UNIVERSAL DIFFUSION-BASED PROBABILISTIC DOWNSCALING

Roberto Molinaro[1*] **Niall Siegenheim**[1*] **Henry Martin**[1] **Mark Frey**[1]
Niels Poulsen[1] **Philipp Seitz**[1] **Marvin Vincent Gabler**[1]

[1]Jua.ai

ABSTRACT

We introduce a universal diffusion-based downscaling framework that lifts deterministic low-resolution weather forecasts into probabilistic high-resolution predictions without any model-specific fine-tuning. A single conditional diffusion model is trained on paired coarse-resolution inputs ($\sim$25 km resolution) and high-resolution regional reanalysis targets ($\sim$5 km resolution), and is applied in a fully zero-shot manner to deterministic forecasts from heterogeneous upstream weather models. Focusing on near-surface variables, we evaluate probabilistic forecasts against independent in situ station observations over lead times up to 90 h. Across a diverse set of AI-based and numerical weather prediction (NWP) systems, the ensemble mean of the downscaled forecasts consistently improves upon each model's own raw deterministic forecast, and substantially larger gains are observed in probabilistic skill as measured by CRPS. These results demonstrate that diffusion-based downscaling provides a scalable, model-agnostic probabilistic interface for enhancing spatial resolution and uncertainty representation in operational weather forecasting pipelines.

1 INTRODUCTION

Accurate weather prediction requires approximating solutions of highly nonlinear partial differential equations governing atmospheric flow. Operational global forecasting systems represent the atmosphere on finite grids, typically at horizontal resolutions of 9–25 km for global deterministic forecasts. At these resolutions, many near-surface processes and land–atmosphere interactions remain only partially resolved. Sub-grid-scale effects associated with terrain, surface roughness, land use, and boundary-layer structure are therefore suppressed or parameterized, leading to systematic loss of small-scale spatial variability. This loss directly impacts forecast skill for near-surface variables such as 2 m temperature and 10 m wind speed, which are strongly controlled by local surface heterogeneity (Maraun, 2010; Benestad, 2008).

A widely used strategy to address this limitation is *downscaling*, i.e. the enrichment of coarse-resolution forecasts with fine-scale spatial detail. Dynamical downscaling resolves the governing equations on a finer grid using a nested regional model driven by coarse boundary conditions (Giorgi & Mearns, 1999). While physically consistent, this approach remains computationally expensive and tightly coupled to specific NWP model configurations.

An alternative class of methods, commonly referred to as *statistical* or data-driven downscaling, seeks to infer fine-scale fields directly from coarse-resolution predictors using empirical or learned mappings (Wilby et al., 1998a; Maraun, 2010). From a modeling perspective, statistical downscaling constitutes an ill-posed inverse problem: the mapping from a coarse atmospheric state to fine-scale realizations is not unique, and multiple high-resolution fields may be consistent with the same large-scale conditions. As a result, deterministic regression-based approaches whether based on linear models, analog methods, or supervised machine learning tend to converge toward conditional means, producing overly smooth fields and systematically under-representing variability and uncertainty (Benestad, 2008; Maraun, 2010). This limitation is particularly pronounced for near-surface variables and in complex terrain.

[*]Equal contribution. Correspondence to *firstname.lastname*@jua.ai.

In addition, access to probabilistic forecasts is essential for many weather-dependent decision-making applications. Nevertheless, operational systems are often restricted to deterministic outputs because uncertainty estimation, typically achieved via ensemble models is computationally expensive.

A natural remedy is to abandon deterministic mappings altogether and instead formulate downscaling as the problem of learning the *conditional distribution* of fine-scale states given a coarse-resolution forecast. However, learning high-dimensional conditional distributions in a stable and scalable manner has remained a major challenge for data-driven downscaling.

Recent advances in generative modeling offer a promising pathway to address this challenge. In particular, diffusion-based models have demonstrated the ability to learn complex, high-dimensional probability distributions and to generate realistic samples conditioned on auxiliary inputs (Ho et al., 2020; Karras et al., 2022; Molinaro et al., 2025a). In this work, we adopt this perspective and formulate high-resolution downscaling as a conditional generative modeling problem. We introduce a diffusion-based framework that learns a stochastic mapping from coarse-resolution atmospheric states to distributions of fine-scale surface fields, trained solely on paired reanalysis data.

A key feature of our approach is its *model-agnostic* design. The downscaler is trained once using low-resolution→high-resolution *analysis or reanalysis* pairs and is subsequently applied *zero-shot* as a post-processing step to a diverse set of *heterogeneous deterministic forecasts*, spanning both AI-based and NWP systems. This decouples the learning of fine-scale surface structure from the upstream forecast model and allows a single probabilistic module to be reused across forecasting pipelines with different numerical formulations, resolutions, and systematic biases.

Using independent in situ station observations, we show that diffusion-based downscaling consistently improves forecast quality across all tested upstream models. Improvements are observed both in point accuracy, measured by RMSE of the ensemble mean, and—more prominently—in probabilistic skill, measured by CRPS. This demonstrates that diffusion-based downscaling enables uncertainty-aware probabilistic forecasting from deterministic inputs, rather than improving point predictions alone.

Our main contributions are:

- We introduce a *single*, model-agnostic conditional diffusion downscaler trained solely on low-resolution→high-resolution reanalysis pairs, which can be applied *zero-shot* as a post-processing step to heterogeneous *operational* deterministic forecasts from both AI-based and NWP models.
- We show that a single deterministic forecast can be lifted to a probabilistic forecast without explicit ensembling, providing uncertainty-aware predictions at a fraction of the computational cost of traditional ensemble systems.
- We demonstrate through independent weather station verification that this plug-and-play downscaling consistently improves forecast quality across diverse upstream systems *at continental (European) scale*, yielding gains in both point accuracy and probabilistic skill across multiple near-surface variables.

2 Related Work

In recent years, there has been substantial interest in learning solution operators to partial differential equations from data, rather than solving the governing equations using numerical integration (Li et al., 2021; Raonić et al., 2023; Herde et al., 2024). Weather and climate modeling has proven itself as a particularly fertile ground for such approaches, given the vast amounts of historical data available from reanalysis products (Hersbach et al., 2020) and the high computational cost of traditional numerical solvers (Pathak et al., 2022; Lam et al., 2023; Lang et al., 2024; Bodnar et al., 2024).

While many AI-based weather models have shown impressive results in terms of error metrics compared to their numerical counterparts such as the ECMWF IFS (European Centre for Medium-Range Weather Forecasts (ECMWF), 2024), their deterministic training objectives (e.g., mean squared error) lead them to regress to the conditional mean of the target distribution, resulting in smoothed predictions that lack physically relevant small-scale structure (Selz et al., 2025). To address this limitation, recent works have explored using diffusion models to learn probabilistic weather

forecasts that capture the full distribution of possible outcomes (Price et al., 2025; Cachay et al., 2025). However, these large models are computationally expensive during inference time.

In practice, short-term high-resolution regional weather forecasts are often preferred to global ones, as they can better resolve local features such as orography, land-sea contrasts, and surface heterogeneity that strongly influence near-surface variables. The most notable examples are NOAA's High Resolution Rapid Refresh (HRRR) model (Dowell et al., 2022) and ICON-EU, a European model operated by DWD (Zängl et al., 2015). These operate as limited-area models (LAMs), which numerically integrate the governing equations over a restricted spatial domain using boundary conditions provided by a global model.

In the case of HRRR, there have been attempts at using machine learning techniques to obtain similar improvements as in global weather forecasting models by replacing the numerical time integration with a learned diffusion-based model (Pathak et al., 2024; Abdi et al., 2025). However, they continue to rely on a global numerical weather prediction model for their boundary conditions. While there has been research on more localised AI-based LAMs for Europe (Oskarsson et al., 2023; Adamov et al., 2025), to the best of our knowledge, there have been no published attempts at building an AI-based LAM which covers the whole European domain as ICON-EU does.

Concurrently, statistical downscaling techniques focus on enhancing the resolution of coarse global weather states by learning a mapping from low-resolution inputs to high-resolution outputs using historical data, without incorporating any time integration. Dating back to the last century (Wilby & Wigley, 1997; Wilby et al., 1998b), machine learning-based approaches have accelerated this line of research in recent years (Leinonen et al., 2021; Mardani et al., 2024; Wan et al., 2023; Bischoff & Deck, 2024; Merizzi et al., 2024; Wan et al., 2025; Tomasi et al., 2025; Glawion et al., 2025; Schillinger et al., 2025). Most of these works use diffusion models to learn these mappings, given their success in generating realistic images in computer vision (Rombach et al., 2022).

While these efforts are formidable, we believe these works have yet to fully explore the potential of diffusion-based downscaling as a universal, model-agnostic interface between upstream forecasting systems and high-resolution probabilistic predictions. Concretely, they suffer from one or more of the following limitations: (i) They do not evaluate their downscaling models when used for operational high-resolution forecasting. (ii) They focus on smaller regions (e.g. Italy (Merizzi et al., 2024; Tomasi et al., 2025), Taiwan (Mardani et al., 2024)) rather than continental scales. (iii) They focus on downscaling single variables (e.g. precipitation (Leinonen et al., 2021; Glawion et al., 2025)).

In this work, we address all of these limitations by exploring statistical downscaling through diffusion models as a means of *improving operational medium-range weather forecasts* on a continental scale across multiple relevant surface variables.

3 Problem Formulation

The large-scale atmospheric state evolves according to the hydrostatic primitive equations, which describe the conservation of momentum, mass, and energy. Following Holton & Hakim (2012), let $\mathbf{v}$ denote the horizontal velocity, T the temperature, p the pressure, and ρ the density. Using pressure as the vertical coordinate and writing ∇ for the horizontal gradient, the governing equations are

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} + f\,\mathbf{k} \times \mathbf{v} = -\frac{1}{\rho}\nabla p + \mathbf{F}, \tag{1}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{2}$$

$$\frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T = \frac{Q}{c_p} + \frac{RT\,\omega}{c_p p}, \tag{3}$$

$$\frac{\partial p}{\partial z} = -\rho g, \qquad p = \rho R T, \tag{4}$$

where $\omega \equiv Dp/Dt$ denotes the pressure vertical velocity. In height coordinates, the material derivative reads

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla + w\,\frac{\partial}{\partial z}, \tag{5}$$

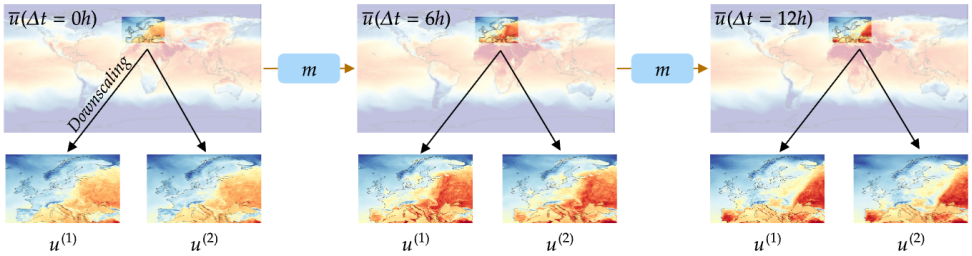


Figure 1: Model-agnostic conditional diffusion downscaling. A coarse-resolution forecast $\bar{u}$, produced by different upstream models $m \in \mathcal{M}$, conditions a reverse diffusion process. Integrating the probability-flow dynamics yields multiple high-resolution realizations sampled from $p_\theta(u \mid \bar{u})$.

yielding

$$\omega = \frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p - \rho g\, w, \tag{6}$$

after invoking hydrostatic balance.

In practice, operational forecasting systems approximate solutions of eqs. (1) to (4) on discrete grids. Let $\bar{u} \in \mathcal{U}_{\mathrm{LR}}$ denote a coarse-resolution atmospheric state produced by an upstream forecasting system, and let $u \in \mathcal{U}_{\mathrm{HR}}$ denote a corresponding high-resolution surface state defined on a finer grid. The downscaling problem consists in inferring plausible realizations of u given $\bar{u}$. Since the mapping $\bar{u} \mapsto u$ is not injective, this inference problem is intrinsically non-unique. We therefore formulate downscaling as a conditional density estimation task and seek to learn the distribution $p(u \mid \bar{u})$, rather than a single deterministic mapping.

Learning this conditional distribution relies on training data consisting of paired low- and high-resolution analysis or reanalysis fields $\mathcal{D} = \{(\bar{u}_i, u_i)\}_{i=1}^{N}$ that are aligned in time and represent the same underlying atmospheric state at different spatial resolutions. Each pair constitutes a *single* high-resolution realization associated with its coarse counterpart; multiple realizations u for the same $\bar{u}$ are not observed. The conditional distribution is therefore not available empirically but must be inferred statistically from the ensemble of such pairs across time.

At inference time, the conditioning input $\bar{u}$ may originate from *any deterministic forecasting system* and from *any forecast lead time* Δt. Let $m \in \mathcal{M}$ index such a system, producing a coarse forecast $\bar{u}^{(m)}$ on its native grid and variable conventions. Our objective is to learn a single conditional model $p_\theta(u \mid \tilde{u})$ that can be applied zero-shot to all $m \in \mathcal{M}$ (see Figure 1 for a schematic representation of the process). This formulation yields a probabilistic downscaler that produces ensembles of high-resolution realizations. The ensemble mean provides a point forecast, while the ensemble variability captures uncertainty arising from unresolved small-scale structure.

4 Diffusion-Based Downscaling Model

To address the learning task depicted in Section 3, we adopt a conditional generative modeling approach and represent the distribution $p(u \mid \bar{u})$ using a score-based diffusion formulation, following Song et al. (2021). Instead of defining sampling from an explicit parametric density, we use a generative stochastic process.

Concretely, we embed the conditional data distribution into a family of noise-perturbed conditionals indexed by an artificial diffusion time $\tau \in [0, K]$. Given paired samples $(\bar{u}, u) \in \mathcal{D}$, the forward diffusion process progressively corrupts the high-resolution target by additive Gaussian noise according to $u_\tau = u + \eta,\ \eta \sim \mathcal{N}(0, \sigma_\tau^2 I)$, where σ_τ denotes a prescribed noise schedule. For sufficiently large σ_τ, the conditional distribution $p_\tau(u_\tau \mid \bar{u})$ becomes indistinguishable from an isotropic Gaussian, providing a tractable terminal distribution from which sampling can be initiated.

The generative task then consists of reversing this noising process, transforming a sample u_K drawn from the terminal noise distribution into a sample from the target conditional distribution $p(u \mid \bar{u})$. In the continuous-time limit, this reverse dynamics is described by the (backward-time) stochastic

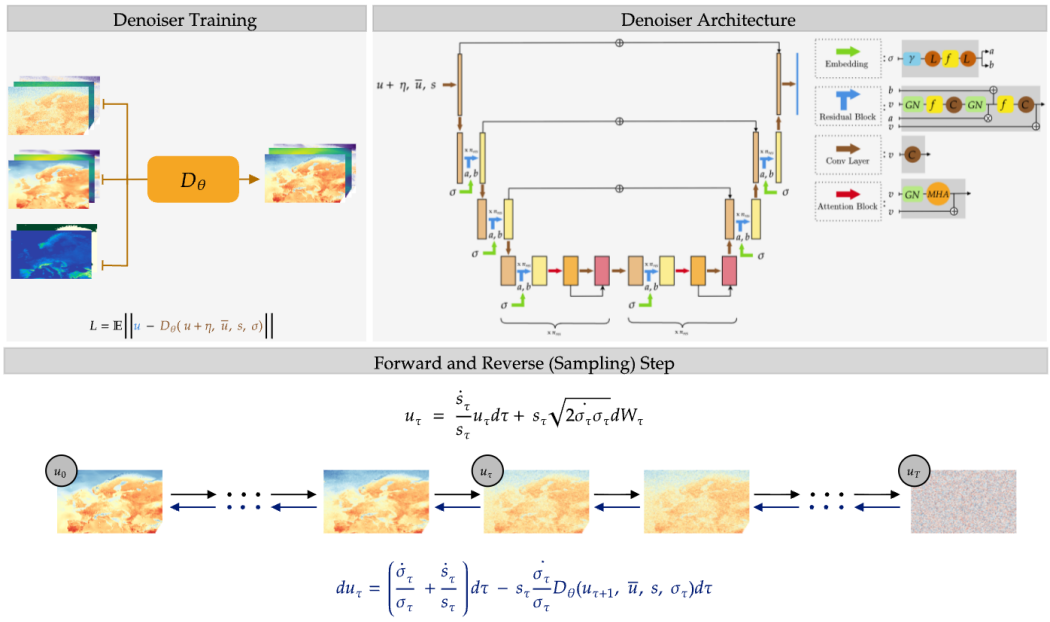


Figure 2: **Diffusion-based downscaling architecture and sampling procedure.** *Top left:* Denoiser training. A clean high-resolution target u is corrupted with Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_\tau^2 I)$ and, together with the conditioning inputs, passed to the denoiser D_θ, which is trained to reconstruct u via a mean-squared denoising objective. *Top right:* Denoiser architecture. A U-Net encoder–decoder with residual blocks at each resolution and attention at the bottleneck. The noisy sample u_τ is concatenated with the coarse-resolution forecast $\bar{u}$ and static high-resolution fields. The noise level σ_τ (and lead time, when applicable) is embedded and injected at all resolutions via adaptive scale–shift modulation. *Bottom:* Forward and reverse diffusion processes. The forward process progressively corrupts u into pure noise, while sampling integrates the reverse-time probability-flow dynamics to generate high-resolution samples from $p_\theta(u \mid \bar{u})$.

differential equation

$$\mathrm{d}u_\tau = -2\,\dot{\sigma}_\tau \sigma_\tau \nabla_{u_\tau} \log p_\tau(u_\tau \mid \bar{u})\,\mathrm{d}\tau + \sqrt{2\,\dot{\sigma}_\tau \sigma_\tau}\,\mathrm{d}\widehat{W}_\tau, \tag{7}$$

where $\widehat{W}_\tau$ denotes Brownian motion in reverse diffusion time.

While the reverse SDE is known, it is not directly tractable in practice because it depends on the conditional score $\nabla_{u_\tau} \log p_\tau(u_\tau \mid \bar{u})$, which is not available in closed form. Following standard score-based diffusion models, we approximate this score using a noise-conditional denoising operator $D_\theta(u_\tau, \bar{u}, \sigma_\tau)$, implemented as a neural network with trainable parameters θ. In particular, the score is approximated using Tweedie's formula as

$$\nabla_{u_\tau} \log p_\tau(u_\tau \mid \bar{u}) \approx \frac{D_\theta(u_\tau, \bar{u}, \sigma_\tau) - u_\tau}{\sigma_\tau^2}. \tag{8}$$

Accordingly, the denoiser is trained to predict the clean high-resolution state from its noisy counterpart by minimizing a mean-squared denoising objective over paired data and over a continuous range of noise levels,

$$\mathcal{L}(D_\theta) = \mathbb{E}_{\sigma \sim p_{\text{train}}} \mathbb{E}_{(\bar{u},u)\sim\mathcal{D}} \mathbb{E}_{\eta\sim\mathcal{N}(0,\sigma^2 I)} \left\| D_\theta(u+\eta, \bar{u}, \sigma) - u \right\|_2^2. \tag{9}$$

In practice, we follow the preconditioning and implementation choices introduced in the Elucidated Diffusion Models framework (Karras et al., 2022); full architectural and training details are provided in Appendix A.

5 Experimental Setup and Results

This section describes the experimental setup used to evaluate the proposed diffusion-based downscaling framework. We detail the training data, model configuration, and evaluation protocol, and assess probabilistic high-resolution forecasts against independent in situ station observations across multiple heterogeneous upstream forecasting models.

Training data. For training, we choose ERA5 reanalysis (Hersbach et al., 2020) fields as low-resolution inputs $\bar{u}$ ($\sim$25 km / 0.25° resolution) and CERRA reanalysis (Ridal et al., 2024) fields as the corresponding high-resolution targets u ($\sim$5 km / 0.05° resolution). ERA5 is well-known global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), combining a state-of-the-art numerical weather prediction model with a comprehensive set of in situ and satellite observations. CERRA is a high-resolution regional reanalysis for Europe and provides surface fields at approximately 5 km horizontal resolution, covering latitudes from 36° to 72° N and longitudes from 15° W to 45° E. Both ERA5 inputs and CERRA targets include 2 m temperature (t2m), eastward and northward 10 m wind components (u10, v10), and mean sea-level pressure (msl). Training is performed on data spanning the period 2014–2023. All input and output channels are standardized using ERA5 statistics computed over the period 1979–2021.

Denoiser training, architecture, and conditioning interface. Figure 2 provides an overview of the denoiser training, architecture, and sampling procedure.

The denoiser D_θ is implemented as a U-Net operating on 2D latitude–longitude fields, augmented with attention at the bottleneck to enable global spatial mixing. The denoiser takes as input the noisy high-resolution sample u_τ on the CERRA grid, the corresponding coarse-resolution ERA5 data $\bar{u}$, and static high-resolution fields $s = (z, \mathrm{lsm})$ consisting of elevation z and land–sea mask lsm. During training, $\bar{u}$ corresponds to ERA5, while at inference time it is provided by a forecasting model.

To account for systematic differences in effective resolution, spectral content, and implicit smoothing across upstream models (particularly over long forecast horizons, where autoregressive AI models are known to progressively smooth their predictions) we augment the conditioning input during training by applying randomized spectral low-pass filtering to the ERA5 fields (see Appendix A.1 for details).

The coarse-resolution atmospheric state is bilinearly upsampled to the CERRA grid and concatenated channel-wise with the noisy sample and static fields, yielding the network input

$$x_\tau = \mathrm{Concat}\big[u_\tau,\ \bar{u},\ s\big] \in \mathbb{R}^{H\times W\times C}.$$

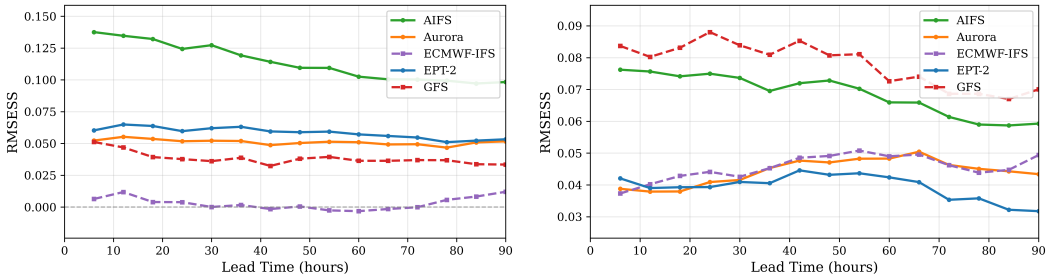
where $H = 721$, $W = 1201$ and $C = 10$. The diffusion noise level σ_τ is embedded using Fourier features and processed by a small multilayer perceptron to produce per-channel scale and shift parameters. These parameters modulate normalization layers within each residual block in a FiLM-style manner and are applied at all resolutions throughout the encoder and decoder (see Appendix A for further details on training, inference and denoiser architecture).

Evaluation protocol. The model is evaluated in a strictly out-of-sample setting spanning the period from 1 July 2024 to 30 June 2025. Forecast skill is assessed against roughly 10.000 independent in situ observations, focusing on 2 m temperature and 10 m wind speed. We use a curated subset of high-quality station networks, including WMO surface synoptic observations (SYNOP), WIS 2.0 surface observations, and aviation METAR reports (including ASOS/AWOS and international METAR feeds). Gridded forecasts are collocated to station locations using bilinear interpolation in latitude and longitude.

Unless stated otherwise, probabilistic forecasts are produced using ensembles of n=16 high-resolution samples. Point forecast skill is quantified using the root-mean-square error (RMSE) between observations y and the ensemble mean $\hat{y} = \frac{1}{n}\sum_{k=1}^{n} y^{(k)}$. Probabilistic forecast quality is quantified using the continuous ranked probability score (CRPS). In practice, for an ensemble forecast $\{y^{(k)}\}_{k=1}^{n}$ of size n, CRPS can be computed in closed form as

$$\mathrm{CRPS} = \frac{1}{n}\sum_{k=1}^{n}|y^{(k)} - y| - \frac{1}{2n^2}\sum_{k=1}^{n}\sum_{\ell=1}^{n}|y^{(k)} - y^{(\ell)}|. \tag{10}$$

The first term measures the average absolute error of the ensemble members with respect to the observation, while the second term accounts for the internal dispersion of the ensemble. For a



(a) 2 m Temperature

(b) 10 m Wind Speed

Figure 3: **RMSE Skill Score vs. lead time.** RMSESS ($1 - \mathrm{RMSE}_{\mathrm{downscaled}}/\mathrm{RMSE}_{\mathrm{base}}$) for all models (AI: Aurora, AIFS, EPT-2; NWP: GFS, ECMWF IFS) over July 2024–June 2025, up to 90 h lead time. Positive values indicate improvement over the raw baseline. The diffusion-based downscaler consistently improves point forecast skill across all upstream models.

deterministic forecast $\hat{y}$ CRPS reduces exactly to the absolute error (and thus averages to MAE over a dataset).

Upstream forecast models. We evaluate the proposed downscaling approach across a heterogeneous set of upstream forecasting systems spanning both data-driven AI models and physics-based numerical weather prediction (NWP) models: Microsoft Aurora (Bodnar et al., 2024), ECMWF AIFS (Lang et al., 2024), Jua EPT-2 (Molinaro et al., 2025b), NOAA GFS, and ECMWF IFS HRES. Aurora, AIFS, and EPT-2 are learned forecasting systems trained on historical reanalysis and analysis data, while GFS and IFS HRES are operational NWP models based on numerical integration of the governing equations. Since Aurora and AIFS produce forecasts at a 6-hourly temporal resolution only, all models are evaluated at a common 6-hourly cadence for consistency. This model set spans distinct spatial resolutions, training objectives, physical parameterizations, and data assimilation strategies, and therefore exhibits markedly different systematic error characteristics and biases.

Baselines and comparisons. For each upstream model, we compare (i) the raw deterministic forecast and (ii) the diffusion-downscaled forecast, and compute *skill scores of the downscaled model relative to the corresponding raw upstream forecast*,

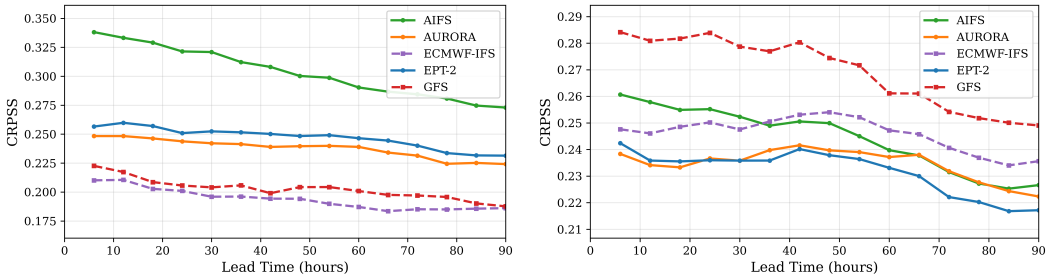
$$\mathrm{RMSESS} = 1 - \frac{\mathrm{RMSE}_{\mathrm{downscaled}}}{\mathrm{RMSE}_{\mathrm{base}}}, \qquad \mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\mathrm{downscaled}}}{\mathrm{CRPS}_{\mathrm{base}}},$$

where positive values indicate improvement over the baseline forecast.

Results. Figure 3 presents the root-mean-square error skill scores (RMSESS) of the diffusion-based downscaler relative to each upstream model's raw forecast. Across all evaluated systems, including both AI-based and NWP models, the downscaler yields predominantly positive RMSESS at nearly all lead times, indicating systematic improvements in point forecast accuracy. Although the magnitude of improvement varies by variable, model, and forecast lead time, positive skill is observed consistently from short-range to medium-range forecasts and, in most cases, extends beyond day 2.

For 2 m temperature (left panel of Figure 3), AI-based models exhibit the largest relative improvements. Temperature RMSESS typically ranges from approximately 5–14% at short lead times and gradually decreases to about 2–6% at longer lead times. For NWP models, improvements are generally smaller but remain positive, with the exception of ECMWF-IFS, for which RMSESS approaches zero beyond approximately 24 hours, indicating little to no systematic improvement at longer lead times.

For 10 m wind speed (right panel of Figure 3), relative improvements are larger and more persistent across lead times. AI-based models achieve wind RMSESS values of approximately 3–8% over the full 0–90 h forecast range. For NWP models, wind RMSESS remains consistently positive, typically ranging from about 4–9% over the 0–90 h lead times considered. Notably, RMSE improvements remain positive and not negligible even for strong operational NWP baselines such as ECMWF IFS which is currently the world's leading NWP model, demonstrating that the method adds value across a broad range of upstream forecast qualities.



(a) 2 m Temperature

(b) 10 m Wind Speed

Figure 4: **CRPS Skill Score vs. lead time.** CRPSS ($1 - \mathrm{CRPS}_{\mathrm{downscaled}}/\mathrm{CRPS}_{\mathrm{base}}$) for all models (AI: Aurora, AIFS, EPT-2; NWP: GFS, ECMWF IFS) over July 2024–June 2025, up to 90 h lead time. Positive values indicate improvement over the deterministic baseline. The diffusion-based downscaler improves probabilistic skill across all upstream models.

On the probabilistic side, Figure 4 reports continuous ranked probability skill scores (CRPSS), evaluating the quality of the full predictive distribution. Across all upstream models and lead times, diffusion-based downscaling yields consistently positive CRPSS. Moreover, the magnitude of probabilistic improvement is substantially larger than that observed for RMSE. In the short range, CRPSS frequently reaches values of 15–30% for both NWP and AI-based models, with positive skill persisting at longer lead times. The systematically larger gains in CRPSS compared to RMSESS indicate that the primary benefit of diffusion-based downscaling lies in improving the full conditional distribution rather than only the ensemble mean.

Overall, these results show that diffusion-based downscaling provides a robust, model-agnostic mechanism for lifting deterministic forecasts into high-resolution ensembles, improving both point skill (RMSE of the ensemble mean) and probabilistic skill (CRPS) in a strictly zero-shot setting across diverse upstream models, including a strong operational high-resolution baseline. Sample visualizations comparing CERRA reference fields against ensemble members generated by the diffusion downscaler are provided in Appendix A.5.

Training and sampling cost. The diffusion downscaler is computationally lightweight both to train and to deploy. The model is trained for 50 epochs on eight NVIDIA H100 GPUs, requiring approximately 8 hours in total. At inference time, generating 16 high-resolution samples at 5 km resolution over the full European domain from a single deterministic forecast step takes approximately 20 seconds on a single NVIDIA H100 GPU. Additionally, multiple diffusion samples can be generated fully in parallel across GPUs. As a result, probabilistic forecasts can be produced with only a modest increase in latency relative to the underlying deterministic model.

Ablation: diffusion vs. regression-based downscaling. Figure 5 isolates the effect of the diffusion training objective by comparing against a regression-based downscaler trained with an MSE loss, using the exact same U-Net architecture. Both learned downscalers improve RMSE relative to the raw upstream forecast for 2 m temperature and 10 m wind speed. For 10 m wind speed, the diffusion-based model achieves consistently lower RMSE than the MSE-based alternative across lead times, whereas for 2 m temperature the two approaches exhibit very similar RMSE. In contrast, pronounced differences emerge in probabilistic skill: the regression-based model yields only limited CRPS improvements, while the diffusion-based downscaler achieves substantially lower CRPS across all lead times and both variables. This demonstrates that the dominant gains in probabilistic performance stem from learning a conditional distribution via diffusion, rather than from architectural capacity alone.

6 Conclusion

We presented a diffusion-based formulation for probabilistic downscaling that acts as a *model-agnostic* resolution-lifting operator for deterministic weather forecasts. A single conditional EDM-style model is trained once on paired reanalysis fields (ERA5→CERRA over Europe) and subsequently applied

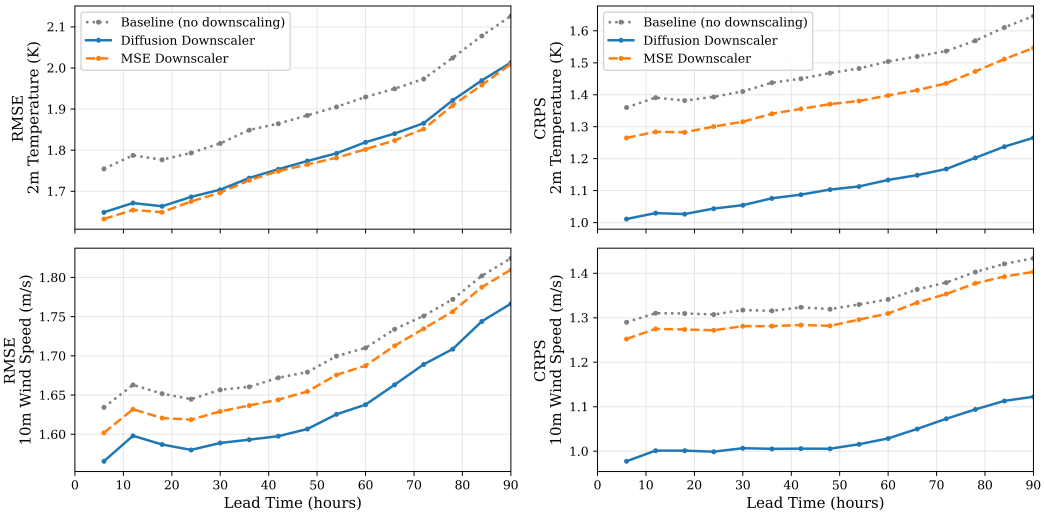


Figure 5: **Ablation: Diffusion vs. regression-based downscaling.** RMSE (left) and CRPS (right) for 2 m temperature (top) and 10 m wind speed (bottom) using EPT-2 as upstream model. Dashed: raw EPT-2 forecast. Dash-dot: regression-based (MSE) downscaler. Solid: diffusion-based downscaler. The diffusion model achieves comparable or better RMSE while substantially improving CRPS, demonstrating that gains arise from learning a conditional distribution rather than from architectural capacity alone.

zero-shot to heterogeneous upstream forecasting systems, spanning both AI-based and numerical weather prediction (NWP) models. In all cases, the same trained model is used without any upstream-specific fine-tuning, producing high-resolution ensembles on the CERRA grid.

In a station-verified, out-of-sample evaluation covering July 2024–June 2025 and lead times up to 90 h, diffusion-based downscaling consistently improves forecast skill relative to each model's own raw deterministic forecast for near-surface variables, including 2 m temperature and 10 m wind speed. These improvements are observed across upstream models with different numerical formulations, resolutions, and error characteristics, demonstrating that the learned downscaler generalizes robustly when used as a post-processing step.

While diffusion-based downscaling yields consistent improvements in point accuracy, the magnitude of these gains is modest compared to the substantial and robust improvements in probabilistic skill. This reflects the inherently ill-posed nature of statistical downscaling, where multiple fine-scale realizations are compatible with the same coarse atmospheric state. The primary benefit of the proposed approach lies not in sharper deterministic predictions, but in its ability to represent flow-dependent uncertainty and generate realistic high-resolution ensembles. By operating zero-shot across heterogeneous upstream forecasts, the method provides a model-agnostic and computationally efficient pathway for uncertainty-aware regional forecasting, producing probabilistic high-resolution fields without the need for costly regional dynamical simulations.

Limitations and future work. The present study focuses on a limited set of near-surface variables and on a single high-resolution training domain defined by the CERRA European reanalysis. While this setting is sufficient to demonstrate the feasibility and model-agnostic nature of diffusion-based downscaling, it does not fully probe the method's behavior under stronger forms of distribution shift, such as transfer to climatologically distinct regions.

From an operational perspective, we do not explicitly optimize ensemble size or inference cost, nor do we perform post hoc calibration of ensemble spread. While the reported CRPS improvements indicate enhanced probabilistic skill, future work will investigate explicit calibration and sharpness diagnostics and their dependence on sampling strategy.

Finally, although the proposed downscaler is intentionally model-agnostic, mild forms of adaptation to upstream forecast structure-such as weak physical constraints, lead-time-dependent conditioning,

lightweight per-model normalization, or limited specialization for regional forecasting systems (e.g., HRRR, ICON-EU, ICON-D2)-may further improve robustness under severe distribution shift. Exploring such extensions while preserving the core objective of a single, reusable probabilistic module remains an important direction for future research.

ACKNOWLEDGEMENTS

We gratefully acknowledge the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Oceanic and Atmospheric Administration (NOAA) for making high-quality reanalysis and forecast data publicly available, without which this work would not have been possible.

REFERENCES

Daniel Abdi, Isidora Jankov, Paul Madden, Vanderlei Vargas, Timothy A. Smith, Sergey Frolov, Montgomery Flora, and Corey Potvin. HRRRCast: A data-driven emulator for regional weather forecasting at convection allowing scales, July 2025. URL http://arxiv.org/abs/2507.05658.

Simon Adamov, Joel Oskarsson, Leif Denby, Tomas Landelius, Kasper Hintz, Simon Christiansen, Irene Schicker, Carlos Osuna, Fredrik Lindsten, Oliver Fuhrer, and Sebastian Schemm. Building Machine Learning Limited Area Models: Kilometer-Scale Weather Forecasting in Realistic Settings, April 2025. URL http://arxiv.org/abs/2504.09340.

R. E. Benestad. Empirical–statistical downscaling in climate modeling. *International Journal of Climatology*, 2008.

Tobias Bischoff and Katherine Deck. Unpaired Downscaling of Fluid Flows with Diffusion Bridges. *Artificial Intelligence for the Earth Systems*, March 2024. ISSN 2769-7525. doi: 10.1175/AIES-D-23-0039.1. URL https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-23-0039.1/AIES-D-23-0039.1.xml.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Vaughan, 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, November 2024. URL http://arxiv.org/abs/2405.13063.

Salva Rühling Cachay, Miika Aittala, Karsten Kreis, Noah Brenowitz, Arash Vahdat, Morteza Mardani, and Rose Yu. Elucidated Rolling Diffusion Models for Probabilistic Weather Forecasting, June 2025. URL http://arxiv.org/abs/2506.20024.

David C. Dowell, Curtis R. Alexander, Eric P. James, Stephen S. Weygandt, Stanley G. Benjamin, Geoffrey S. Manikin, Benjamin T. Blake, John M. Brown, Joseph B. Olson, Ming Hu, Tatiana G. Smirnova, Terra Ladwig, Jaymes S. Kenyon, Ravan Ahmadov, David D. Turner, Jeffrey D. Duda, and Trevor I. Alcott. The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description. *Weather and Forecasting*, 37(8):1371–1395, August 2022. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-21-0151.1. URL https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-21-0151.1.xml.

European Centre for Medium-Range Weather Forecasts (ECMWF). Ifs upgrade improves near-surface wind and temperature. ECMWF Newsletter, No. 181, 2024. URL https://www.ecmwf.int/en/newsletter/181/earth-system-science/ifs-upgrade-improves-near-surface-wind-and-temperature. Accessed 2026-01-21.

F. Giorgi and L. O. Mearns. Introduction to special section: Regional climate modeling revisited. *Journal of Geophysical Research*, 1999.

Luca Glawion, Julius Polz, Harald Kunstmann, Benjamin Fersch, and Christian Chwala. Global spatio-temporal ERA5 precipitation downscaling to km and sub-hourly scale using generative AI. *npj Climate and Atmospheric Science*, 8(1):219, June 2025. ISSN 2397-3722. doi: 10.1038/s41612-025-01103-y. URL https://www.nature.com/articles/s41612-025-01103-y.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient Foundation Models for PDEs, November 2024. URL http://arxiv.org/abs/2405.19101.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana

Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. ISSN 1477-870X. doi: 10.1002/qj.3803. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

James R. Holton and Gregory J. Hakim. *An Introduction to Dynamic Meteorology*. Academic Press, December 2012. ISBN 978-0-12-384867-3.

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.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, December 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/10.1126/science.adi2336.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. AIFS – ECMWF's data-driven forecasting system, August 2024. URL http://arxiv.org/abs/2406.01465.

Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields With a Generative Adversarial Network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, September 2021. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2020.3032790. URL https://ieeexplore.ieee.org/document/9246532/.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations, May 2021. URL http://arxiv.org/abs/2010.08895.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs] version: 3.

D. et al. Maraun. Precipitation downscaling under climate change. *Reviews of Geophysics*, 2010.

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, August 2024. URL http://arxiv.org/abs/2309.15214.

Fabio Merizzi, Andrea Asperti, and Stefano Colamonaco. Wind speed super-resolution and validation: From ERA5 to CERRA via diffusion models. *Neural Computing and Applications*, 36(34): 21899–21921, December 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-10139-9. URL https://doi.org/10.1007/s00521-024-10139-9.

Roberto Molinaro, Samuel Lanthaler, Bogdan Raonić, Tobias Rohner, Victor Armegioiu, Stephan Simonis, Dana Grund, Yannick Ramic, Zhong Yi Wan, Fei Sha, Siddhartha Mishra, and Leonardo Zepeda-Núñez. Generative ai for fast and accurate statistical computation of fluids, 2025a. URL https://arxiv.org/abs/2409.18359.

Roberto Molinaro, Niall Siegenheim, Niels Poulsen, Jordan Dane Daubinet, Henry Martin, Mark Frey, Kevin Thiart, Alexander Jakob Dautel, Andreas Schlueter, Alex Grigoryev, Bogdan Danciu, Nikoo Ekhtiari, Bas Steunebrink, Leonie Wagner, and Marvin Vincent Gabler. Ept-2 technical report. arXiv preprint, 2025b. URL https://arxiv.org/abs/2507.09703.

Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Graph-based Neural Weather Prediction for Limited Area Modeling, November 2023. URL http://arxiv.org/abs/2309.17370.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, February 2022. URL http://arxiv.org/abs/2202.11214.

Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, and Michael Pritchard. Kilometer-Scale Convection Allowing Model Emulation using Generative Diffusion Modeling, August 2024. URL http://arxiv.org/abs/2408.10958.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08252-9. URL https://www.nature.com/articles/s41586-024-08252-9.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional Neural Operators for robust and accurate learning of PDEs, December 2023. URL http://arxiv.org/abs/2302.01178.

Martin Ridal, Eric Bazile, Patrick Le Moigne, Roger Randriamampianina, Semjon Schimanke, Ulf Andrae, Lars Berggren, Pierre Brousseau, Per Dahlgren, Lisette Edvinsson, Adam El-Said, Michael Glinton, Susanna Hagelin, Susanna Hopsch, Ludvig Isaksson, Paulo Medeiros, Esbjörn Olsson, Per Unden, and Zheng Qi Wang. CERRA, the Copernicus European Regional Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, 150:3385–3411, July 2024. ISSN 0035-9009. doi: 10.1002/qj.4764. URL https://ui.adsabs.harvard.edu/abs/2024QJRMS.150.3385R.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022. URL http://arxiv.org/abs/2112.10752.

Maybritt Schillinger, Maxim Samarin, Xinwei Shen, Reto Knutti, and Nicolai Meinshausen. EnScale: Temporally-consistent multivariate generative downscaling via proper scoring rules, December 2025. URL http://arxiv.org/abs/2509.26258.

Tobias Selz, Wessel Bruinsma, George C. Craig, Stratis Markou, Richard Turner, and Anna Vaughan. On the effective resolution of AI weather prediction models, March 2025. URL https://essopenarchive.org/users/645836/articles/1274105-on-the-effective-resolution-of-ai-weather-prediction-models?commit=02cef8b30e0b88f4240701b89e3271c9f734342c.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

Elena Tomasi, Gabriele Franch, and Marco Cristoforetti. Can AI be enabled to dynamical downscaling? A Latent Diffusion Model to mimic km-scale COSMO5.0_CLM9 simulations. *Geoscientific Model Development*, 18(6):2051–2078, April 2025. ISSN 1991-9603. doi: 10.5194/gmd-18-2051-2025. URL http://arxiv.org/abs/2406.13627.

Zhong Yi Wan, Ricardo Baptista, Yi-fan Chen, John Anderson, Anudhyan Boral, Fei Sha, and Leonardo Zepeda-Núñez. Debias Coarsely, Sample Conditionally: Statistical Downscaling through Optimal Transport and Probabilistic Diffusion Models, October 2023. URL http://arxiv.org/abs/2305.15618.

Zhong Yi Wan, Ignacio Lopez-Gomez, Robert Carver, Tapio Schneider, John Anderson, Fei Sha, and Leonardo Zepeda-Núñez. Regional climate risk assessment from climate models using probabilistic machine learning, June 2025. URL http://arxiv.org/abs/2412.08079.

R. L. Wilby, T. M. L. Wigley, and D. Conway. Statistical downscaling of general circulation model output: A comparison of methods. *Journal of Climate*, 11(4):530–548, 1998a.

R. L. Wilby, T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, 34(11):2995–3008, 1998b. ISSN 1944-7973. doi: 10.1029/98WR02577. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/98WR02577.

R.L. Wilby and T.M.L. Wigley. Downscaling general circulation model output: A review of methods and limitations. *Progress in Physical Geography: Earth and Environment*, 21(4): 530–548, December 1997. ISSN 0309-1333. doi: 10.1177/030913339702100403. URL https://doi.org/10.1177/030913339702100403.

Günther Zängl, Daniel Reinert, Pilar Rípodas, and Michael Baldauf. The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, 141(687):563–579, 2015. ISSN 1477-870X. doi: 10.1002/qj.2378. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.2378.

A Supplementary Information

A.1 Spectral smoothing of the conditioning input

In this section, we further elaborate on the ERA5 input spectral smoothing mechanism introduced in the main text.

Specifically, the conditioning input $\bar{u}$ is obtained by Fourier-domain smoothing of the original ERA5 fields, with the smoothing strength sampled uniformly from the interval $[0,\ 0.8]$, where zero corresponds to no smoothing. This procedure exposes the denoiser to a continuum of effective input resolutions during training and improves robustness to heterogeneous upstream models at inference time.

Let $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the two-dimensional discrete Fourier transform and its inverse. We define

$$\bar{u} = \mathcal{F}^{-1}\left[\mathcal{F}(\bar{u}_0)\ \cdot\ G_\alpha\right], \tag{11}$$

where $\cdot$ denotes element-wise multiplication and G_α is an isotropic Gaussian low-pass filter defined in frequency space.

Specifically, letting (k_x, k_y) denote discrete spatial frequencies centered at zero, the filter takes the form

$$G_\alpha(k_x, k_y) = \exp\left(-\frac{k_x^2 + k_y^2}{2\sigma_\alpha^2}\right), \tag{12}$$

with bandwidth

$$\sigma_\alpha = \frac{1}{2}\max\left(1,\ \frac{\min(H, W)}{2}\,(1-\alpha)\right). \tag{13}$$

The smoothing strength $\alpha \in [0,\ 0.8]$ is sampled uniformly for each training instance. The limiting case $\alpha = 0$ corresponds to the identity operator (no smoothing), while increasing values of α progressively suppress high-frequency components, yielding conditioning inputs with coarser effective resolution.

A.2 Denoiser Training

The parameters θ of the conditional denoiser D_θ are learned from paired low- and high-resolution reanalysis data. Training samples are constructed from time-aligned fields, where $\bar{u}(t)$ denotes a coarse-resolution atmospheric state from ERA5 and $u(t)$ the corresponding high-resolution target from CERRA at the same valid time t. Paired samples $(\bar{u}, u)$ are obtained by matching timestamps and sampled from long reanalysis time series; we denote by $\mathcal{D}$ the resulting empirical distribution of paired fields.

For a given noise level $\sigma > 0$ and noise realization $\eta \sim \mathcal{N}(0, \sigma^2 I)$, a noisy high-resolution field $u + \eta$ is constructed. The conditional denoiser $D_\theta(\cdot, \bar{u}, \sigma)$ is then trained to reconstruct the clean target u by minimizing a denoising objective. Specifically, the denoiser parameters are obtained as (local) minimizers of

$$\mathcal{L}(D_\theta, \sigma) = \mathbb{E}_{(\bar{u},u)\sim\mathcal{D}} \mathbb{E}_{\eta\sim\mathcal{N}(0,\sigma^2 I)} \left\| D_\theta(u + \eta, \bar{u}, \sigma) - u \right\|_2^2. \tag{14}$$

The full training objective averages this loss over a distribution of noise levels,

$$\mathcal{L}(D_\theta) = \mathbb{E}_{\sigma\sim p_{\text{train}}} \left[\lambda(\sigma)\, \mathcal{L}(D_\theta, \sigma) \right], \tag{15}$$

where p_{train} denotes the noise-level sampling distribution and $\lambda(\sigma)$ a noise-dependent weighting factor.

Preconditioning and effective weighting. For all experiments, we adopt the preconditioning and weighting scheme of Elucidated Diffusion Models (EDM). The denoiser is parameterized as

$$D_\theta(u + \eta, \bar{u}, \sigma) = c_{\text{skip}}(\sigma)(u + \eta) + c_{\text{out}}(\sigma)\, F_\theta(c_{\text{in}}(\sigma)(u + \eta),\, \bar{u};\, c_{\text{noise}}(\sigma)), \tag{16}$$

where F_θ denotes the raw neural network.

The preconditioning coefficients are given by

$$c_{\text{skip}}(\sigma) = \frac{\sigma_{\text{data}}^2}{\sigma^2 + \sigma_{\text{data}}^2}, \tag{17}$$

$$c_{\text{out}}(\sigma) = \frac{\sigma\, \sigma_{\text{data}}}{\sqrt{\sigma^2 + \sigma_{\text{data}}^2}}, \tag{18}$$

$$c_{\text{in}}(\sigma) = \frac{1}{\sqrt{\sigma^2 + \sigma_{\text{data}}^2}}, \tag{19}$$

$$c_{\text{noise}}(\sigma) = \tfrac{1}{4} \log(\sigma), \tag{20}$$

with $\sigma_{\text{data}} = 0.5$.

The corresponding regression target is

$$F_{\text{target}} = \frac{1}{c_{\text{out}}(\sigma)} \left(u - c_{\text{skip}}(\sigma)(u + \eta) \right), \tag{21}$$

and the resulting training loss can be written as

$$\mathcal{L} = \mathbb{E}_{\sigma,(\bar{u},u),\eta} \left[w(\sigma) \left\| F_\theta(c_{\text{in}}(\sigma)(u + \eta),\, \bar{u};\, c_{\text{noise}}(\sigma)) - F_{\text{target}} \right\|_2^2 \right], \tag{22}$$

with effective weighting

$$w(\sigma) = \frac{\sigma^2 + \sigma_{\text{data}}^2}{(\sigma\, \sigma_{\text{data}})^2}. \tag{23}$$

Noise levels are sampled during training according to a log-normal distribution,

$$\log \sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2), \tag{24}$$

with $P_{\text{mean}} = -0.5$ and $P_{\text{std}} = 1.5$ in all experiments.

A.2.1 Inference and sample generation

At inference time, samples from the learned conditional distribution $p_\theta(u \mid \bar{u})$ are generated by reversing the diffusion process. Starting from the conditional reverse-time stochastic differential equation Eq.(7), we consider its associated *deterministic* probability-flow formulation:

$$\mathrm{d}x = \left[\frac{\dot{\sigma}(\tau)}{\sigma(\tau)} + \frac{\dot{s}(\tau)}{s(\tau)} \right] x\, \mathrm{d}\tau - \frac{s(\tau)\, \dot{\sigma}(\tau)}{\sigma(\tau)} D_\theta\left(\frac{x}{s(\tau)},\, \bar{u},\, \sigma(\tau) \right) \mathrm{d}\tau, \tag{25}$$

where the score function is approximated by the trained conditional denoiser D_θ. Both the stochastic formulation in Eq. (7) and the deterministic probability-flow ODE in Eq. (25) induce the same family of conditional distributions $p_\tau(u \mid \bar{u})$.

Following the Elucidated Diffusion Models (EDM) framework (Karras et al., 2022), we adopt a variance-exploding diffusion with identity signal scaling,

$$s(\tau) = 1, \qquad \sigma(\tau) = \tau, \tag{26}$$

so that diffusion time directly parameterizes the noise level and Eq. (25) simplifies accordingly.

In practice, inference is performed by discretizing the probability-flow ODE over a monotonically decreasing sequence of noise levels $\{\sigma_i\}_{i=1}^{N} \subset [\sigma_{\max}, \sigma_{\min}]$. The discrete noise levels are constructed using the EDM power-law schedule,

$$\sigma_i = \left(\sigma_{\max}^{1/\rho} + \frac{i-1}{N-1}\left(\sigma_{\min}^{1/\rho} - \sigma_{\max}^{1/\rho}\right)\right)^{\rho}, \qquad i = 1, \ldots, N, \tag{27}$$

with $N = 128$, $\sigma_{\max} = 80$, and $\sigma_{\min} = 0.002$.

The resulting ODE is integrated using a first-order explicit Euler method. Integration is initialized from

$$x_{\sigma_{\max}} \sim \mathcal{N}(0, \sigma_{\max}^2 I), \tag{28}$$

and proceeds until $\sigma_{\min}$, at which point the final state is taken as a sample from the conditional distribution $p_\theta(u \mid \bar{u})$. Multiple independent realizations are obtained by repeating the integration with different initial noise samples.

A.3 Denoiser Architecture and Configuration

The denoiser D_θ is implemented as a 2D encoder–decoder U-Net operating on latitude–longitude fields on the CERRA grid (Figure 2). At each forecast lead time, the network takes as input the concatenation $x_\tau = \mathrm{Concat}[u_\tau, \bar{u}, s]$, where $u_\tau \in \mathbb{R}^{H\times W\times 4}$ is the noisy high-resolution sample (t2m, u10, v10, msl), $\bar{u} \in \mathbb{R}^{H\times W\times 4}$ is the bilinearly upsampled coarse forecast, and $s \in \mathbb{R}^{H\times W\times 2}$ contains static fields (elevation and land–sea mask). Hence $x_\tau \in \mathbb{R}^{H\times W\times 10}$.

Encoder–decoder layout. The network begins with an input projection (3×3 convolution) to C_0=32 channels. It then applies L=4 resolution stages with channel multipliers $[2, 4, 8, 8]$, resulting in channel widths

$$C_0 \to 2C_0 \to 4C_0 \to 8C_0 \to 8C_0 \quad \text{i.e.} \quad 32 \to 64 \to 128 \to 256 \to 256.$$

Downsampling between stages is performed by a strided convolution (kernel 3, stride 2), halving spatial resolution at each stage. The decoder mirrors the encoder with symmetric upsampling stages. Upsampling is performed via bilinear interpolation by a factor of 2, followed by a 3×3 convolution.

Residual blocks and skip connections. Each resolution stage contains n_{res}=2 residual blocks (Figure 2, blue "T" blocks), and standard U-Net skip connections connect encoder and decoder features at matching resolutions. Within each residual block, we use GroupNorm–activation–Conv twice (GN–f–C, GN–f–C), with dropout set to 0. Unless stated otherwise, GroupNorm uses 32 groups.

Conditioning interface. Diffusion noise conditioning is injected into every residual block through adaptive scale–shift modulation. We embed σ using Fourier features with embedding dimension d_σ=256, followed by a two-layer MLP that outputs per-channel affine parameters (a, b). These parameters modulate a normalization layer in FiLM form,

$$\mathrm{GN}(x) \;\mapsto\; (1 + a)\,\mathrm{GN}(x) + b, \tag{29}$$

where (a, b) are broadcast over spatial dimensions.

Attention blocks. The denoiser follows a standard U-Net encoder–decoder structure with residual blocks at each resolution and a single self-attention block at the bottleneck (see Figure 2). Self-attention is not used at intermediate encoder or decoder resolutions; all attention-based global mixing is confined to the bottleneck representation.

Configuration (for reproducibility). For clarity, we summarize the key architectural hyperparameters used throughout:

- Input channels: 10 (noisy high-resolution sample u_τ: 4; coarse-resolution forecast provided as conditioning: 6).
- Output channels: 4 (2 m temperature, 10 m wind components, and mean sea-level pressure on the CERRA grid).
- Base channels: $C_0 = 32$, with channel multipliers $[2, 4, 8, 8]$, corresponding to feature widths $32 \to 64 \to 128 \to 256 \to 256$.
- Residual blocks per resolution: $n_{\text{res}} = 2$; dropout $= 0$.
- Noise embedding dimension: 256, used to generate adaptive scale and shift parameters for conditioning normalization layers.

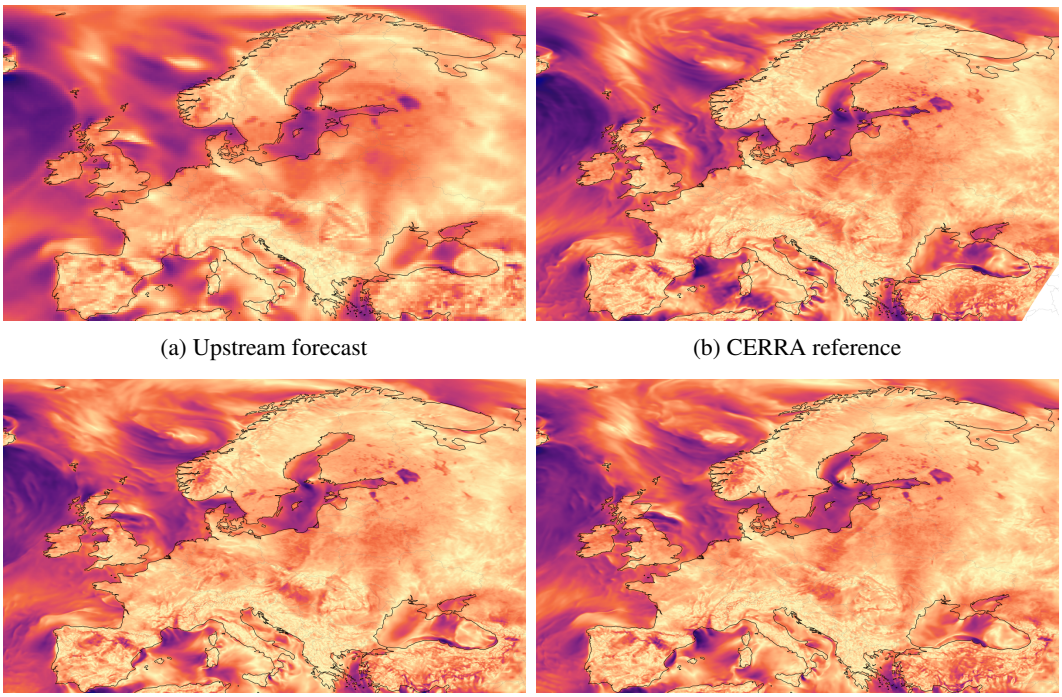
Overall, the model counts a total number of 24M parameters.

A.4 Denoiser Training Details

During training, we used paired samples from ERA5 and CERRA at three-hourly resolution, spanning the years 2014–2023, resulting in 26280 samples. We used the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 10^{-4}, a weight decay of 10^{-5}, and cosine anneal the learning rate to 10^{-5}. The batch size was set to 1.

A.5 Sample Visualizations

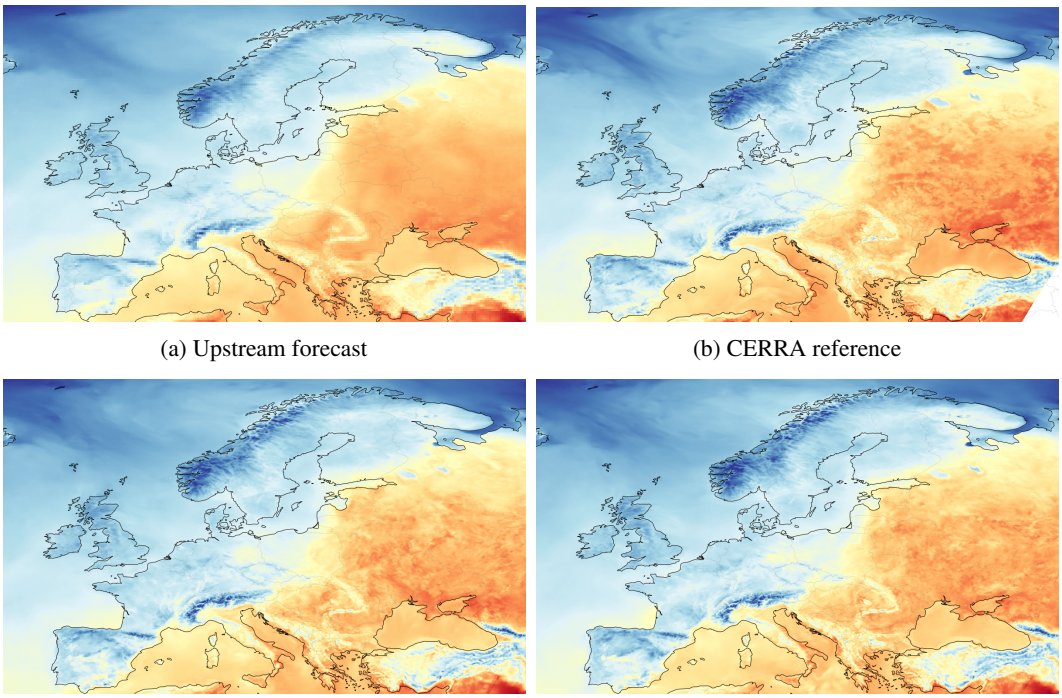
Figures 6 and 7 show example downscaled fields for 10 m wind speed and 2 m temperature at a 6-hour lead time, comparing the coarse upstream forecast, the CERRA reference, and two independent ensemble members generated by our diffusion model.



(a) Upstream forecast (b) CERRA reference

(c) Ensemble member 1 (d) Ensemble member 2

Figure 6: 10 m wind speed for 1 July 2024 at 6-hour lead time. (a) Coarse upstream forecast, (b) CERRA reference, (c–d) two independent ensemble members from the diffusion downscaler.



(a) Upstream forecast

(b) CERRA reference

(c) Ensemble member 1

(d) Ensemble member 2

Figure 7: 2 m temperature for 1 July 2024 at 6-hour lead time. (a) Coarse upstream forecast, (b) CERRA reference, (c–d) two independent ensemble members from the diffusion downscaler.