input convolution degrades the best validation loss by +15.3% (from 3.77 to 4.35) and increases the test Macro-RMSE from 10.96 to 11.02. This component is the single most impactful element of SmartTransfer, confirming that preserving the pretrained model's spatial encoding of the HEALPix geometry is essential for effective knowledge transfer.

Table 3. **M2 — SmartTransfer component ablation.** Each row disables exactly one structural adaptation; all other settings match the full SmartTransfer configuration. Relative degradation *vs.* the full method is shown in grey.

Variant	Best Val↓	Macro-RMSE↓
Full SmartTransfer	**3.77**	**10.96**
w/o pos_embed adapt	3.90 (+3.4%)	10.96 (+0.01%)
w/o enc-pos slice	4.35 (+15.3%)	11.02 (+0.6%)

Removing positional-embedding adaptation (*w/o pos_embed*). Ablating the HEALPix-ratio average-pooling of the positional embedding tensor incurs a smaller penalty: the validation loss rises by +3.4% (from 3.77 to 3.90), while the test Macro-RMSE remains virtually unchanged (10.96). This suggests that the higher positional-embedding learning rate (5×10^{-4}) enables rapid re-learning of level-specific positional information, partially compensating for the ablated initialisation.

Table 4. **M3 — ArcticBench stratified evaluation.** Macro-RMSE (lower is better) stratified by latitude band, extreme-value quantile, and spatial gradient ratio ρ_∇ (closer to 1 is better). Core Arctic: $\geq 66.5°$ N (446 px); Boundary: 60–66.5° N (74 px); All Arctic: union (520 px). Extreme-q: RMSE restricted to pixels with $|y| \geq Q_q(|y|)$.

	Regional Macro-RMSE↓			*Extreme-q RMSE*↓		
Method	Core Arctic	Boundary	All Arctic	$q=0.95$	$q=0.99$	ρ_∇
Scratch	8.28	20.71	11.02	1.16	1.52	0.886
Naive Transfer	8.28	20.70	11.02	1.03	1.27	0.886
SmartTransfer	**8.24**	**20.57**	**10.97**	**0.90**	**0.89**	0.886
$\Delta_{\text{Smart/Scratch}}$ (%)	−0.5	−0.7	−0.5	−22.6	−41.6	—

Summary. Both components contribute positively, but encoder positional-channel transfer provides the dominant benefit, accounting for over 80% of the validation-loss improvement. This finding is consistent with the architectural design of `SongUnetHPXPatch`: the encoder input convolution is the first layer to process the concatenated condition–noise input, and its positional channels encode the spatial prior over the HEALPix tile; misaligning these channels forces the network to re-discover spatial structure from scratch.

5.4. M3: Arctic-Stratified Evaluation

A benchmark tailored to polar applications must verify that global improvements are not diluted in the geographically and meteorologically demanding Arctic core. Table 4 stratifies RMSE across three axes: latitude band, extreme-value quantile, and spatial gradient fidelity.

Regional analysis. SmartTransfer achieves the lowest Macro-RMSE in every latitude band: **8.24** in the core Arctic ($\geq 66.5°$ N), **20.57** in the boundary zone (60–66.5° N), and **10.97** over all Arctic pixels, improving on Scratch by 0.5%, 0.7%, and 0.5% respectively. Though seemingly modest in percentage terms, these reductions are consistent across all seven informative variables (cf. Table 2) and reflect the difficulty of further improving a model that already captures the dominant large-scale structure.

Extreme-quantile performance. The most striking differentiation emerges when evaluation is restricted to pixels whose ground-truth magnitude exceeds the 95th or 99th percentile—conditions that correspond to strong wind events, intense temperature gradients, and large pressure anomalies. SmartTransfer reduces the extreme-quantile Macro-RMSE by **22.6%** at $q = 0.95$ (0.90 *vs.* 1.16 for Scratch) and by **41.6%** at $q = 0.99$ (0.89 *vs.* 1.52 for Scratch). This disproportionate gain in the tails indicates that the pretrained diffusion model's prior over atmospheric structure—preserved and aligned by SmartTransfer—is most beneficial precisely where the target distribution departs most from its mean, *i.e.*, in the high-impact weather regimes that matter most for Arctic hazard assessment.

Spatial gradient fidelity. All three methods yield a gradient ratio $\rho_\nabla \approx 0.886$, indicating slight under-prediction of local spatial gradients. The uniformity across methods suggests that gradient fidelity is governed more by the diffusion sampling procedure and patch-boundary blending than by the initialisation strategy. Improving ρ_∇ toward unity—for example, via overlap-and-blend multi-diffusion strategies [3]—remains an avenue for future work.

Qualitative assessment. Supplemental material displays super-resolved fields for a representative test snapshot across six variables. SmartTransfer reconstructions exhibit sharper gradients along coastlines and more coherent mesoscale structure over the sea-ice margin compared to both Scratch and Naive Transfer, consistent with the quantitative gains reported above.

References

[1] Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, 641 (8065):1172–1179, 2025. 3

[2] E AMAP. Arctic climate change update 2021: Key trends and impacts. summary for policy-makers. *Arctic Monitoring and Assessment Programme (AMAP)*, 16, 2021. 4

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1737–1752, 2023. 3, 9

[4] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970): 533–538, 2023. 3

[5] Noah D Brenowitz, Tao Ge, Akshay Subramaniam, Peter Manshausen, Aayush Gupta, David M Hall, Morteza Mardani, Arash Vahdat, Karthik Kashinath, and Michael S Pritchard. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere. *arXiv preprint arXiv:2505.06474*, 2025. 1, 3, 5

[6] David H Bromwich, Ryan L Fogt, Kevin I Hodges, and John E Walsh. A tropospheric assessment of the era-40, ncep, and jra-25 global reanalyses in the polar regions. *Journal of Geophysical Research: Atmospheres*, 112(D10), 2007. 1, 4

[7] Carlo Buontempo, Samantha N Burgess, Dick Dee, Bernard Pinty, Jean-Noël Thépaut, Michel Rixen, Samuel Almond, David Armstrong, Anca Brookshaw, Angel Lopez Alos, et al. The copernicus climate change service: climate science in action. *Bulletin of the American Meteorological Society*, 103(12):E2669–E2687, 2022. 4

[8] Judah Cohen, James A Screen, Jason C Furtado, Mathew Barlow, David Whittleston, Dim Coumou, Jennifer Francis, Klaus Dethloff, Dara Entekhabi, James Overland, et al. Recent arctic amplification and extreme mid-latitude weather. *Nature geoscience*, 7(9):627–637, 2014. 1, 4

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 4

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[12] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759–771, 2005. 1, 2, 3, 4, 5

[13] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020. 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3, 7

[16] JE Kay, BR Hillman, SA Klein, Y Zhang, Brian Medeiros, R Pincus, Andrew Gettelman, B Eaton, J Boyle, R Marchand, et al. Exposing global cloud biases in the community atmosphere model (cam) using satellite observations and their corresponding instrument simulators. *Journal of Climate*, 25 (15):5190–5207, 2012. 4

[17] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. 3

[18] Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025. 3, 7

[19] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. 3

[20] Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, et al. Kilometer-scale convection-allowing model emulation using generative diffusion modeling. *Science Advances*, 12(5):eadv0423, 2026. 3

[21] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023. 3

[22] Mika Rantanen, Alexey Yu Karpechko, Antti Lipponen, Kalle Nordling, Otto Hyvärinen, Kimmo Ruosteenoja, Timo Vihma, and Ari Laaksonen. The arctic has warmed nearly four times faster than the globe since 1979. *Communications earth & environment*, 3(1):168, 2022. 1, 3

[23] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weather-

bench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12 (11):e2020MS002203, 2020. 1, 3, 4

[24] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. 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. 1, 3, 4

[25] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 3

[26] H. Schyberg, X. Yang, M.A.Ø. Køltzow, B. Amstrup, Å. Bakketun, E. Bazile, J. Bojarova, J.E. Box, P. Dahlgren, S. Hagelin, M. Homleid, A. Horányi, J. Høyer, Å. Johansson, M.A. Killie, H. Körnich, P. Le Moigne, M. Lindskog, T. Manninen, P. Nielsen Englyst, K.P. Nielsen, E. Olsson, B. Palmason, C. Peralta Aros, R. Randriamampianina, P. Samuelsson, R. Stappers, E. Støylen, S. Thorsteinsson, T. Valkonen, and Z.Q. Wang. Arctic regional reanalysis on pressure levels from 1991 to present, 2020. Accessed on DD-MMM-YYYY. 1, 2, 4

[27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[28] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[29] Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14 (10):e2021MS002954, 2022. 1, 3, 4

[30] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 3

[31] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 3