Assessing the Geographic Generalization and Physical Consistency of Generative Models for Climate Downscaling

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Kilometer-scale weather data is crucial for real-world applications but remains computationally intensive to produce using traditional weather simulations. An emerging solution is to use deep learning models, which offer a faster alternative for climate downscaling. However, their reliability is still in question, as they are often evaluated using standard machine learning metrics rather than insights from atmospheric and weather physics. This paper benchmarks recent state-of-the-art deep learning models and introduces physics-inspired diagnostics to evaluate their performance and reliability, with a particular focus on geographic generalization and physical consistency. Our experiments show that, despite the seemingly strong performance of models such as CorrDiff, when trained on a limited set of European geographies (e.g., central Europe), they struggle to generalize to other regions such as Iberia, Morocco in the south, or Scandinavia in the north. They also fail to accurately capture second-order variables such as divergence and vorticity derived from predicted velocity fields. These deficiencies appear even in in-distribution geographies, indicating challenges in producing physically consistent predictions. We propose a simple initial solution: introducing a power spectral density loss function that empirically improves geographic generalization by encouraging the reconstruction of small-scale physical structures.

1 Introduction

Weather forecasts at kilometer-scale resolution are essential for many applications, including energy system planning [34], agriculture [16], and natural disaster preparedness [30]. However, global reanalysis datasets like ERA5 [21], commonly used to train models of weather dynamics, are available only at coarse spatial resolutions (e.g., 25 km for ERA5 [21]). Consequently, global machine learning (ML)-based weather forecasting systems trained on these datasets [4, 26, 33, 31, 3] are inherently limited to the same coarse scale. This prevents them from capturing fine-scale dynamics and physical processes vital for accurate local impact assessments, leading to systematic discrepancies from real-world measurements [27].

Traditionally, dynamical downscaling is widely used to address the aforementioned limitation by deriving high-resolution information from coarse global forecasts [36]. This is achieved by numerically integrating the governing equations on a nested, limited-area grid, allowing simulations to capture small-scale physical processes. Still, computational cost grows steeply with domain size and simulation length, making this approach infeasible for large scale and detailed runs [2]. Recently, machine learning has been proposed as a faster alternative: statistical downscaling (often called super-resolution in computer vision [18]) learns a data-driven coarse-to-fine mapping. Once trained, the models can produce kilometer-scale fields several orders of magnitude faster than traditional dynamical downscaling [42]. However, most existing deep learning approaches neither verify whether

Submitted to the AI for Science workshop (NeurIPS 2025).

their outputs are physically consistent nor test whether the learned mapping generalizes to climates and geographic terrains unseen during training. This represents a literature gap as high performance on in-distribution results may not fully reflect their reliability in other conditions.

In this work, we critically evaluate state-of-the-art deep-learning downscalers on physical consistency and generalization and show that physics-informed loss can improve both. Our contributions can be summarized as follows:

1. We assess generalization performance by benchmarking three state-of-the-art deep-learning downscalers trained on a Central-European subset and tested on two held-out regions, the Iberia, Morocco region and Northern Scandinavia. We assess geographic transfer using standard machine-learning metrics and observe that the considered models significantly underperform in out-of-distribution (OOD) geographic domains.
2. Guided by diagnostics from atmospheric-physics literature, we assess physical consistency by examining power spectra of the following second order derived variables: horizontal kinetic energy, mass-continuity/divergence, and relative vorticity. This broadens evaluation beyond conventional ML scores and provides novel insights from atmospheric physics, revealing a lack of physical consistency in the predictions.
3. We introduce a Power spectral density (PSD) loss that explicitly rewards the reconstruction of small-scale physical dynamics, acts as a frequency-domain regularizer, and alleviates both OOD generalization gaps and physics-consistency shortcomings.

2 Related work

Generalization Statistical downscaling focuses on learning the conditional distribution $p(\mathbf{q}|\mathbf{x})$ from a set of paired data $\{(\mathbf{x}_i, \mathbf{q}_i)\}_{i=1}^{N}$, where $\mathbf{x} \in \mathbb{R}^{C\times h\times w}$ and $\mathbf{q} \in \mathbb{R}^{c\times H\times W}$ represent the input and target, respectively. Here, N denotes the number of paired samples, C is the number of meteorological variables given as input, and c is the number of meteorological variables to be downscaled. Also, (h, w) and (H, W) denote the height and width of the coarse and fine grids, respectively. Global physics-based simulations are computationally expensive, so any statistical downscaling scheme is useful only if it can be applied beyond the few regions where high-quality training data exist [15]. This scarcity demands models that generalize across disparate climates and topographies; otherwise, distribution shifts between the training and target domains can sharply degrade skill and even produce physically implausible outputs. Nevertheless, most prior studies still train and evaluate on the same geography, often limited to a single country [28], leaving the cross-climate robustness of ML downscalers largely unexplored.

Physical consistency Assessing physical consistency is crucial as it is common to diagnose (i.e., derive) one set of variables from the set of prognosed (i.e., directly estimated) variables based on known physical relationships [5]. An example is vertical wind speed, which is linked to horizontal wind speed through the fundamental principle of continuity. However, as documented in previous studies [39, 5, 17], physical consistency is not always given.

In atmospheric science, developing physics-informed approaches, such as incorporating full partial differential equations (PDEs) like the Navier–Stokes equations directly into the loss function as done by Raissi et al. [35], can be difficult. In practice, production-grade weather models like the Weather Research and Forecasting Model [40] solve not only the momentum equations (Navier–Stokes), but also the conservation equations for heat and moisture, all in three dimensions and over time. Consequently, a faithful PDE constraint in atmospheric physics would have to enforce this full, coupled 3D, time-dependent system. By contrast, the PDE-based constraints considered in the literature are often limited to idealized 2D flow cases (e.g., [35, 25, 11]). Another complication is that these equations are typically solved on terrain-following native grids, whereas widely used reanalysis products such as ERA5 are provided on pressure levels after interpolation from the native model grid. This transformation can disrupt physical balances, making direct application of atmospheric PDEs as constraints unreliable if the interpolation effects are ignored. Furthermore, atmospheric flows are influenced by many processes beyond the resolved dynamics, including orography, surface properties (e.g., roughness, heat capacity), solar radiation, cloud formation, and precipitation. Many of these are parameterized in numerical models and enter the PDE solver as forcings or as auxiliary equations coupled to the governing dynamics. Applying the full set of atmospheric PDEs with all relevant forcings on the correct native grid would therefore require substantial additional development. A practical compromise is to incorporate physical principles into the loss function design, as we explore in this paper.

Spectral loss Many statistical downscaling studies rely on deterministic mappings [19, 24, 10, 22], where the task reduces to estimating the conditional mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{q}|\mathbf{x}]$. Deterministic estimates of μ are obtained by minimizing a standard pixel-space loss (ℓ_1 or ℓ_2 loss), which often produce blurry predictions [6]. Recently, attention has shifted to probabilistic approaches [29, 13, 28] that aim to learn the full conditional distribution $p(\mathbf{q}|\mathbf{x})$, inspired by the remarkable success of diffusion models for super-resolution tasks in computer vision. The probabilistic nature of these models allows them to quantify uncertainty and achieve lower pixel-space errors (e.g., MAE) [29, 13, 28]. However, whether these predictions are also physically consistent remains an open question.

Previous work has shown that incorporating Fourier-based losses helps restore high-frequency details in efficient super-resolution models [14, 43]. Our PSD-based loss motivation is rooted in the physics of weather and climate data: large-scale structures (e.g., pressure systems) and fine-scale details (e.g., local wind shifts, sharp temperature gradients) coexist, and the PSD quantifies how much variance is present at each spatial scale. If a downscaling model reproduces the large-scale features but underestimates smaller ones, its PSD will decay too rapidly at high frequencies, indicating missing small-scale physics. By incorporating a PSD-based loss, we explicitly encourage the model to match the observed scale-by-scale variance distribution, aiming to improve physical realism and mitigate both generalization and physics-consistency shortcomings.

3 Methods

The next subsections present physical consistency checks comparing key variables like kinetic energy, divergence, and vorticity, followed by a spectral loss term based on PSD to improve fine-scale detail. Together, they evaluate the physical realism and spatial accuracy of the model predictions.

3.1 Physical consistency checks

In this work, to evaluate the physical consistency of the model predictions, we employ variable interdependencies known from fluid mechanics and meteorology, similar to [5]. More specifically, we compare each ground truth sample and its corresponding model prediction with respect to the following variables:

1. Mean horizontal kinetic energy, defined as

$$\mathbf{E}_h = \frac{\mathbf{u}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{v}}{2}. \tag{1}$$

2. Horizontal continuity/divergence, given by

$$\boldsymbol{\delta}_h = \nabla_h \cdot \mathbf{v}_h = \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial y}. \tag{2}$$

3. Relative horizontal vorticity, expressed as

$$\boldsymbol{\zeta}_h = \nabla_h \times \mathbf{v}_h = \frac{\partial \mathbf{v}}{\partial x} - \frac{\partial \mathbf{u}}{\partial y}. \tag{3}$$

Here, $\mathbf{u}$ and $\mathbf{v}$ are the mean horizontal wind components in the zonal (x) and meridional (y) directions, respectively. Since statistical downscaling does not consider the temporal evolution of a meteorological state, more complex dynamical physics-consistency checks (see, e.g., [39, 17]) are not applicable.

Physically and mathematically, the three variables allow checking for different constraints that a realistic and physically consistent wind field must satisfy. Comparing $\mathbf{E}_h$ between ground truth and prediction, for example, assesses whether the nonlinear relationship between $\mathbf{u}$ and $\mathbf{v}$ is preserved. Checking if divergence $\boldsymbol{\delta}_h$ and vorticity $\boldsymbol{\zeta}_h$ match is crucial, as a realistic differentiable wind vector field can be decomposed into a divergence-free and a vorticity-free component via the Helmholtz decomposition [9, 5]. Mathematically, the gradients in $\boldsymbol{\delta}_h$ and $\boldsymbol{\zeta}_h$ allow us to test whether the models capture the differential relationships between their outputs $\mathbf{u}$ and $\mathbf{v}$.

We note that, strictly speaking, all gradients should be horizontal gradients, which is not satisfied in complex terrain when $\mathbf{u}$ and $\mathbf{v}$ are at $10\,\mathrm{m}$ height above the terrain (see Table 1). Consequently, numerical values in mountainous regions are overestimated [23], but the metrics remain useful for broadly assessing physical consistency.

3.2 PSD Loss Term

To encourage fidelity at high spatial frequencies, we include a PSD penalty alongside the pre-defined loss of the considered baselines: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{baseline}} + \lambda\, \mathcal{L}_{\text{PSD}}$. The specific baselines are described in Section 4.1.

To elaborate, let $\mathbf{q} \in \mathbb{R}^{H \times W}$ denote a single-channel atmospheric variable field on a grid of height H and width W (here $\mathbb{R}^{H \times W}$ denotes the set of $H \times W$ real matrices). For multi-channel data (e.g., multiple atmospheric variables), the PSD is computed independently for each channel. The PSD is obtained from the squared magnitude of the Fourier coefficients, normalized by the spatial dimensions and the grid spacing Δx:

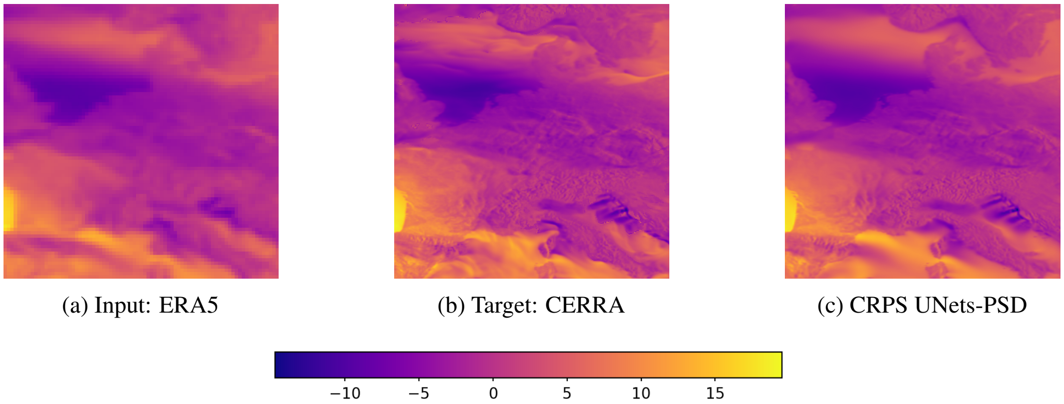
$$\mathrm{PSD}(\mathbf{q})(k_h, k_w) = \frac{|\mathcal{F}(\mathbf{q})(k_h, k_w)|^2}{H\,W\,\Delta x}\,, \tag{4}$$

where $k_h = 0, \ldots, H-1$ and $k_w = 0, \ldots, W-1$ are the discrete wavenumbers in the vertical and horizontal directions, respectively, and $\mathcal{F}(\mathbf{q})$ is the 2D discrete Fourier transform.

To define $\mathcal{L}_{\text{PSD}}$, we weigh the contribution of each PSD coefficient in proportion to its normalized isotropic wavenumber squared. The isotropic wavenumber k is obtained from k_h and k_w via $k = \sqrt{k_h^2 + k_w^2}$, and the weights are computed as $w(k_h, k_w) = (k/k_{\max})^2$. This quadratic weighting emphasizes high-frequency components, where the largest discrepancies occur, and therefore guides the model to reduce over-smoothing at small spatial scales.

The weighted PSD loss between ground truth $\mathbf{q}$ and prediction $\hat{\mathbf{q}}$ is then calculated as:

$$\mathcal{L}_{\text{PSD}} = \sqrt{\frac{1}{H\,W} \sum_{k_h=0}^{H-1} \sum_{k_w=0}^{W-1} w(k_h, k_w)\ \left[\log\big(\mathrm{PSD}(\mathbf{q})(k_h, k_w)\big) - \log\big(\mathrm{PSD}(\hat{\mathbf{q}})(k_h, k_w)\big)\right]^2}. \tag{5}$$


(a) Input: ERA5

(b) Target: CERRA

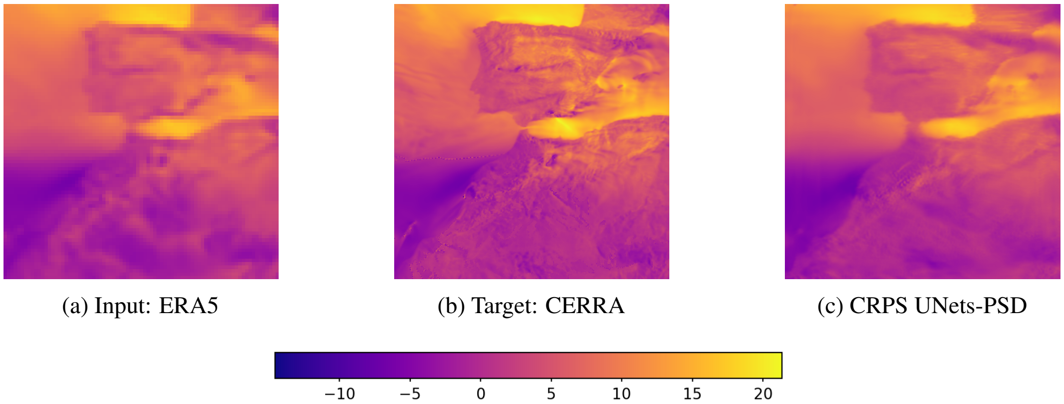
(c) CRPS UNets-PSD

Figure 1: Input, target, and CRPS UNets-PSD prediction of $\mathbf{u}$ in Central Europe (in-distribution region) on 1st of January 2021

4 Experimental results

4.1 Experiments setup

We benchmark three state-of-the-art downscalers: the full probabilistic *CorrDiff* model of Mardani et al. [28], its deterministic regression backbone (*Regression-CorrDiff*), and a probabilistic U-Net


(a) Input: ERA5

(b) Target: CERRA

(c) CRPS UNets-PSD

Figure 2: Input, target, and CRPS UNets-PSD prediction of $\mathbf{u}$ in the Iberia, Morocco region (OOD) on 1st of January 2021. The prediction is much blurrier compared to in-distribution as shown Figure 1.

ensemble trained with a continuous ranked probability score (CRPS) loss (*CRPS-UNets*) following Alet et al. [1]. Each architecture, except for *CorrDiff*, is retrained from scratch in two variants, one using its standard loss and one augmented with the extra PSD loss term. All models take as input the set of coarse-resolution variables listed in Table 1, taken from the ERA5 reanalysis at 25 km resolution [21] on multiple pressure levels. Their task is to predict the corresponding high-resolution surface fields from the CERRA reanalysis at 5.5 km resolution [37]: 10 m U-component of wind ($\mathbf{u}$), 10 m V-component of wind ($\mathbf{v}$), and 2 m temperature ($\mathbf{t2m}$). Training is performed on an ERA5–CERRA subset cropped to Central Europe. To test geographic generalization, we create two additional, non-overlapping ERA5–CERRA subsets, one over the Iberia, Morocco region and one over Northern Scandinavia. They are used only for OOD evaluation and not for training or fine-tuning. We assess physical consistency as indicated in Section 3.1. Further information on the datasets and training details can be found in Appendix B.

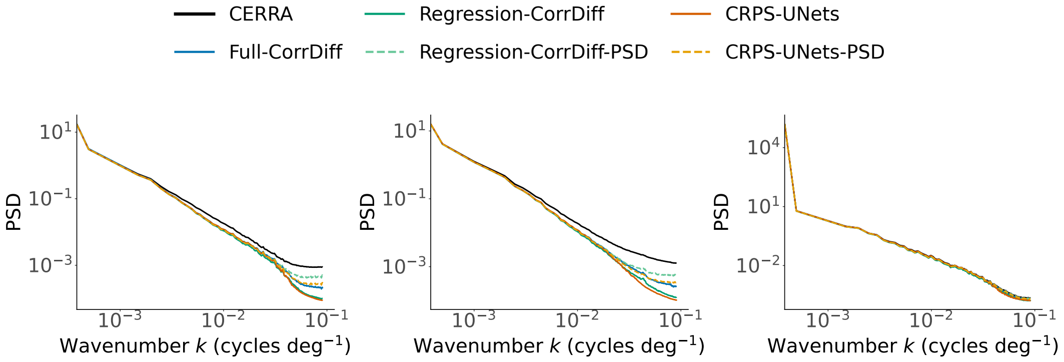
Table 1: Meteorological variables used as inputs and outputs for the downscaling models.

Input variables	**Output variables**
Pressure levels (500 hPa and 850 hPa)	
Temperature	
U-wind component	
V-wind component	
Surface level	
10 m U-component of wind	10 m U-component of wind
10 m V-component of wind	10 m V-component of wind
2 m Temperature	2 m Temperature

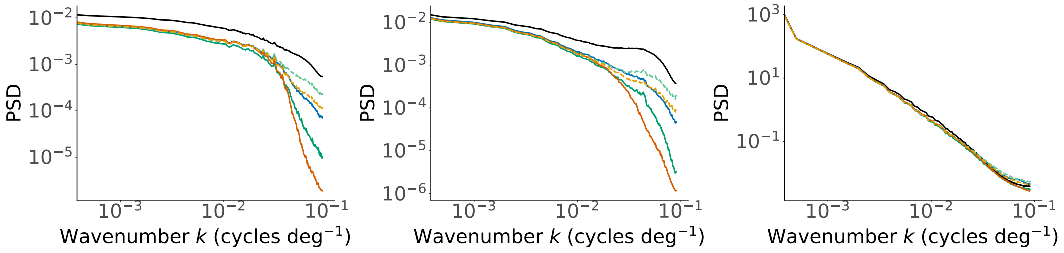
4.2 Results

Table 2 shows the mean-absolute error (MAE) and root-mean-square error (RMSE) for different models over Central Europe, the Iberia, Morocco region, and Northern Scandinavia, while Table 3 reports the CRPS. Results for *Regression-CorrDiff* and *Regression-CorrDiff-PSD* are omitted from this latter table as these models are deterministic, for which CRPS is equivalent to MAE [20]. Figure 3 shows the comparison of PSD curves computed between CERRA (ground truth) and model predictions.

Table 2 indicates that, in the in-distribution test over Central Europe, the three baseline models produce accurate downscalings showing low MAE and RMSE. Table 3 further shows that CRPS values for *CorrDiff* and *CRPS UNets* are also low, consistent with the results reported by Mardani et al. [28]. From Figure 3, we note that among the target variables, $\mathbf{t2m}$ is the most well-behaved and easy to predict: it exhibits the lowest errors and shows little sensitivity to the PSD loss, reflecting its smoother spatial structure and weaker small-scale variability compared to wind components [41]. The PSDs of $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{E}_h$ closely follow the reference spectrum, whereas those of $\boldsymbol{\zeta}_h$ and $\boldsymbol{\delta}_h$ deviate markedly. Particularly, the second row of Figure 3 indicates that these predictions are not fully physically consistent. This gap narrows when both *Regression-CorrDiff* and *CRPS-UNets* are trained with the PSD loss term, returning spectral curves that follow the reference more closely. Notably, after adding the PSD loss, *Regression-CorrDiff* surpasses the full probabilistic *CorrDiff* architecture in spectral fidelity, suggesting that frequency-domain regularization can offset some of the advantages typically associated with probabilistic approaches. Training *CorrDiff*'s diffusion component on the residuals of the *Regression-CorrDiff-PSD* backbone yields no noticeable improvement in our experiments. We hypothesize that training *Regression-CorrDiff* with a PSD loss already captures high-frequency content, leaving little signal for the diffusion stage. Consequently, the diffusion step adds negligible skill and we therefore omit that variant.



(a) PSD of $\mathbf{u}$ (b) PSD of $\mathbf{v}$ (c) PSD of temperature $\mathbf{t2m}$



(d) PSD of $\boldsymbol{\delta_h}$ (e) PSD of $\boldsymbol{\zeta}_h$ (f) PSD of $\mathbf{E}_h$

Figure 3: PSD comparison of down-scaled (top row) and physics-derived (bottom row) variables from CERRA (ground truth) and model predictions in Central Europe (in-distribution domain). Wavenumber k is plotted on a logarithmic scale.

In OOD domains (the Iberia, Morocco region and Northern Scandinavia), MAE and RMSE rise sharply relative to the in-distribution case (Table 2), and CRPS values (Table 3) also increase. Models trained with the PSD loss term achieve slightly better scores, yet, their performance remains well below the in-distribution level and predictions look blurrier (see Figure 1 and 2). For more extensive visual results see Appendix C, D, and E.

We also observe that variables the model is directly optimized for, such as $\mathbf{u}$ and $\mathbf{v}$, tend to match the ground truth more closely, particularly at low frequencies. This is expected, as loss functions like MSE emphasize minimizing errors in low-frequency components and tend to suppress high-frequency features such as sharp gradients. In contrast, derived variables such as $\boldsymbol{\delta}_h$ and $\boldsymbol{\zeta}_h$ exhibit a consistent offset from the ground truth, even at low frequencies. This suggests that the model is not merely missing high-frequency details, but is also failing to capture important underlying correlations between physical variables: the model can learn to reproduce wind components while missing the physics embedded in their spatial derivatives.

4.3 Qualitative discussion

In atmospheric dynamics, preserving the PSD slope across different scales is crucial: deviations imply that the model redistributes variance incorrectly among scales, leading to either excessive

Table 2: MAE and RMSE of $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{t2m}$ over Central Europe (in-distribution) and the Iberia, Morocco region / Northern Scandinavia (OOD).

Region	Variable	CorrDiff		Reg-CorrDiff		Reg-CorrDiff-PSD		CRPS UNets		CRPS UNets-PSD	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Central Europe	$\mathbf{u}$, m/s	**0.663**	**0.971**	0.674	0.978	0.676	0.987	0.671	0.977	0.678	0.982
	$\mathbf{v}$, m/s	**0.669**	**1.014**	0.681	1.026	0.685	1.040	0.679	1.037	0.690	1.037
	$\mathbf{t2m}$, K	**0.602**	**0.826**	0.622	0.852	0.619	0.857	0.619	0.855	0.617	0.853
Iberia, Morocco	$\mathbf{u}$, m/s	1.067	1.446	1.074	1.459	1.063	1.446	1.044	1.407	**1.035**	**1.398**
	$\mathbf{v}$, m/s	1.094	1.528	1.080	1.523	1.071	1.517	1.044	1.469	**1.042**	**1.460**
	$\mathbf{t2m}$, K	1.486	2.118	1.459	2.088	1.426	2.062	1.361	1.937	**1.323**	**1.857**
Northern Scandinavia	$\mathbf{u}$, m/s	1.021	1.355	1.042	1.398	0.999	1.332	0.932	1.258	**0.858**	**1.156**
	$\mathbf{v}$, m/s	1.001	1.338	1.010	1.365	0.981	1.344	0.925	1.225	**0.875**	**1.174**
	$\mathbf{t2m}$, K	1.481	1.883	1.471	1.908	1.390	1.823	**1.304**	**1.677**	1.405	1.835

Table 3: CRPS of $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{t2m}$ over Central Europe (in-distribution) and the Iberia, Morocco region / Northern Scandinavia (OOD).

Region	Variable	CorrDiff	CRPS UNets	CRPS UNets-PSD
Central Europe	$\mathbf{u}$, m/s	**0.480**	0.515	0.516
	$\mathbf{v}$, m/s	**0.476**	0.520	0.520
	$\mathbf{t2m}$, K	**0.432**	0.470	0.467
Iberia, Morocco	$\mathbf{u}$, m/s	0.828	0.801	**0.786**
	$\mathbf{v}$, m/s	0.849	0.815	**0.797**
	$\mathbf{t2m}$, K	1.197	1.077	**1.033**
Northern Scandinavia	$\mathbf{u}$, m/s	0.760	0.718	**0.661**
	$\mathbf{v}$, m/s	0.750	0.727	**0.692**
	$\mathbf{t2m}$, K	1.102	1.017	**1.075**

smoothing (loss of small-scale structure) or spurious noise. Standard pixel-space losses (ℓ_1 or ℓ_2 loss) implicitly weight low-k modes more heavily, since large-scale errors dominate the sum of squared differences. This biases deep learning downscalers toward fitting coarse features accurately at the expense of mesoscale and sub-mesoscale dynamics.

Further, $\boldsymbol{\delta}_h$ and $\boldsymbol{\zeta}_h$ are derived from spatial derivatives of the prognostic variables $\mathbf{u}$ and $\mathbf{v}$. In Fourier space, differentiation multiplies each mode by its wavenumber magnitude (see Appendix A), so derivative-based diagnostics amplify high-frequency components. Thus, if the high-wavenumber part of the PSD is under-represented, these derived variables systematically lose variance, even if the base fields $\mathbf{u}$ and $\mathbf{v}$ have satisfactory accuracy in pixel-space metrics. This explains why our baselines reproduce $\mathbf{E}_h$ more faithfully than $\boldsymbol{\delta}_h$ or $\boldsymbol{\zeta}_h$: $\mathbf{E}_h$ depends on the square of the velocity itself (low- and mid-k dominated), whereas divergence and vorticity are explicitly high-k weighted.

By directly optimizing a PSD-based term, we reward matching the observed scale-by-scale variance distribution, rather than only minimizing pixel-space errors. This frequency-domain regularization reduces the high-k deficit in $\mathbf{u}$ and $\mathbf{v}$ and, through the k-weighting of derivatives, improves the physical consistency of $\boldsymbol{\delta}_h$ and $\boldsymbol{\zeta}_h$ as well.

5 Conclusion

Our findings confirm that the considered baselines struggle to generalize to unseen geographic regions, underscoring the need for downscaling models explicitly designed for both spatial transferability and physical fidelity. We have also shown that a simple spectral loss has a clear and consistent impact on the physical realism of the predictions: it substantially improves the match between predicted and reference spectra, particularly at high spatial frequencies, and narrows the gap for physically derived variables such as $\boldsymbol{\delta}_h$ and $\boldsymbol{\zeta}_h$. These results indicate that, while geographic transferability remains challenging, explicitly encouraging the reconstruction of scale-by-scale variance is an effective way to preserve small-scale physical structures. In future work, it would be interesting to incorporate kinetic energy, divergence, and vorticity as either soft constraints in the loss function to guide the optimization of the velocities, following Sekiyama et al. [38] or integrating the constraints as part of the model architecture akin to other research in the fluid dynamics and machine learning literature [12, 7, 8].

References

[1] Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R. Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals, 2025. URL `https://arxiv.org/abs/2506.10772`.

[2] David C. Bader et al. Climate models: An assessment of strengths and limitations. U.S. Climate Change Science Program Synthesis and Assessment Product 3.1, U.S. Department of Energy, Washington, D.C., 2009. U.S. Climate Change Science Program Synthesis and Assessment Product 3.1.

[3] 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.

[4] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the Earth system. *Nature*, pages 1–8, 2025.

[5] Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024. ISSN 1944-8007. doi: 10.1029/2023GL107377.

[6] Annalisa Bracco, Julien Brajard, Henk A Dijkstra, Pedram Hassanzadeh, Christian Lessig, and Claire Monteleoni. Machine learning for the physics of climate. *Nature Reviews Physics*, 7(1): 6–20, 2025.

[7] Haitz Sáez de Ocáriz Borde, David Sondak, and Pavlos Protopapas. Convolutional neural network models and interpretability for the anisotropic reynolds stress tensor in turbulent one-dimensional flows. *Journal of Turbulence*, 23(1-2):1–28, 2022. doi: 10.1080/14685248.2021.1999459. URL `https://doi.org/10.1080/14685248.2021.1999459`.

[8] Haitz Sáez de Ocáriz Borde, Pietro Innocenzi, and Flavio Savarino. Aerothermodynamic simulators for rocket design using neural fields, 2023. URL `https://arxiv.org/abs/2303.10283`.

[9] John A. Dutton. *The Ceaseless Wind: An Introduction to the Theory of Atmospheric Motion*. Dover Publications, Inc., USA, 2 edition, 1986. ISBN 0-486-65096-0.

[10] Bijan Fallah, Masoud Rostami, Emmanuele Russo, Paula Harder, Christoph Menz, Peter Hoffmann, Iulii Didovets, and Fred F Hattermann. Climate model downscaling in central Asia: a dynamical and a neural network approach. *Geoscientific Model Development*, 18(1):161–180, 2025.

[11] Xiantao Fan, Deepak Akhare, and Jian-Xun Wang. Neural differentiable modeling with diffusion-based super-resolution for two-dimensional spatiotemporal turbulence. 433:117478. ISSN 0045-7825. doi: 10.1016/j.cma.2024.117478. URL `https://linkinghub.elsevier.com/retrieve/pii/S0045782524007321`. Publisher: Elsevier BV.

[12] Rui Fang, David Sondak, Pavlos Protopapas, and Sauro Succi. Neural network models for the anisotropic reynolds stress tensor in turbulent channel flow. *Journal of Turbulence*, 21(9-10): 525–543, 2020.

[13] Stathi Fotiadis, Noah D Brenowitz, Tomas Geffner, Yair Cohen, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Adaptive flow matching for resolving small-scale physics. In *Forty-second International Conference on Machine Learning*.

[14] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2360–2369, 2021.

[15] 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, 2025.

[16] W. J. Gutowski, P. A. Ullrich, A. Hall, L. R. Leung, T. A. O'Brien, C. M. Patricola, R. W. Arritt, M. S. Bukovsky, K. V. Calvin, Z. Feng, A. D. Jones, G. J. Kooperman, E. Monier, M. S. Pritchard, S. C. Pryor, Y. Qian, A. M. Rhoades, A. F. Roberts, K. Sakaguchi, N. Urban,

and C. Zarzycki. The ongoing need for high-resolution regional climate models: Process understanding and stakeholder information. *Bulletin of the American Meteorological Society*, 101(5):E664 – E683, 2020. doi: 10.1175/BAMS-D-19-0113.1. URL `https://journals.ametsoc.org/view/journals/bams/101/5/bams-d-19-0113.1.xml`.

[17] Gregory J. Hakim and Sanjit Masanam. Dynamical tests of a deep-learning weather prediction model. *Artificial Intelligence for the Earth Systems*, May 2024. ISSN 2769-7525. doi: 10.1175/AIES-D-23-0090.1.

[18] Paula Harder, Alex Hernandez-Garcia, Venkatesh Ramesh, Qidong Yang, Prasanna Sattegeri, Daniela Szwarcman, Campbell Watson, and David Rolnick. Hard-constrained deep learning for climate downscaling. *Journal of Machine Learning Research*, 24(365):1–40, 2023.

[19] Nidhin Harilal, Bri-Mathias S Hodge, Claire Monteleoni, and Aneesh Subramanian. EnhancedSD: Downscaling solar irradiance from climate model projections. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022. URL `https://www.climatechange.ai/papers/neurips2022/67`.

[20] Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

[21] 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 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 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, July 2020. ISSN 0035-9009, 1477-870X. doi: 10.1002/qj.3803.

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

[23] Mark Z. Jacobson. *Fundamentals of Atmospheric Modeling*. Cambridge University Press, Cambridge, second edition edition, 2005. ISBN 978-1-139-16538-9. doi: 10.1017/CBO9781139165389.

[24] Azin Al Kajbaf, Michelle Bensi, and Kaye L Brubaker. Temporal downscaling of precipitation from climate model projections using machine learning. *Stochastic Environmental Research and Risk Assessment*, 36(8):2173–2194, 2022.

[25] Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. 118(21):e2101784118. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2101784118. URL `https://pnas.org/doi/full/10.1073/pnas.2101784118`.

[26] 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.

[27] Douglas Maraun, Frederick Wetterhall, Anderson M Ireson, Richard E Chandler, Elizabeth J Kendon, Martin Widmann, S Brienen, Henning W Rust, Tobias Sauter, Matthias Themeßl, et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, 48(3), 2010.

[28] 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.

[29] 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, 2024.

[30] Evan Mills. Insurance in a climate of change. *Science*, 309(5737):1040–1044, 2005.

[31] Joel Oskarsson, Tomas Landelius, Marc Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with hierarchical graph neural networks. *Advances in Neural Information Processing Systems*, 37:41577–41648, 2024.

[32] PhysicsNeMo Contributors. NVIDIA PhysicsNeMo: An open-source framework for physics-based deep learning in science and engineering. `https://github.com/NVIDIA/physicsnemo`, 2023. Version released 24 February 2023.

[33] 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. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

[34] Liying Qiu, Rahman Khorramfar, Saurabh Amin, and Michael F Howland. Decarbonized energy system planning with high-resolution spatial representation of renewables lowers cost. *Cell Reports Sustainability*, 1(12), 2024.

[35] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 00219991. doi: 10.1016/j.jcp.2018.10.045.

[36] Eric P Salathé Jr, L Ruby Leung, Yun Qian, and Yongxin Zhang. Regional climate model projections for the state of washington. *Climatic Change*, 102(1):51–75, 2010.

[37] S. Schimanke, M. Ridal, P. Le Moigne, L. Berggren, P. Undén, R. Randriamampianina, U. Andrea, E. Bazile, A. Bertelsen, P. Brousseau, P. Dahlgren, L. Edvinsson, A. El Said, M. Glinton, S. Hopsch, L. Isaksson, R. Mladek, E. Olsson, A. Verrelle, and Z.Q. Wang. CERRA sub-daily regional reanalysis data for Europe on single levels from 1984 to present, 2021.

[38] Tsuyoshi Thomas Sekiyama, Syugo Hayashi, Ryo Kaneko, and Ken-ichi Fukui. Surrogate downscaling of mesoscale wind fields using ensemble superresolution convolutional neural networks. *Artificial Intelligence for the Earth Systems*, 2(3), July 2023. ISSN 2769-7525. doi: 10.1175/aies-d-23-0007.1.

[39] T. Selz and G. C. Craig. Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophysical Research Letters*, 50(20):e2023GL105747, 2023. ISSN 1944-8007. doi: 10.1029/2023GL105747.

[40] William C. Skamarock, Joseph B. Klemp, Jimy Dudhia, David O. Gill, Zhiquan Liu, Judith Berner, Wei Wang, Jordan G. Powers, Michael G. Duda, Dale M. Barker, and Xiang-Yu Huang. A description of the Advanced Research WRF Model Version 4.3. Technical report, UCAR/NCAR, July 2021.

[41] Elena Tomasi, Gabriele Franch, and Marco Cristoforetti. Can ai be enabled to perform dynamical downscaling? a latent diffusion model to mimic kilometer-scale cosmo5. 0_clm9 simulations. *Geoscientific Model Development*, 18(6):2051–2078, 2025.

[42] Robert L Wilby, TML Wigley, D Conway, PD Jones, BC Hewitson, J Main, and DS Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, 34(11):2995–3008, 1998.

[43] Chiu-Wai Yan, Shi Quan Foo, Van Hoan Trinh, Dit-Yan Yeung, Ka-Hing Wong, and Wai-Kin Wong. Fourier amplitude and correlation loss: Beyond using l2 loss for skillful precipitation nowcasting. *Advances in Neural Information Processing Systems*, 37:100007–100041, 2024.

Appendix

A Fourier Differentiation Amplifies High Frequencies

Fourier transform and wavenumbers. Let $\mathbf{q} \in \mathbb{R}^{H \times W}$ be a real-valued scalar field defined on a uniform $H \times W$ grid with isotropic grid spacing Δx. We denote its 2D discrete Fourier transform by $\mathcal{F}(\mathbf{q})(k_h, k_w)$, where k_h and k_w are discrete mode indices. The corresponding physical wavenumbers are

$$\kappa_x = \frac{2\pi k_w}{W \Delta x}, \qquad \text{and} \qquad \kappa_y = \frac{2\pi k_h}{H \Delta x}.$$

We define the horizontal wavenumber vector $\boldsymbol{\kappa} = (\kappa_x, \kappa_y)$ and its magnitude $\kappa = \|\boldsymbol{\kappa}\| = \sqrt{\kappa_x^2 + \kappa_y^2}$. High values of κ correspond to high-frequency, small-scale features. Then, the discrete PSD of $\mathbf{q}$ is given by

$$\mathrm{PSD}(\mathbf{q})(k_h, k_w) = \frac{\left|\mathcal{F}(\mathbf{q})(k_h, k_w)\right|^2}{H\, W\, \Delta x}.$$

This quantity measures the contribution of each wavenumber mode to the field's total variance.

Now we examine how spatial derivatives affect its spectral content.The discrete Fourier transform of a spatial derivative is obtained by multiplying by i times the corresponding wavenumber:

$$\mathcal{F}\left(\frac{\partial \mathbf{q}}{\partial x}\right) = i\, \kappa_x\, \mathcal{F}(\mathbf{q}), \qquad \text{and} \qquad \mathcal{F}\left(\frac{\partial \mathbf{q}}{\partial y}\right) = i\, \kappa_y\, \mathcal{F}(\mathbf{q}).$$

Therefore, the PSDs of the derivatives are

$$\mathrm{PSD}\left(\frac{\partial \mathbf{q}}{\partial x}\right) = \kappa_x^2\, \mathrm{PSD}(\mathbf{q}), \qquad \text{and} \qquad \mathrm{PSD}\left(\frac{\partial \mathbf{q}}{\partial y}\right) = \kappa_y^2\, \mathrm{PSD}(\mathbf{q}).$$

Consequently, for the horizontal gradient $\nabla \mathbf{q} = (\partial_x \mathbf{q},\ \partial_y \mathbf{q})$, we have

$$\left\|\mathcal{F}(\nabla \mathbf{q})\right\|^2 = \left\|\left[\mathcal{F}\left(\tfrac{\partial \mathbf{q}}{\partial x}\right) \quad \mathcal{F}\left(\tfrac{\partial \mathbf{q}}{\partial y}\right)\right]\right\|^2 = \kappa^2 \left|\mathcal{F}(\mathbf{q})\right|^2,$$

so that the corresponding PSD is

$$\mathrm{PSD}(\nabla \mathbf{q}) = \kappa^2\, \mathrm{PSD}(\mathbf{q}).$$

Since κ increases with frequency, the κ^2 factor explicitly shows that high-frequency modes are multiplied by larger factors, meaning they are *amplified*.

Example: horizontal divergence. Let $\boldsymbol{\delta}_h$ be the horizontal divergence defined as

$$\boldsymbol{\delta}_h = \nabla_h \cdot \mathbf{v}_h = \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial y}.$$

In Fourier space,

$$\mathcal{F}(\boldsymbol{\delta}_h) = i\, \kappa_x\, \mathcal{F}(\mathbf{u}) + i\, \kappa_y\, \mathcal{F}(\mathbf{v}).$$

Its power spectrum is

$$\left|\mathcal{F}(\boldsymbol{\delta}_h)\right|^2 = \kappa_x^2 \left|\mathcal{F}(\mathbf{u})\right|^2 + \kappa_y^2 \left|\mathcal{F}(\mathbf{v})\right|^2 + 2\, \kappa_x \kappa_y\, \Re\big(\mathcal{F}(\mathbf{u})\, \mathcal{F}(\mathbf{v})^*\big).$$

The leading κ_x^2 and κ_y^2 factors again make the amplification of high-κ modes explicit, demonstrating why high-frequency modes are strongly amplified.

Differentiation in Fourier space multiplies each mode by its wavenumber magnitude κ, so the power in derivative-based diagnostics is weighted by κ^2. For the horizontal wind components $\mathbf{u}$ and $\mathbf{v}$, any loss of high-frequency variance in $\mathrm{PSD}(\mathbf{u})$ or $\mathrm{PSD}(\mathbf{v})$ is magnified in the spectra of $\boldsymbol{\delta}_h$ and $\boldsymbol{\zeta}_h$, since their definitions involve first-order derivatives in both spatial directions. Consequently, models that over-smooth $\mathbf{u}$ and $\mathbf{v}$ will not only misrepresent their own high-κ PSDs, but will show even larger discrepancies in the PSDs of $\boldsymbol{\delta}_h$ and $\boldsymbol{\zeta}_h$. By explicitly constraining the model to match the full scale-dependent variance of $\mathbf{u}$ and $\mathbf{v}$ through a PSD-based loss, we reduce this high-frequency deficit and thereby improve the physical consistency of all derivative-based diagnostics.

B Dataset and training details

ERA5 [21] and CERRA [37] are both multi-decade reanalysis products produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) using state-of-the-art numerical weather models with data assimilation. CERRA provides high-resolution fields over Europe at 5.5 km grid spacing and a temporal resolution of three hours, whereas ERA5 covers the globe at 25 km resolution and hourly intervals. This corresponds to a spatial super-resolution factor of approximately $4.54\times$ in each horizontal direction. Since the downscaling task requires temporally aligned input–target pairs, we sub-sample ERA5 to every third hour to match the temporal availability of CERRA.

Training, validation, and test splits. For training, we select a Central European domain bounded by $[60°\mathrm{N}, -2°\mathrm{E}, 40°\mathrm{N}, 18°\mathrm{E}]$ The training period spans from 1 January 2014 to 31 December 2020. We reserve the year 2021 for validation and in-distribution testing: January, March, May, July, September, and November are used for testing, while the remaining months serve as the validation set.

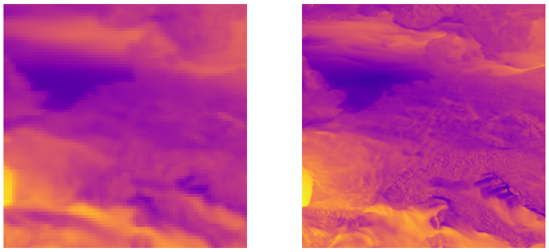
Geographic generalization experiments. To evaluate cross-regional generalization, we use the same 2021 monthly split but replace the Central European domain with two non-overlapping target regions:

- **Iberia, Morocco:** $[45°\mathrm{N}, -15°\mathrm{E}, 25°\mathrm{N}, 5°\mathrm{E}]$
- **Northern Scandinavia:** $[70°\mathrm{N}, 24°\mathrm{E}, 50°\mathrm{N}, 44°\mathrm{E}]$

These domains differ substantially in climate and topography from the training region, providing a robust test of spatial transferability.

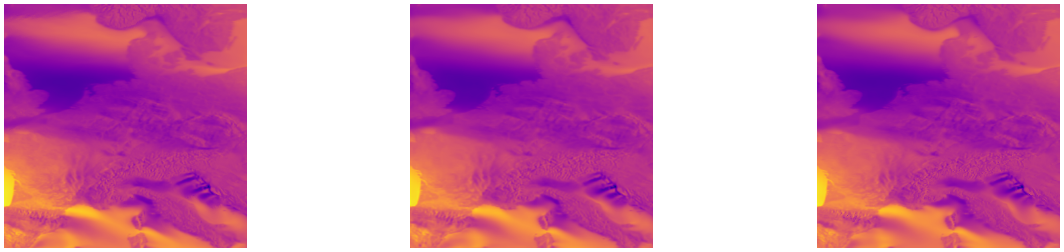
Training details. We train *CorrDiff* and its regression component, *Regression-CorrDiff*, by adapting the original *PhysicsNeMo* code repository provided by the authors [32]. To train the CRPS-based U-Net ensemble, we follow the methodology of Alet et al. [1], but still adapt the U-Net architecture from the *PhysicsNeMo* repository to construct the ensemble. Unlike Alet et al. [1], who train four models (M1, M2, M3, M4) each with its own ensemble, we restrict our experiments to a single model configuration (M1).

C Results Central Europe



(a) Input: ERA5

(b) Target: CERRA

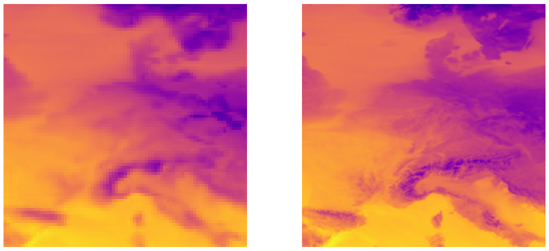


(c) CorrDiff

(d) CRPS UNets-PSD

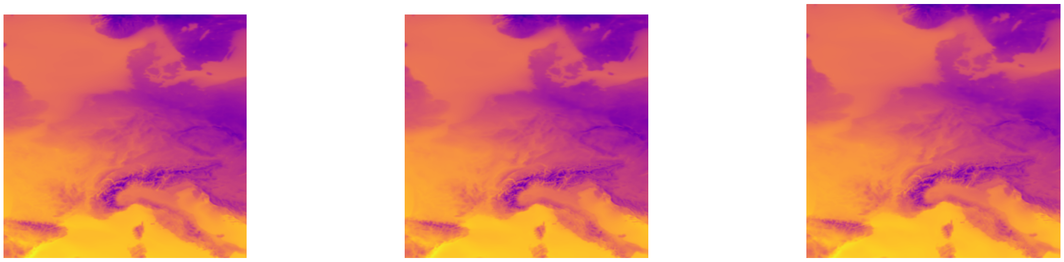
(e) Reg-CorrDiff-PSD

Figure 4: Input, target, and predictions of $\mathbf{u}$ in Central Europe (in-distribution) using *CorrDiff*, *Reg-CorrDiff-PSD*, and *CRPS UNets-PSD*. *CorrDiff* yields a visually sharper prediction, while the others appear slightly blurrier.



(a) Input: ERA5

(b) Target: CERRA



(c) CorrDiff

(d) CRPS UNets-PSD

(e) Reg-CorrDiff-PSD

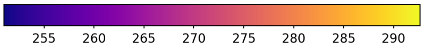
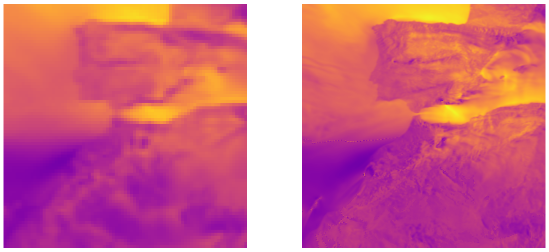
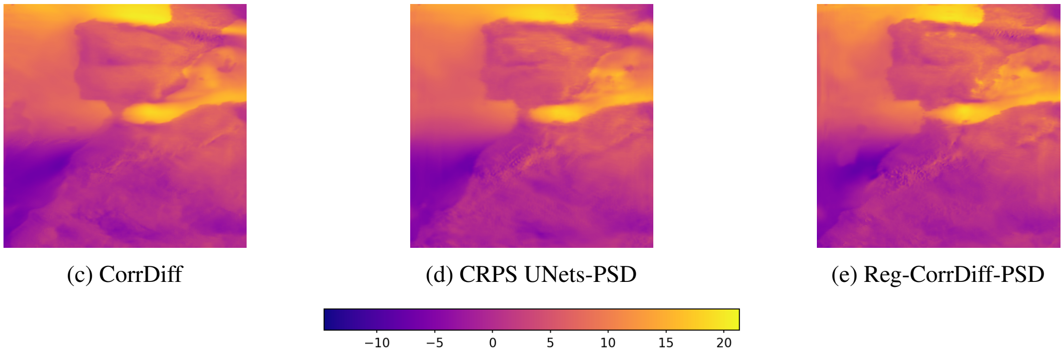


Figure 5: Input, target, and predictions of $\mathtt{t2m}$ in Central Europe (in-distribution) using *CorrDiff*, *Reg-CorrDiff-PSD*, and *CRPS UNets-PSD*. $\mathtt{t2m}$ is easier to predict, hence differences between outputs are hard to spot visually.

D Results Iberia, Morocco

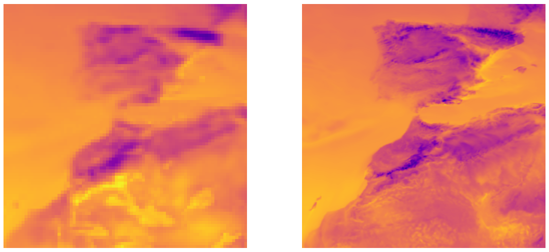


(a) Input: ERA5 (b) Target: CERRA

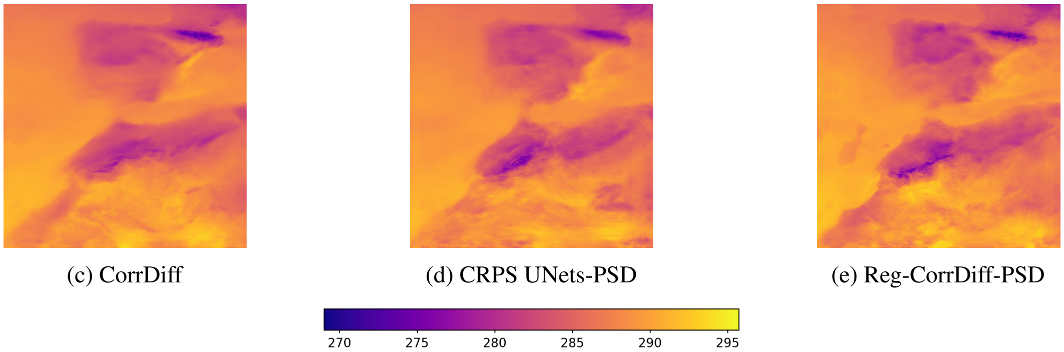


(c) CorrDiff (d) CRPS UNets-PSD (e) Reg-CorrDiff-PSD

Figure 6: Input, target, and predictions of $\mathbf{u}$ in Iberia, Morocco (OOD) using *CorrDiff*, *Reg-CorrDiff-PSD*, and *CRPS UNets-PSD*. *CRPS UNets-PSD* yields a visually sharper prediction, but still significantly worse compared to the in-distribution scenario.

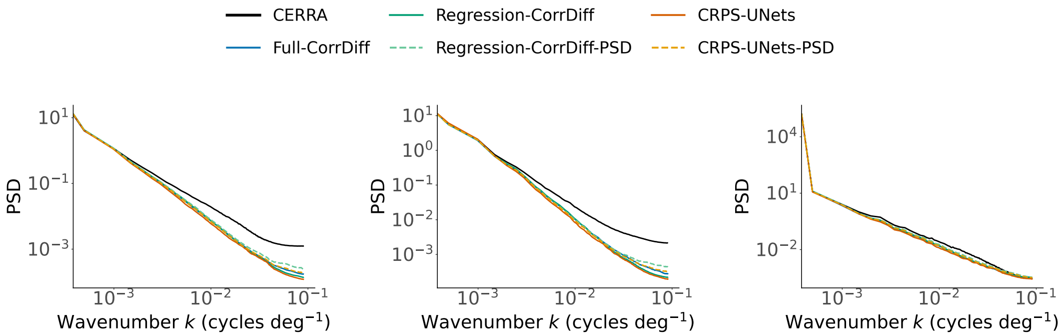


(a) Input: ERA5 (b) Target: CERRA

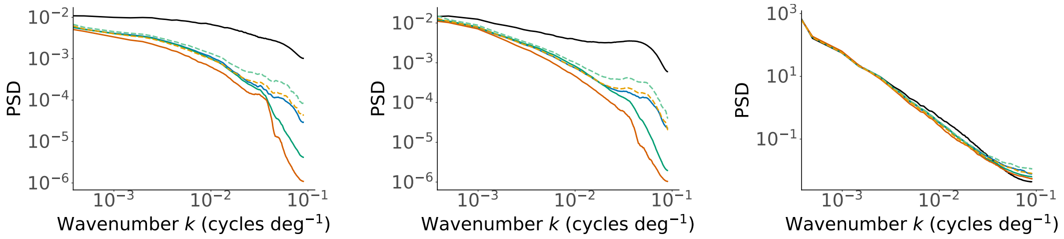


(c) CorrDiff (d) CRPS UNets-PSD (e) Reg-CorrDiff-PSD

Figure 7: Input, target, and predictions of $\mathbf{t2m}$ in Iberia, Morocco (OOD) using *CorrDiff*, *Reg-CorrDiff-PSD*, and *CRPS UNets-PSD*. *CRPS UNets-PSD* yields a visually sharper prediction, but still significantly worse compared to the in-distribution scenario.



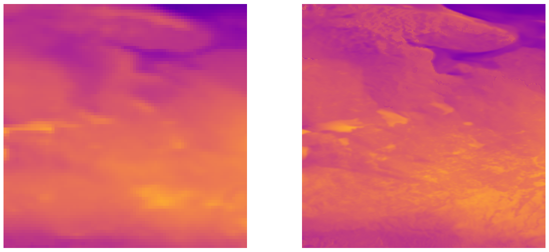
(a) PSD of $\mathbf{u}$ (b) PSD of $\mathbf{v}$ (c) PSD of temperature $\mathbf{t2m}$



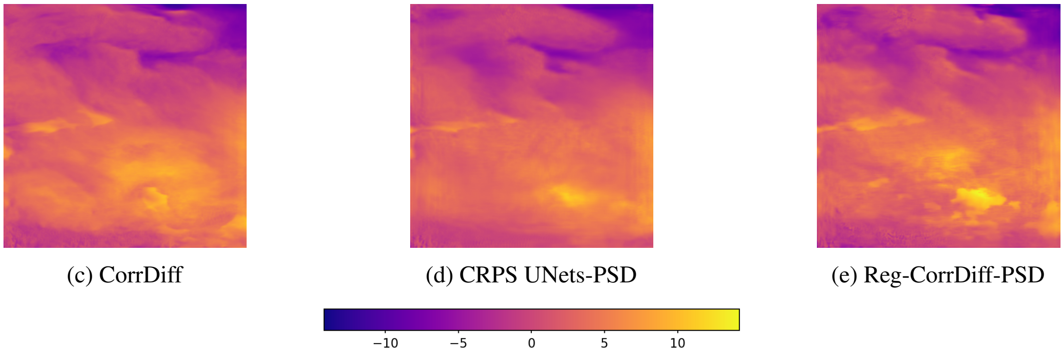
(d) PSD of $\boldsymbol{\delta}_h$ (e) PSD of $\boldsymbol{\zeta}_h$ (f) PSD of $\mathbf{E}_h$

Figure 8: PSD comparison of down-scaled and physics-derived variables from CERRA (ground truth) and model predictions in the Iberia, Morocco region (OOD). For all models, PSD curves deviate more from CERRA at high frequencies than in the in-distribution case, with an even larger gap for $\boldsymbol{\delta_h}$ and $\boldsymbol{\zeta_h}$.

E Results Northern Scandinavia

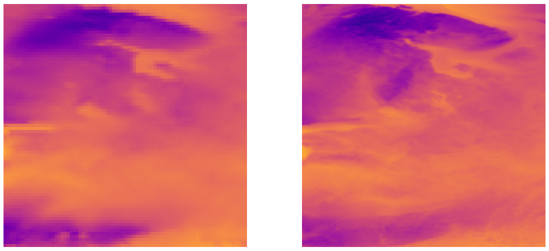


(a) Input: ERA5 (b) Target: CERRA

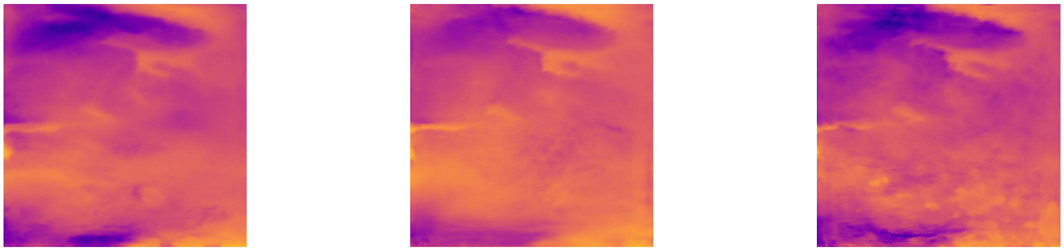


(c) CorrDiff (d) CRPS UNets-PSD (e) Reg-CorrDiff-PSD

Figure 9: Input, target, and predictions of $\mathbf{u}$ in Northern Scandinavia (OOD) using *CorrDiff*, *Reg-CorrDiff-PSD*, and *CRPS UNets-PSD*. Visually, all models miss many details of the target $\mathbf{u}$, making it difficult to see clear differences in quality.



(a) Input: ERA5 (b) Target: CERRA



(c) CorrDiff (d) CRPS UNets-PSD (e) Reg-CorrDiff-PSD

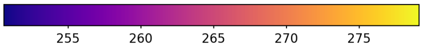
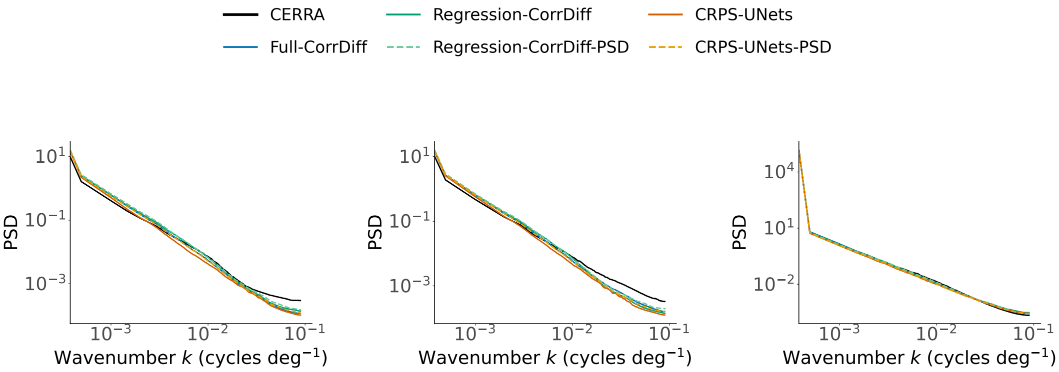
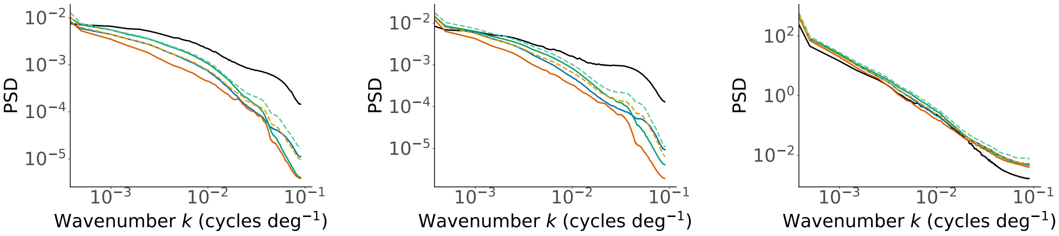


Figure 10: Input, target, and predictions of $\mathbf{t2m}$ in Northern Scandinavia (OOD) using *CorrDiff*, *Reg-CorrDiff-PSD*, and *CRPS UNets-PSD*. Visually, all models miss many details of the target $\mathbf{t2m}$, making it difficult to see clear differences in quality.



(a) PSD of $\mathbf{u}$ (b) PSD of $\mathbf{v}$ (c) PSD of temperature $\mathbf{t2m}$



(d) PSD of $\boldsymbol{\delta}_h$ (e) PSD of $\boldsymbol{\zeta}_h$ (f) PSD of $\mathbf{E}_h$

Figure 11: PSD comparison of down-scaled and physics-derived variables from CERRA (ground truth) and model predictions in Northern Scandinavia (OOD). For all models, PSD curves deviate more from CERRA at high frequencies than in the in-distribution case, with an even larger gap for $\boldsymbol{\delta_h}$ and $\boldsymbol{\zeta_h}$.