FROM GLOBAL TO LOCAL: EFFICIENT REGIONAL WEATHER DOWNSCALING WITH GLOBAL WEATHER FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate regional weather prediction requires resolving fine-scale structure while remaining consistent with global dynamics. Traditional limited-area models rely on computationally expensive simulations, while many learning-based approaches frame the problem as super-resolution, overlooking statistical and physical mismatches across scales. We propose a foundation-model-driven downscaling framework that learns regional refinements of global forecasts by augmenting a pretrained weather model backbone with lightweight, multi-scale prediction heads operating directly in its latent space. Despite being pretrained on substantially coarser data, the backbone supports regional adaptation at a tenfold finer grid spacing, without retraining it. The proposed approach uses operational global analyses as inputs and high-resolution regional NWP fields as training targets. It is evaluated not only against gridded datasets but also against ground-based weather station observations, enabling analysis of systematic biases between global analyses, regional WRF forecasts, and real world mesurements. Our model outperforms or matches operational WRF-ARW forecasts on all surface variables and achieves lower RMSE on upper-air variables, at a fraction of the computational cost. It also shows markedly better rollout stability.

1 INTRODUCTION

Accurate weather prediction, normally solved by Numerical Weather Prediction (NWP) models (NWP, 1922) became one of the key technologies in today's applications spanning from precision farming (Maier & Dandy, 2000) and urban flood management (Papagiannaki et al., 2015) to stabilization of the energy grid by accurate predictions of renewable energy sources (Ukoba et al., 2024). However, all of the aforementioned applications require forecasts that are not only precise but also produced quickly. NWP methods address this challenge through localized mesoscale approaches, such as the Weather Research and Forecasting (WRF) model (Skamarock et al., 2008), including its most recent state-of-the-art solvers: Advanced Research WRF (WRF-ARW) (Skamarock et al., 2019) and the nonhydrostatic mesoscale model (WRF-NMM) (Janjic, 2003). Their widespread adoption in operational forecasting systems demonstrates their reliability as a source of meteorological data. However, the computationally intensive nature of NWP simulations limits their suitability for applications requiring rapid turnaround (Vaughan et al., 2024). This limitation has driven growing interest in alternative methods for weather prediction, such as Data-Driven Models (DDM)s. Recently, AI models such as GraphCast, Aurora or FourCastNet (Lam et al., 2023; Bodnar et al., 2025; Pathak et al., 2022) have matched NWP accuracy, providing global weather forecasts at 0.25° resolution, limited by the ERA5 dataset (Hersbach et al., 2018), while being orders of magnitude faster. Moreover, even subjects responsible for providing global forecasts using traditional methods, took interest in DDMs and created their own models (Lang et al., 2024). On the same level, ML-based Limited Area Models (ML LAMs) became a substitute of traditional mesoscale methods, aimed at modelling weather phenomena at finer scales, providing both accuracy and speed. However, until recently they heavily relied on input from NWP models for stabilizing training, by providing the most accurate Local Boundary Condition (LBC) (Xu et al., 2025; Larsson et al., 2025). This, and the fact that most of the ML LAM models need to be trained from scratch (Adamov et al., 2025) or require full model fine-tuning (Munir et al., 2024) is the motivation of our work. In this work we propose a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

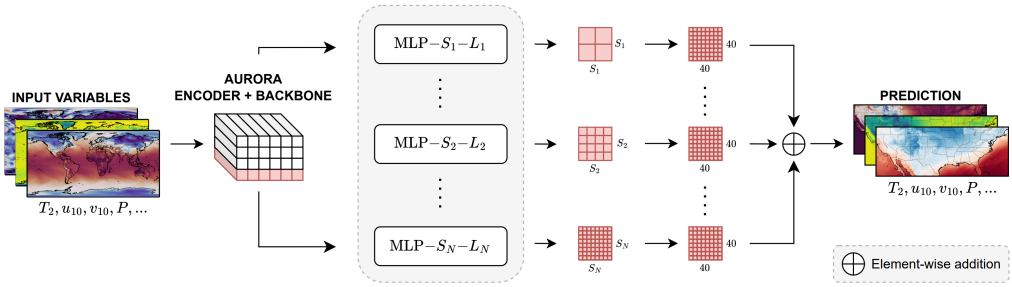


Figure 1: Overview of the workflow. Decoder variants augment or replace Aurora's read-out with additional linear or MLP heads operating at different spatial resolutions. For simplicity, decoder heads are denoted as MLP-S-L, where S indicates the native output resolution and L the number of hidden layers. The special case $L = 0$ corresponds to a linear decoder. In general, each head may use arbitrary (S, L), and multiple heads ($n = 1, \ldots, N$) can be composed. The figure shows only the case of parallel heads with elementwise summation after upsampling to 40×40, while sequential compositions are omitted. The architecture is agnostic to the specific choice of input and target variables, subject to the constraints imposed by the pretrained backbone; the variables shown correspond to those used in our experiments.

new approach for building weather downscaling models that utilizes pretrained foundation model – Aurora (Bodnar et al., 2025), without the need of fine-tuning most of the architecture involving heavy computation. We use standard IFS HRES $0.25°$ input states at given timestamps to generate outputs at the WRF $0.025°$ resolution for surface-level weather variables: temperature, u and v wind components, mean sea level pressure, and upper-air temperature, u and v wind components, geopotential and specific humidity. We evaluate our approach on all US-based HadISD (Dunn et al., 2012) in-situ observations, including coastline stations and HRRR forecasts dataset produced by the WRF-ARW model (Dowell et al., 2022). We compare also to other available ML-LAM approaches. Additionally, we present extensive physical consistency analysis.

Our main contributions are:

- **No explicit regional LBC forcing:** We propose a framework that eliminates dependency on NWP-based Local Boundary Conditions (LBCs) for Limited Area Models. Unlike traditional ML-LAMs that require region-specific retraining and fail when driving regional forecasts are unavailable, our method uses global foundation model representations directly, improving operational robustness and simplifying the training pipeline.
- **Latent-Space Regional Adaptation:** We demonstrate that latent-space decoupling enhances adaptation from global to regional scales, allowing a frozen $0.25°$ model to perform $0.025°$ mesoscale forecasting without backbone retraining, using lightweight decoder heads.
- **Operational Benchmarking:** We evaluate against operational NWP (WRF-ARW) on surface and atmospheric variables across both gridded and in-situ data, demonstrating stable 48h rollouts and improved accuracy at inference speeds orders of magnitude faster than NWP.
- **Multi-region Training:** We show that lightweight decoder heads trained jointly on CONUS and Denmark regions (7000 km apart, different hemispheres) maintain accuracy comparable to region-specific models, eliminating the need for separate training pipelines per geographic domain.

2 Related Works

Early machine learning approaches were explored as data-driven alternatives to numerical weather prediction (NWP) due to the high computational and memory costs of traditional models. Using historical reanalysis datasets such as ERA-Interim and ERA5, these methods framed weather forecasting as a supervised learning problem, learning mappings between atmospheric states and enabling short- to medium-range forecasts at significantly reduced inference cost (Dueben & Bauer, 2018).

CNN-based models (Weyn et al., 2020) subsequently achieved promising performance for variables such as geopotential height, temperature, and wind, while offering substantial speed-up over numerical baselines inference. However, their reliance on purely spatial convolutions and limited temporal modeling constrained forecast lead times and generalization. To address these limitations, recent work has shifted toward large-scale deep learning systems that jointly model atmospheric dynamics across space and time using GNNs and transformers (Scarselli et al., 2009; Vaswani et al., 2017), aiming to match operational NWP skill at substantially lower inference cost.

Among end-to-end systems, Pangu-Weather used a hierarchical transformer to capture multiscale atmospheric structure and achieved state-of-the-art performance (Bi et al., 2023). GraphCast formulated global forecasting on a spherical graph for improved spatial modeling (Lam et al., 2023), while ClimaX introduced a foundation model approach emphasizing scalability and adaptability across weather and climate tasks (Nguyen et al., 2023). One of the most successful global models is Aurora (Bodnar et al., 2025), a foundation model trained on vast amounts of data, enabling it to learn a strong latent representation of the climate. Despite these advances, such models operate at relatively coarse spatial resolutions (e.g., $0.25°$), limiting their applicability to mesoscale forecasting. Our work builds on Aurora by adopting its $0.25°$ resolution base model, which was originally designed for fine-tuning across a range of downstream tasks.

A complementary line of research adapts pretrained weather foundation models to downstream tasks rather than training from scratch. Prithvi WxC (Schmude et al., 2024), a 2.3-billion-parameter model pretrained on MERRA-2 with a joint masked-reconstruction and forecasting objective, exemplifies the full fine-tuning route: the entire backbone is retrained for each downstream task, including downscaling of reanalysis fields and CORDEX regional climate projections. Since such retraining is computationally prohibitive at scale, Cao et al. (2026) examine parameter-efficient fine-tuning (PEFT) for weather foundation models and introduce WeatherPEFT, which combines task-adaptive dynamic prompting with Fisher-information-guided selection of trainable backbone parameters. Their downscaling experiments with the Aurora backbone show that standard PEFT methods, such as LoRA and DoRA, yield 25–36% higher RMSE than full fine-tuning, and that the deficit persists even when the PEFT parameter budget is increased by an order of magnitude. Parity with full fine-tuning is reached only by the purpose-built WeatherPEFT, and only once its parameter budget is relaxed accordingly. These results indicate that sparse backbone modifications struggle to capture fine-grained spatial transformations required for downscaling, motivating our approach of keeping the backbone entirely frozen and confining all trainable parameters to lightweight, resolution-specialized output heads. A related frozen-backbone strategy was explored by Lehmann et al. (2025), who attach shallow MLP decoders to Aurora's latent representations to predict new hydrological and energy variables at the model's native $0.25°$ resolution. They observe that decoder skill tracks the physical correlation between new and pretraining variables – evidence that pretrained latent spaces encode transferable physical relationships. However, their work extends the model along the variable axis (new outputs at fixed resolution), whereas we extend it along the resolution axis (existing outputs at $0.25°$, a 100 increase in grid-cell density). This distinction is consequential: predicting new variables preserves spatial scales present during pretraining, while downscaling requires generating fine-scale structure absent from the pretrained output space. To our knowledge, this is the first demonstration of weather foundation model adaptation achieving order-of-magnitude resolution enhancement with a completely frozen backbone.

In recent years Machine Learning based Limited Area Models (ML LAMs) started to emerge as a solution for replacing mesoscale simulations, offering fine-scale predictions at a fraction of time consumed by NWP models. First ML LAMs required input from traditional forecasts as a way to stabilize training. They employ ground truth weather states as boundary forcing conditions (Oskarsson et al., 2023; Xu et al., 2025) to keep errors from accumulating across next roll-out steps. The most recent work (Adamov et al., 2025) managed to bring ML LAM closer to practical scenarios by using ML-based forecasts as LBC. They needed, however, to include future boundary forcing i.e. a state of local boundary conditions from time $t + 1$, making it easier for model to predict region of interest. In other approaches stretched-grid models were applied to combine global and local meshes of data to achieve promising results on regional forecasts (Nipen et al., 2024). While all aforementioned methods solely depend on training new models rather than using existing ones developed for coarser scale, an approach of fine-tuning pretrained foundation model ClimaX was presented in Munir et al. (2024) effectively downscaling weather forecasts $5.652° \rightarrow 1.40652°$, showing that fine-tuning can

be leveraged to achieve better performance, but failing to demonstrate it on a more challenging fine-scaled grid.

3 DATASETS

3.1 DATA

We pair a global input dataset with a kilometer-scale target dataset for each of two geographic domains. As input we use the IFS HRES t=0 "analysis" product from WeatherBench2 (ECMWF, 2023; Rasp et al., 2024), which provides global fields at 6-hourly resolution on a regular 0.25° grid, comprising atmospheric state variables on 13 pressure levels from 50 to 1000 hPa (geopotential, temperature, specific humidity, wind components, and vertical velocity), surface variables (2 m temperature, 10 m wind components), and mean sea level pressure. However, less accurate, we use it instead of a retrospective reanalysis because it is available in near-real time, keeping the full pipeline operational. Static fields (e.g., topography, land-sea mask) are provided with Aurora. As prediction targets we use HRRR (Dowell et al., 2022) over the Contiguous United States (CONUS) domain ($\approx$ 3 km, hourly) and DANRA (Yang et al., 2026) over the Danish domain (2.5 km, 3-hourly). For both domains we predict 2m temperature, 10m wind components, and mean sea level pressure (we decided to not predict surface pressure because of the well known issue with HRRR data – Appendix F), together with atmospheric temperature, horizontal wind components, and geopotential on 10 pressure levels from 200 to 1000,hPa for CONUS and on 11 pressure levels from 100 to 1000 hPa for the Danish domain. Additionally, specific humidity is predicted only for CONUS region. We restrict our analysis to these pressure levels due to known systematic biases in stratospheric humidity above 150hPa, for further details see Appendix E. We train on 2017–2021 for CONUS and 2017–2019 for the Danish domain. Evaluation is performed on 2022 for CONUS and on 2020 (excluding November and December, which DANRA does not provide) for the Danish domain. The HRRR archive contains WRF-ARW forecasts (+6,h to +48,h), which we use as an operational NWP baseline for model comparison. As an additional, model-independent reference, we use HadISD (Dunn et al., 2012), a quality-controlled collection of in-situ surface station observations over the CONUS region ($\approx$ 1,800 stations), exclusively for evaluation in 2022. Detailed dataset descriptions are provided in Appendix B.

3.2 HRRR AND DANRA REGRIDDING.

Both HRRR and DANRA datasets are originally defined using the LCC curvilinear projection, whereas the IFS HRES dataset utilizes a 0.25° latitude-longitude grid. For our analysis, the HRRR and DANRA datasets are reprojected onto a regular latitude–longitude grid aligned with the IFS HRES coordinate system, while preserving their native high spatial resolution.

The regridding process was carried out in a physically grounded manner, preserving both spatial and physical consistency for each variable. Intensive variables were treated pointwise using bilinear interpolation, extensive variables were regridded to conserve totals, and wind variables were decomposed into their orthogonal vector components. This careful, variable-aware handling of the regridding process is crucial for preserving meaningful fine-scale information without introducing spurious artifacts, and it ensures that the resulting dataset is suitable for super-resolution modelling. Additional implementation details along with the calculation of two-step regridding error are provided in Appendix C and D.

4 METHOD

4.1 PROBLEM FORMULATION

We build on the state-of-the-art Aurora foundation model, which learns generalizable representations of global weather and climate dynamics from diverse reanalysis, forecast, and climate data. Rather than modifying its pretrained backbone, we use Aurora as a fixed simulator and adapt its decoder to produce outputs at the desired spatial resolution and 6-hour temporal resolution. We extend the model with specialized output heads for each target variable, enabling each head to leverage shared features for domain-specific tasks.

Aurora encodes global atmospheric dependencies at the level of the backbone $\mathcal{E}$AUR, which processes two consecutive IFS HRES analysis fields at $t-6$h and $t0$h as input to produce a latent embedding for the $+6$h forecast horizon. Patch embeddings $\mathbf{e}i = \mathcal{E}\text{AUR}(\mathbf{X}t-6h, \mathbf{X}_{t0h})_i \in \mathbb{R}^D$ (with $D = 1024$, where index i corresponds to a $1^\circ \times 1^\circ$ spatial patch) are computed jointly and contextualized through cross-patch interactions. As a result, each embedding represents a local spatial region conditioned on large-scale and non-local atmospheric structure learned during pretraining. Our decoder heads then map these embeddings directly to high-resolution outputs at 0.025° resolution without modifying the pretrained backbone.

In contrast, Aurora's decoder $\text{AUR-MLP-4-0}(\mathbf{e}_i)$ factorizes over patches: given these contextualized embeddings, outputs are produced independently for each patch. This separation between global coupling in the backbone and patch-local decoding not only allows decoder heads to be trained as lightweight, modular adapters without sacrificing access to global weather dynamics, but also allows to have equally accurate outputs in the whole prediction domain, without the drop of quality along boundaries. In principle, this patch-local decoding interface also allows a single decoder head to be trained on patches drawn from multiple geographic regions or datasets. Such multi-regional training proves effective, with no substantial increase in loss. We provide examples of these training runs in Appendix J.

We introduce a family of MLP-S-L decoder heads that map Aurora's 1024-dimensional embeddings directly to an $S \times S$ output defined as:

$$\text{MLP-S-L}(\mathbf{e}_i) = \text{Reshape}_{S\times S}\left(\mathbf{W}_{\text{out}} \cdot \sigma\left(\mathbf{W}_L \cdots \sigma\left(\mathbf{W}_1 \mathbf{e}_i\right)\right)\right),$$

where $\mathbf{W}_1, \ldots, \mathbf{W}_L \in \mathbb{R}^{D\times D}$ are hidden layer weight matrices, $\mathbf{W}_{\text{out}} \in \mathbb{R}^{S^2\times D}$ is the output projection, and $\sigma = \text{GELU}$. The special case $L = 0$ recovers a linear decoder without hidden layers: $\text{MLP-S-0}(\mathbf{e}_i) = \text{Reshape}_{S\times S}\left(\mathbf{W}_{\text{out}}\mathbf{e}_i\right)$. In both cases, the output is a spatial field at the head's native resolution $S \times S$.

For heads with native resolution $S < 40$, we upsample to the target 40×40 grid using two interpolation schemes:

$$\uparrow_{40}^{\text{NN}}\colon \mathbb{R}^{S\times S} \to \mathbb{R}^{40\times40}, \quad \left(\uparrow_{40}^{\text{NN}}(\mathbf{Y})\right)_{i,j} = \mathbf{Y}_{\lfloor i\cdot S/40\rfloor, \lfloor j\cdot S/40\rfloor},$$

$$\uparrow_{40}^{\text{Bicubic}}\colon \mathbb{R}^{S\times S} \to \mathbb{R}^{40\times40}, \quad \left(\uparrow_{40}^{\text{Bicubic}}(\mathbf{Y})\right)_{i,j} = \sum_{m=-1}^{2}\sum_{n=-1}^{2} w_{i,j}(m,n)\cdot \mathbf{Y}_{m,n},$$

where $w_{i,j}(m,n)$ are bicubic interpolation weights computed from a 4×4 neighborhood around position $(i \cdot S/40, j \cdot S/40)$. Both operations are non-learnable and applied uniformly before parallel composition. We denote the chosen upsampling method as $\uparrow_{40}^{*} \in \{\uparrow_{40}^{\text{NN}}, \uparrow_{40}^{\text{Bicubic}}\}$.

Decoder heads can be combined through parallel composition pattern:

$$(\text{MLP-S-L}_1 \parallel \text{MLP-S-L}_2)(\mathbf{e}_i) = (\text{MLP-S-L}_1(\mathbf{e}_i)) + (\text{MLP-S-L}_2(\mathbf{e}_i)),$$

where all heads first have to be upsampled to a common $S = 40$ resolution using deterministic upsampling $\uparrow_{40}^{*}$.

Following NWP initialization workflows, we incorporate high-resolution orography field $\mathbf{O}_h \in \mathbb{R}^{40\times40}$ through two mechanisms:

$$\text{MLP}_{\text{H}}\text{-S-L}(\mathbf{e}_i, \mathbf{O}_h) = \text{MLP}_{\text{H}}\text{-S-L}\left([\mathbf{e}_i; \text{vec}(\mathbf{O}_h)]\right),$$
$$\text{MLP}_{\text{LH}}\text{-S-L}(\mathbf{e}_i, \mathbf{O}_h) = \text{MLP}_{\text{LH}}\text{-S-L}\left([\mathbf{e}_i; \text{MLP}_{\text{oro}}(\text{vec}(\mathbf{O}_h))]\right),$$

where $\text{MLP}_{\text{oro}} : \mathbb{R}^{1600} \to \mathbb{R}^D$ is a shallow projection network with one hidden layer of width 1024, and $[\cdot;\cdot]$ denotes concatenation.

To isolate the contribution of latent-space representations, we define also a $\text{MLP}_{\text{P}}\text{-S-L}(\mathbf{p}_i)$ variant operating directly on Aurora's native 4×4 output patches $\mathbf{p}_i = \text{AUR-MLP-4-0}(\mathbf{e}_i) \in \mathbb{R}^{4\times4}$ produced by the pretrained linear decoder. This formulation represents a learnable post-processing module that refines Aurora's predictions without accessing the full embedding. $\text{MLP}_{\text{P}}\text{-S-L}$ is combined with Aurora's original decoder through sequential composition:

$$\text{AUR-MLP-4-0}(\mathbf{e}_i) + \text{MLP}_{\text{P}}\text{-S-L}(\mathbf{e}_i) = \text{MLP}_{\text{P}}\text{-S-L}\left(\text{AUR-MLP-4-0}(\mathbf{e}_i)\right).$$

This model is meant to provide direct comparison of the same decoder trained once on latent-space embeddings and once on decoded images.

4.2 Training Objective

Following Aurora's design each variable-specific decoder is trained independently using mean absolute error over all spatial locations in the target domain. Decoder parameters are optimized to minimize the average absolute difference between predictions and high-resolution targets from regional NWP simulations. We do not apply spatial loss weighting, as our regional domains are geographically compact with limited latitudinal variation.

4.3 Baselines

We compare our approach against multiple baseline categories. For non-learned interpolation, we evaluate nearest-neighbor ($\mathrm{IFS\,HRES}\uparrow_{40}^{\mathrm{NN}}$) and bicubic interpolation ($\mathrm{IFS\,HRES}\uparrow_{40}^{\mathrm{Bicubic}}$) of IFS HRES, plus Aurora's pretrained linear decoder with NN upsampling ($\mathrm{AUR\text{-}MLP\text{-}4\text{-}0}\uparrow_{40}^{\mathrm{NN}}$) and its HRRR fine-tuned alternative. For learned patch-based post-processing, we test composite baselines applying learned mappings to Aurora's 4×4 output patches: $\mathrm{AUR\text{-}MLP\text{-}4\text{-}0} + \mathrm{MLP_P\text{-}40\text{-}1}$ (non-linear MLP) and $\mathrm{AUR\text{-}MLP\text{-}4\text{-}0} + \mathrm{MLP_P\text{-}40\text{-}0}$ (linear), isolating the contribution of latent-space versus decoded-output operations.

4.4 Experimental setup

We evaluate our approach on both gridded and in-situ datasets to assess downscaling accuracy and generalization across complementary settings. IFS HRES provides low-resolution inputs, while HRRR and DANRA serve as high-resolution ground truth for training. Evaluation is performed at a +6h lead time on (i) the HRRR grid, enabling dense, spatially aligned comparisons, and (ii) HadISD in-situ station observations, which test robustness against real-world measurement sparsity and representativeness errors. Evaluation on both datasets is performed using RMSE for near-surface temperature (T_2), horizontal wind components at 10 meters (u_{10}, v_{10}), and mean sea level pressure ($MSLP$) together with atmospheric temperature, horizonal wind components, geopotential and specific humidity (for HRRR). We also investigate rollout performance of our models to WRF-ARW NWP model to assess the stability of the approach.

All experiments were conducted on NVIDIA H100 GPUs. A single training run takes approximately 5 hours on 4 H100 GPUs (or 20 hours on a single GPU). Inference of the full pipeline, including Aurora's backbone, on a single GPU takes approximately 57 seconds for a +6 h prediction and 65 seconds for a +24 h rollout. However, this runtime is dominated by model loading, which takes nearly 50 seconds. The inference itself takes only 5.5 seconds for a +6 h prediction and 12.74 seconds for a +24 h rollout. To reduce I/O overhead and memory footprint during training, we operate directly on intermediate embeddings. The models were implemented in PyTorch (Paszke et al., 2019) and trained with AdamW (Loshchilov & Hutter, 2017) using an initial learning rate of 0.001 and cosine annealing with a 1000-step warm-up and a minimum learning rate of 0.0001. We use a batch size of 16, a weight decay of 0.01, and a mean absolute error (MAE) loss. To ensure training stability we use EMA (Morales-Brotons et al., 2024) with decay of 0.999. We do not apply spatial loss weighting, as we focus on local regional forecasting where the domains are geographically close and latitudinal effects are limited. All computations were performed in bfloat16 mixed precision. Additional details on data scale and preprocessing are available in in Appendix K.

5 Results

Table 1 summarizes the results for our defined methods.

Non-learned interpolation baselines ($\mathrm{IFS\,HRES}\uparrow_{40}^{\mathrm{NN}}$, $\mathrm{IFS\,HRES}\uparrow_{40}^{\mathrm{Bicubic}}$) exhibit the largest errors across all variables, confirming that purely geometric upscaling is insufficient to recover fine-scale atmospheric structure. Bicubic interpolation provides marginal improvements over nearest-neighbor. While Aurora's pretrained linear decoder with upsampling further reduces errors, indicating that learned representations add value even without task-specific fine-tuning. The HRRR-biased variant, fine-tuned on HRRR fields achieves further improvements by correcting systematic biases between the IFS HRES input and HRRR target datasets that remain unaddressed in the base Aurora model.

Table 1: Performance evaluation comparison between HRRR grid and HadISD in-situ observations using RMSE at +6h lead time. T_2 denotes 2-m air temperature (°C), u_{10} and v_{10} wind component speeds (m/s), and MSLP mean sea level pressure (hPa).

	HRRR				HadISD		
Model	T_2	u_{10}	v_{10}	$MSLP$	T_2	u_{10}	v_{10}
IFS HRES$\uparrow_{40}^{\text{NN}}$	1.9963	1.7120	1.7221	2.2526	2.2727	1.7820	1.8626
IFS HRES$\uparrow_{40}^{\text{Bicubic}}$	1.9134	1.6621	1.6745	2.2432	2.2137	1.7401	1.8261
AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$	1.8865	1.5927	1.5956	2.2363	2.2056	1.6952	1.7808
AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$ (HRRR biased)	1.6112	1.4948	1.5031	0.8073	1.9637	1.6045	1.6625
AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$ +MLP$_{\text{P}}$-40-0	1.7242	1.5407	1.5442	1.7746	2.1024	1.6364	1.7139
AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$ +MLP$_{\text{P}}$-40-1	1.6860	1.4968	1.5131	1.3745	2.0742	1.6002	1.6759
MLP-40-0	1.8297	1.5165	1.5404	1.1460	2.1713	1.6162	1.6850
MLP-40-1	1.2276	1.3113	1.3425	0.5717	1.6431	1.4946	1.5483
AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$							
| MLP-40-0	1.4951	1.4533	1.4684	0.7582	1.8862	1.5787	1.6382
| MLP-40-1	1.2183	1.3087	1.3412	0.5682	1.6328	1.4962	1.5499
| MLP-40-0 | MLP-8-0$\uparrow_{40}^{\text{NN}}$	1.4948	1.4532	1.4682	0.7579	1.8856	1.5787	1.6381
| MLP-40-1 | MLP-8-1$\uparrow_{40}^{\text{NN}}$	1.2055	1.2954	1.3297	0.5694	1.6243	1.4901	1.5435
| MLP-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$	1.1832	1.2731	1.3024	0.5653	1.5886	1.4796	1.5303
| MLP$_{\text{H}}$-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$	1.1753	1.2697	1.3019	0.5721	1.5898	1.4789	1.5296
| MLP$_{\text{LH}}$-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$	**1.1743**	1.2644	1.2952	0.5664	**1.5793**	**1.4759**	**1.5273**
AUR-MLP-4-0$\uparrow_{40}^{\text{Bicubic}}$							
| MLP$_{\text{LH}}$-40-2 | MLP-8-2$\uparrow_{40}^{\text{Bicubic}}$	1.1823	**1.2630**	**1.2942**	**0.5645**	1.6691	1.4931	1.5440

Table 2: Performance evaluation of our approach against WRF-ARW, StormCast 5-member Ensemble and Aurora's $0.25°$ base model. Results are presented for the HRRR grid and HadISD in-situ observations using RMSE at +6h lead time. T_2 denotes 2-m air temperature (°C), u_{10} and v_{10} wind component speeds (m/s), and $MSLP$ mean sea level pressure (hPa).

	HRRR Grid				HadISD (in-situ)		
Model	T_2	u_{10}	v_{10}	$MSLP$	T_2	u_{10}	v_{10}
AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$							
| MLP$_{\text{LH}}$-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$	**1.1743**	**1.2644**	**1.2952**	**0.5664**	**1.5793**	**1.4759**	**1.5273**
WRF-ARW	**1.1749**	1.3784	1.4248	0.7475	1.8832	1.7663	1.8229
Aurora $0.25°$	1.8865	1.5927	1.5956	2.2362	2.2056	1.6952	1.7808
StormCast 5-member Ensemble	2.2355	1.5715	1.6308	1.3786	2.4279	1.8346	1.8939
	DANRA Grid				HadISD (in-situ)		
AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$							
| MLP$_{\text{LH}}$-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$	0.6759	**0.8988**	**0.8749**	**0.2604**	**1.1430**	1.7019	**1.5711**
Aurora $0.25°$	1.0796	1.2516	1.1819	0.3839	1.5858	**1.5927**	1.5956
ML-LAM(DANRA)	**0.5759**	1.0533	1.1524	0.9900	1.2042	1.9964	2.0103

Patch-based post-processing methods that apply learned mappings to Aurora's native 4×4 output patches show mixed results. Both learned patch-based mappings (MLP$_{\text{P}}$-40-0, MLP$_{\text{P}}$-40-1) improve over standard upsampling but underperform compared to Aurora's fine-tuned decoder with nearest-neighbor interpolation. This indicates that operating on decoded patch outputs rather than full latent embeddings limits the capacity to recover fine-scale atmospheric structure.

Embedding-based models that operate directly on Aurora's latent representations show clear advantages over all defined baselines. The only small exception is linear decoder, even though

it has better scores for MSLP and wind components, is not able to capture temperature correctly. Simple one-head non linear $\mathrm{MLP_P}$-40-0 exceeds the best performing baselines reducing errors across all variables. Importantly, showing much improved error in comparison to direct $\mathrm{MLP_P}$-40-1 counterpart decoding from images instead of latent space. Further increasing models depth gives another improvements, however, according to our tests adding more than two layers yields no further improvements. This confirms that accessing the full embedding, rather than post-processing Aurora's decoded outputs, enables more effective downscaling.

Multi-head architectures. Adding the $\mathrm{MLP_P}$-8-L head provides small but consistent improvements across evaluated configurations, though increasing head depth has a substantially larger impact on final performance. The best-performing configuration without orography is (MLP-40-2 | MLP-8-2), achieving the lowest errors for most variables. Incorporating MLP-encoded orography into the decoder further improves temperature and wind performance while slightly increasing pressure error. In contrast, directly stacking raw orography with the same architecture yields minimal improvement, indicating that even a Lightweight learned projection of orographic features is consistently more effective than using raw inputs, but some of the gains are within the seeding variability (Appendix J). Changing interpolation mode from $\uparrow_{40}^{\mathrm{NN}}$ to $\uparrow_{40}^{\mathrm{Bicubic}}$ gives minor improvements for wind and pressure, but visibly lower the score for temperature. Incorporating a lightweight, pre-trained decoder yields consistent improvements for shallow architectures but provides limited benefit for deeper configurations (Appendix J). For sample prediction plots of our best performing model we refer to Appendix L.

Results on HadISD in-situ observations (Table 1) closely mirror trends observed on the HRRR grid, despite the increased difficulty of sparse and potentially noisy point measurements. The performance gap between single-head and multi-head configurations remains consistent with gridded evaluations, with deeper heads providing the largest improvements for temperature and wind components. Importantly, the relative ranking of models is largely preserved between gridded and in-situ evaluations, indicating that improvements generalize across evaluation methodologies rather than reflecting grid-specific artifacts. The only visible exception is model with $\uparrow_{40}^{\mathrm{Bicubic}}$ upscaling which underperforms compared to HRRR. Baseline models show comparable degradation patterns when moving from grid to point evaluations, with IFS HRES and Aurora exhibiting similar error increases. Additional analysis of error distributions across station subsets is provided in Appendix I.

Comparison with other methods is shown in Table 2, which compares our strongest configuration against WRF-ARW, StormCast 5-member Ensemble (Pathak et al., 2024), Aurora's $0.25°$ base model, and the graph-based model of Adamov et al. (2025), trained on DANRA dataset (denoted as ML-LAM(DANRA)). On HadISD, our approach outperforms WRF-ARW across temperature and both wind components, while both StormCast Ensemble and Aurora exhibit higher errors for all variables. On the HRRR grid, our model achieves the best scores for wind components and pressure, and performs comparably to WRF-ARW on temperature. The deterministic prediction achieves lower pointwise RMSE than the StormCast ensemble statistic under this evaluation protocol across all variables. According to Appendix J, we observe no degradation from regridding; nevertheless, we report both StormCast and our results on LCC and latitude–longitude (°) grids for fair comparison. On the DANRA grid, our model substantially improves over Aurora's base model across all variables and offers better wind and pressure predictions than ML-LAM(DANRA), with slightly higher temperature error. At HadISD stations, our model achieves lower temperature and v-wind RMSE than both Aurora and ML-LAM, while u-wind error remains higher than Aurora's native model but lower than ML-LAM(DANRA). Appendix 20 includes rollout comparison to ML-LAM(DANRA) on surface and atmospheric variables.

Rollout results are shown in Figure 2, comparing relative difference (the lower the better) between RMSE values of AUR-MLP-4-0 $\uparrow_{40}^{\mathrm{Bicubic}}$| $\mathrm{MLP_{LH}}$-40-2 | MLP-8-2 $\uparrow_{40}^{\mathrm{Bicubic}}$ model against WRF-ARW predictions over a 48-hour horizon. Even though $\uparrow_{40}^{\mathrm{Bicubic}}$ model slightly underperforms comparing to $\uparrow_{40}^{\mathrm{NN}}$ on raw RMSE, we decided to showcase its results as according to Appendix H it is more physically consistent. Results are averaged over the full year 2022 and presented for three surface variables, mean sea level pressure and five atmospheric variables. These results indicate much better rollout stability of our method in comparison to WRF-ARW forecasts.

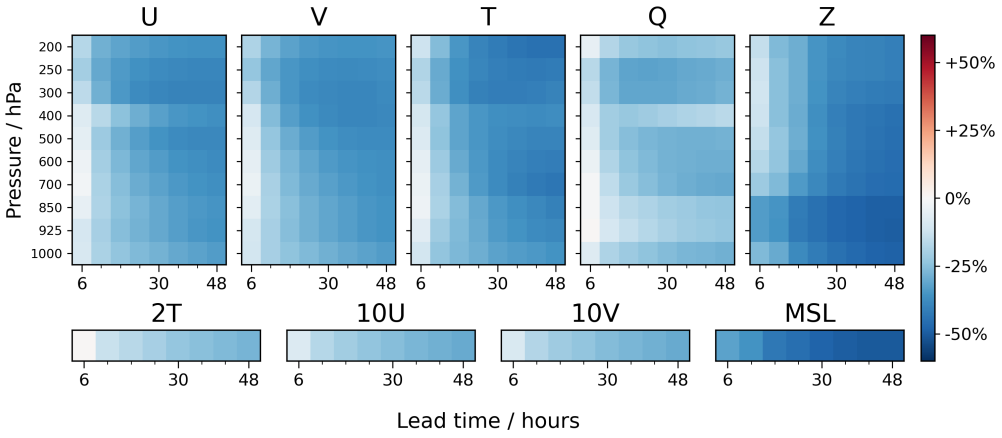


Figure 2: Performance comparison over a 48-hour horizon for our model relative to the WRF-ARW.

Physical consistency checks. Beyond standard verification metrics, we evaluate whether our model produces physically consistent predictions by testing fundamental physical relationships. We examine geostrophic and ageostrophic wind components separately, along with their ratio to assess the balance between large-scale balanced flow and unbalanced ageostrophic motions. We verify hydrostatic balance via the hypsometric equation residual and examine the kinematic ratio of divergence to vorticity. We further analyze multi-scale structure through horizontal energy spectra and evaluate wind vector accuracy using paired u/v component diagnostics (vector RMSE, mean vector error magnitude, wind-speed RMSE, and vector correlation). Finally, we compute bias profiles across all pressure levels for temperature, humidity, wind components, and geopotential. These diagnostics confirm that our learned models preserve key physical properties consistent with atmospheric dynamics. Full details and analysis are provided in Appendix H.

6 Limitations

Our evaluation is restricted to forecast lead times up to +48h, leaving longer-range predictive skill and error growth beyond this horizon uncharacterised. The patch-based architecture of Aurora provides practical benefits, including lightweight decoder design and the ability to model multiple geographic regions within a unified framework. However, this design constrains architectural choices. Addressing output discontinuities at patch boundaries requires specific upsampling and composition strategies. This is particularly important for regions with low spatial gradients, where patching artifacts become apparent, thus becoming visible in energy spectra analysis. Although we empirically verify that key physical properties remain consistent, our evaluation relies on HRRR as ground truth for upper-air variables, extending validation to independent upper-air observational datasets remains an important direction for future work. Consequently, our outputs should be interpreted as complementary to, rather than replacements for, established numerical weather prediction systems and expert analysis.

7 Conclusion and future directions

The proposed approach builds on Aurora's foundation model backbone, which is designed to capture multi-scale physical structure, and combines it with simple downscaling heads to produce high-resolution outputs. Our results show an advantage of this approach over the commonly used WRF-ARW NWP model, defined baselines and StormCast. In addition, the use of a frozen backbone makes training efficient, requiring less than 20 hours on a single H100 GPU, enabling fast inference. Future work includes (1) validating model generalization beyond the training domain using diverse datasets, and (2) enhancing the architecture with multi-scale attention heads for improved spatial reasoning. (3) We plan to investigate the influence of physics-informed losses and constraints on our model and assess their impact on the physical consistency checks in our evaluation suite and 4) utilize a newly introduced Aurora1.5 model (Weyn et al., 2026).

AI USE STATEMENT

In this work, we used generative AI tools for text editing and rephrasing, including rewriting individual sentences and paragraphs to improve clarity and conciseness. We also used generative AI tools for generating code mock-ups and assisting with code refactoring; however, the substantial portion of our codebase was written and designed by the authors. We have not used generative AI tools for formulating research ideas, conducting literature review or drafting full sections of the paper. All scientific claims, experimental design, and methodological decisions were made by the authors. Tasks related to data collection, model training, and evaluation were performed by the authors without AI assistance.

We have reviewed all AI-assisted work. To ensure accuracy, all AI-generated or AI-modified content underwent manual verification by at least two authors.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not raise significant ethical concerns under the ICLR Code of Ethics. Our research uses only publicly available meteorological datasets (HRRR, DANRA, IFS HRES, HadISD) that do not contain personal, sensitive, or privately identifiable information. No human subjects were involved in this study.

The methods and models developed in this work are intended for weather prediction and downscaling applications, which serve beneficial societal purposes such as improved forecasting and climate monitoring. We are not aware of any dual-use concerns or potential for harmful applications arising from this research.

REPRODUCIBILITY STATEMENT

We consider reproducibility a fundamental requirement for responsible research. To this end, we provide comprehensive materials to enable independent reproduction of all results presented in this paper. Our anonymous code repository is available at: `https://anonymous.4open.science/r/iclr2027-inference-4440`. All datasets used in this work are publicly available. Broad appendix B provides complete descriptions of data processing steps, data splits, and any filtering applied. Numerical results for the best model reported in tables and figures can be regenerated by running the provided evaluation scripts on the pre-trained checkpoints. We have verified that the code produces consistent results across multiple runs with fixed random seeds.

REFERENCES

Weather prediction by numerical process. by Lewis F. Richardson. cambridge (university press), 1922. 4°. pp. xii + 236. 30s. *Quarterly Journal of the Royal Meteorological Society*, 48 (203):282–284, 1922. doi: https://doi.org/10.1002/qj.49704820311. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49704820311.

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, 2025. URL https://arxiv.org/abs/2504.09340.

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.

Jacqueline E. Bland, Suzanne Louise Gray, John Methven, and Richard G. Forbes. Characterising extratropical near-tropopause analysis humidity biases and their radiative effects on temperature forecasts. *Quarterly Journal of the Royal Meteorological Society*, 147:3878 – 3898, 2021. URL https://api.semanticscholar.org/CorpusID:238722913.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641:1180–1187, May 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

Shilei Cao, Hehai Lin, Jiashun Cheng, Yang Liu, Guowen Li, Xuehe Wang, Juepeng Zheng, Haoyuan Liang, Meng Jin, Chengwei Qin, Hong Cheng, and Haohuan Fu. Task-adaptive parameter-efficient fine-tuning for weather foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eFExhM3tKr.

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, 2022. doi: 10.1175/WAF-D-21-0151.1. Available as part of the AWS Open Data Program. Data are in the public domain; no copyright restrictions. Users should provide appropriate attribution to NOAA.

Peter D Dueben and Peter Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10):3999–4009, 2018.

R. J. H. Dunn, K. M. Willett, P. W. Thorne, E. V. Woolley, I. Durre, A. Dai, D. E. Parker, and R. S. Vose. Hadisd: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, 8(5):1649–1679, October 2012. ISSN 1814-9332. doi: 10.5194/cp-8-1649-2012. URL http://dx.doi.org/10.5194/cp-8-1649-2012. Data are available without charge under the UK Met Office Non-Commercial Government Licence.

ECMWF. *IFS Documentation CY48R1 - Part VIII: Atmospheric Composition*. Number 8. ECMWF, 06/2023 2023. doi: 10.21957/749dc09059.

ECMWF. Reintroducing the analysis of humidity in the stratosphere, 2025. URL https://www.ecmwf.int/en/newsletter/183/earth-system-science/reintroducing-analysis-humidity-stratosphere.

H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. Era5 hourly data on single levels from 1940 to present, 2018. Accessed on 02-APR-2026.

James R. Holton, Peter H. Haynes, Michael E. McIntyre, Anne R. Douglass, Richard B. Rood, and Leonhard Pfister. Stratosphere-troposphere exchange. *Reviews of Geophysics*, 33(4):403–439, 1995. doi: https://doi.org/10.1029/95RG02097. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/95RG02097.

Zaviša I. Janjic. A nonhydrostatic model based on a new approach. *Meteorology and Atmospheric Physics*, 82:271–285, 2003. doi: 10.1007/s00703-001-0587-6.

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, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/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, 2024. URL https://arxiv.org/abs/2406.01465.

Erik Larsson, Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Diffusion-lam: Probabilistic limited area weather forecasting with diffusion, 2025. URL https://arxiv.org/abs/2502.07532.

Fanny Lehmann, Firat Ozdemir, Benedikt Soja, Torsten Hoefler, Siddhartha Mishra, and Sebastian Schemm. Finetuning a weather foundation model with lightweight decoders for unseen physical processes, 2025. URL https://arxiv.org/abs/2506.19088.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Holger R. Maier and Graeme C. Dandy. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*, 15(1):101–124, 2000. ISSN 1364-8152. doi: https://doi.org/10.1016/S1364-8152(99)00007-9. URL https://www.sciencedirect.com/science/article/pii/S1364815299000079.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits, 2024. URL https://arxiv.org/abs/2411.18704.

Muhammad Akhtar Munir, Fahad Shahbaz Khan, and Salman Khan. Efficient localized adaptation of neural weather forecasting: A case study in the MENA region, 2024. URL https://arxiv.org/abs/2409.07585.

National Weather Service and NCEP Environmental Modeling Center. Service change notice 22-68: Correction of the specification of elevation of the great lakes in the high resolution rapid refresh (hrrr). https://www.weather.gov/media/notification/pdf2/scn22-68_hrrr_great_lakes.pdf, 2022. Issued July 19, 2022.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Thomas Nils Nipen, Håvard Homleid Haugen, Magnus Sikora Ingstad, Even Marius Nordhagen, Aram Farhad Shafiq Salihi, Paulina Tedesco, Ivar Ambjørn Seierstad, Jørn Kristiansen, Simon Lang, Mihai Alexe, Jesper Dramsch, Baudouin Raoult, Gert Mertes, and Matthew Chantry. Regional data-driven weather modeling with a global stretched-grid, 2024. URL https://arxiv.org/abs/2409.02891.

Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Graph-based neural weather prediction for limited area modeling, 2023. URL https://arxiv.org/abs/2309.17370.

K. Papagiannaki, K. Lagouvardos, V. Kotroni, and A. Bezes. Flash flood occurrence and relation to the rainfall hazard in a highly urbanized area. *Natural Hazards and Earth System Sciences*, 15(8):1859–1871, 2015. doi: 10.5194/nhess-15-1859-2015. URL https://nhess.copernicus.org/articles/15/1859/2015/.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

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

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, 2024. URL https://arxiv.org/abs/2408.10958.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2024. URL https://arxiv.org/abs/2308.15560.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E Phillips, Romeo Kienzler, Daniela Szwarcman, Vishal Gaur, Rajat Shinde, Rohit Lal, Arlindo Da Silva, Jorge Luis Guevara Diaz, Anne Jones, Simon Pfreundschuh, Amy Lin, Aditi Sheshadri, Udaysankar Nair, Valentine Anantharaj, Hendrik Hamann, Campbell Watson, Manil Maskey, Tsengdar J Lee, Juan Bernabe Moreno, and Rahul Ramachandran. Prithvi wxc: Foundation model for weather and climate, 2024. URL https://arxiv.org/abs/2409.13598.

Yingkai Sha, John S. Schreck, William Chapman, and David John Gagne II. Improving ai weather prediction models using global mass and energy conservation schemes. *Journal of Advances in Modeling Earth Systems*, 17(11):e2025MS005138, 2025. doi: https://doi.org/10.1029/2025MS005138. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025MS005138. e2025MS005138 2025MS005138.

William Skamarock, Joseph Klemp, Jimy Dudhia, David Gill, Dale Barker, Michael Duda, Xiang-Yu Huang, Wei Wang, and Jordan Powers. A description of the advanced research wrf version 3, 06 2008.

William C. Skamarock, Joseph B. Klemp, Jimy Dudhia, David O. Gill, Zhiquan Liu, Judith Berner, Wei Wang, John G. Powers, Michael G. Duda, Dale M. Barker, and Xiang-Yu Huang. A description of the advanced research wrf version 4. Technical Report NCAR/TN-556+STR, National Center for Atmospheric Research, 2019.

Kingsley Ukoba, Kehinde O. Olatunji, Eyitayo Adeoye, Tien-Chien Jen, and Daniel M. Madyira. Optimizing renewable energy systems through artificial intelligence: Review and future prospects. *Energy & Environment*, 35(7):3833–3879, 2024. doi: 10.1177/0958305X241256293. URL https://doi.org/10.1177/0958305X241256293.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Anna Vaughan, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. Aardvark weather: end-to-end data-driven weather forecasting, 2024. URL https://arxiv.org/abs/2404.00411.

Jonathan Weyn, Zekun Ni, Amit Misra, Will Fein, Haiyu Dong, Wessel P. Bruinsma, Richard E. Turner, Matt Corey, Kit Thambiratnam, Kevin White, and Kenji Takeda. Aurora 1.5: Fine-tuning a foundation model for medium-range ensemble weather prediction. July 2026. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2026/07/Aurora_1_5_Paper.pdf.

Jonathan A Weyn, Dale R Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, 2020.

Pengbo Xu, Xiaogu Zheng, Tianyan Gao, Yu Wang, Junping Yin, Juan Zhang, Xuanze Zhang, San Luo, Zhonglei Wang, Zhimin Zhang, Xiaoguang Hu, and Xiaoxu Chen. Yinglong-weather:

Ai-based limited area models for forecasting of non-precipitation surface meteorological variables, 2025. URL https://arxiv.org/abs/2401.16254.

X. Yang, C. Peralta, B. Amstrup, K. S. Hintz, S. B. Thorsen, L. Denby, S. Kamuk Christiansen, H. Schulz, S. Pelt, and M. Schreiner. Danra: the kilometer-scale danish regional atmospheric reanalysis. *Earth System Science Data*, 18(3):2251–2264, 2026. doi: 10.5194/essd-18-2251-2026. URL https://essd.copernicus.org/articles/18/2251/2026/.

A Appendix

B Dataset Details

B.1 High-Resolution Rapid Refresh (HRRR).

The High-Resolution Rapid Refresh (HRRR) (Dowell et al., 2022) dataset is an NWP product generated using the WRF-ARW modeling system with rapid-update data assimilation. It provides three-dimensional atmospheric fields over the CONUS domain at ≈3 km horizontal resolution and hourly temporal resolution, defined on a Lambert Conformal Conic (LCC) projected grid. Although the dataset starts in 2014, early years contain substantial monthly gaps, therefore, we use data from 2017–2022. The analysis fields are produced through frequent cycling data assimilation incorporating diverse observational inputs, yielding dynamically consistent estimates of the atmospheric state. In this work, HRRR serves a role of training dataset, except that 2022 is reserved for evaluation to enable direct comparison with the Aurora and StormCast models. In addition to assimilation data, the HRRR dataset includes accurate, state-of-the-art NWP forecasts produced using the WRF-ARW model, which we use to evaluate our data-driven approach. For simplicity, we refer to the lead-zero (+0 h) fields as HRRR and the 2022 HRRR forecasts (+6 h–+48 h), generated with the WRF-ARW model, as WRF-ARW. We predict 2 m temperature, 10 m wind components, and mean sea level pressure at the surface, together with atmospheric temperature, horizontal wind components, specific humidity, and geopotential on 10 pressure levels from 200 to 1000 hPa (200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa). We omit surface pressure due to a known issue affecting this variable between December 2020 and July 2022, documented in a public note (National Weather Service & NCEP Environmental Modeling Center, 2022). See Appendix F for more details. Our comparison of HRRR and IFS HRES mean sea level pressure does not show the corresponding 2020-2022 discontinuity. We therefore retain mean sea level pressure (MSLP) as the pressure variable used for training and evaluation.

B.2 DANRA.

The DANish regional atmospheric ReAnalysis (DANRA) (Yang et al., 2026) is a kilometer-scale regional reanalysis dataset produced by the Danish Meteorological Institute (DMI), covering Denmark and its surrounding regions over the period from 1990 to 2023. It provides atmospheric fields at 3-hourly analysis times on the DKA domain, an 800×600 grid with a 2.5 km horizontal grid spacing and 65 vertical model levels. The dataset is generated with DMI's operational HARMONIE-AROME model, a non-hydrostatic, convection-permitting NWP system that runs a 3-hourly cycle of data assimilation and short-range forecasting. In each cycle, upper-air observations are assimilated with 3D-Var, near-surface states are updated with multivariate optimal interpolation, and a short-range forecast provides the background state for the next cycle. Lateral boundary conditions are taken from a ERA5 global reanalysis. In our approach, DANRA serves as the high-resolution training target for this domain, playing a role analogous to HRRR for the CONUS region. We predict mean sea level pressure, 2 m temperature, and 10 m wind components at the surface, together with atmospheric temperature, horizontal wind components, and geopotential on 11 pressure levels (100, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa). We use years 2017–2019 for training and 2020 for evaluation, excluding November and December. This evaluation period is a subset of the 2019-10 to 2020-10 period used by Adamov et al. (2025), ensuring no data leakage while enabling direct comparison.

B.3 IFS HRES.

The Integrated Forecasting System (IFS) (ECMWF, 2023) is the operational global numerical weather prediction system of the European Centre for Medium-Range Weather Forecasts (ECMWF), whose high-resolution configuration (HRES) produces global forecasts at a native horizontal resolution of approximately 9 km ($\sim 0.1^\circ$). We use the IFS HRES t=0 "analysis" dataset distributed through WeatherBench2 (Rasp et al., 2024), which stores the initial conditions of the operational HRES forecasts, i.e., the forecast fields at lead time zero rather than the dedicated ECMWF analysis product, spanning the period from 2016 to early 2023. The data are provided at 6-hourly temporal resolution (00, 06, 12, and 18 UTC) on a regular rectilinear grid with a spatial resolution of 0.25° (approximately 28 km at the equator), which is coarser than the native model resolution. Set of variables used in our approach consists of atmospheric state variables on 13 pressure levels from 50 to 1000 hPa (geopotential, temperature, specific humidity, horizontal wind components) and surface variables (2 m temperature, 10 m wind, mean sea level pressure). Static variables (e.g., land–sea mask or orography) are not included in the dataset, so we use the static fields provided with Aurora. The atmospheric state is reconstructed using four-dimensional variational data assimilation (4D-Var), which combines a numerical weather prediction model with temporally distributed observations to produce a physically consistent estimate of the atmospheric state. Unlike retrospective reanalysis products, these initial states become available in near-real time as part of ECMWF's operational production, which enables an operational setup. In our approach, IFS HRES t=0 serves as input for extracting latent $1^\circ \times 1^\circ$ spatial embeddings from Aurora's backbone, from which weather variables are predicted at higher spatial resolution. We use years 2017–2021 for training and 2022 for testing.

B.4 HadISD.

The Hadley Integrated Surface Dataset (HadISD) (Dunn et al., 2012) is a quality-controlled collection of in-situ surface meteorological observations derived from global weather station records. For this study, we focus on two subsets of stations located within the CONUS and DANRA regions. These subsets comprise approximately 1,800 and 500 stations, respectively, with irregular spatial distributions. HadISD provides point-based measurements of key near-surface variables and serves as an independent observational reference that is not influenced by model dynamics. In our approach, this dataset is used exclusively as a benchmark for evaluation, all the assessments are conducted on data from the year 2022 for CONUS and from January through October 2020 for DANRA. The evaluation metric (RMSE) is computed between station measurements and nearest 0.025° grid points, with station values averaged within grid cells for intensive variables.

C Regridding - detailed description

Regridding must be physically aware: different atmospheric variables behave differently under spatial averaging or interpolation. For example, radiative fluxes or precipitation rates are reported per unit area or as equivalent depth rates so convservative regridding must be used to preserve their area-weighted integral over a region. In contrast, variables such as temperature or pressure are intensive, defined pointwise and independent of cell size. Wind-related variables pose an additional challenge, the direction is circular and horizontal wind components must be expressed on a common coordinate basis. To ensure that our downscaling model learns meaningful relationships rather than artifacts of inconsistent interpolation, the regridding strategy must respect these physical distinctions. Depending on the task, we map fields from their native grids to regular grids or from regular grids back to native grids. For each transformation, we select an interpolation method suited to the variable and, where necessary, express wind components in a common coordinate system before interpolating them. The resulting datasets pair IFS fields with HRRR or DANRA fields over their respective regions for the selected variables. Both HRRR and DANRA are regridded from their native Lambert Conformal Conic grids. An example is shown in Figure 3.

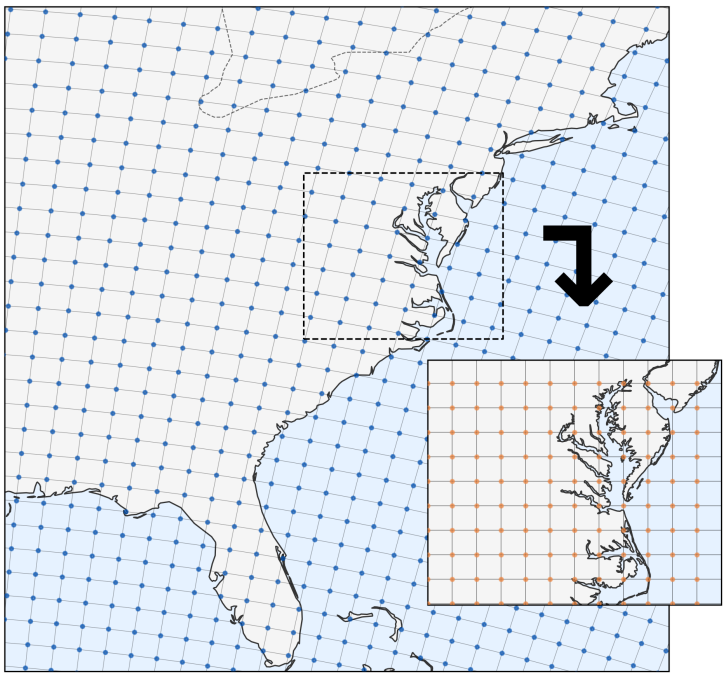


Figure 3: Illustration of the spatial transformation applied during regridding.

D REGRIDDING – ROUNDTRIP ERROR CALCULATION

The data is first interpolated from the native grid to the regular grid, and subsequently interpolated back to the original native grid using the corresponding reverse mapping.

This bidirectional interpolation framework enables isolation of the numerical error associated solely with the regridding procedure, avoiding discrepancies that would arise from comparisons between independent datasets. The resulting error represents the combined effects of information loss during the forward interpolation and additional distortions introduced during the backward interpolation. The reconstructed field is therefore directly compared to the original field on the native grid cells where the reconstruction is valid.

The round-trip interpolation error provides a conservative estimate of the regridding uncertainty, as it reflects the cumulative effects of both forward and backward mappings. In the training workflow, only the forward interpolation is applied; therefore, the effective error is expected to be smaller in practice.

The resulting errors are summarized in Tables 3 and 4. Across fields, their magnitude follows the spatial roughness of each quantity. 2-m air temperature, shaped by terrain-driven gradients, shows the largest round-trip error at 0.248 ± 0.051 °C, with the 10-m wind components close behind near 0.18 m/s. Both remain small compared with the natural variability of the fields. Mean sea-level pressure varies on scales of hundreds of kilometers, far exceeding the grid spacing, so neighboring grid points hold nearly identical values. Its round-trip error of 0.0018 hPa is consistent with its weak terrain-related variations. Upper-air errors are an order of magnitude below their surface counterparts, in line with the weaker horizontal gradients away from the ground. Temperature reaches 0.010 K and winds 0.011 m/s, while the geopotential error corresponds to about 1 cm of equivalent height. Specific humidity is the exception, with a standard deviation across levels that exceeds its mean error. Table **??** explains this pattern. The per-level errors span nearly five orders of magnitude and decay monotonically with height, peaking in the moist lower troposphere and shrinking toward the dry upper levels. This mirrors the behavior of humidity itself, which falls by roughly four orders of magnitude from the boundary layer to the lower stratosphere as rising air is freeze-dried at the cold tropical tropopause (Holton et al., 1995). The errors thus scale with the field rather than with regridding quality. For temperature and winds, the tight spread across the 13 pressure levels suggests

that regridding quality is comparable throughout the troposphere and lower stratosphere, with no clear degradation at upper levels, where geometric grid distortion might amplify interpolation errors. Taken together, the errors remain small relative to the signals the downstream model aims to capture, indicating that regridding is unlikely to be the limiting factor in the preprocessing chain.

Table 3: Roundtrip regridding error for surface variables. Reported values are mean $\pm$ standard deviation. T_2 denotes 2-m air temperature (°C), u_{10} and v_{10} the 10-m wind components (m/s), and $MSLP$ mean sea-level pressure (hPa).

Variable	RMSE	MAE
T_2	0.2477 ± 0.0512	0.1172 ± 0.0241
u_{10}	0.1835 ± 0.0500	0.0887 ± 0.0248
v_{10}	0.1819 ± 0.0529	0.0899 ± 0.0276
$MSLP$	0.0018 ± 0.0001	0.0013 ± 0.0001

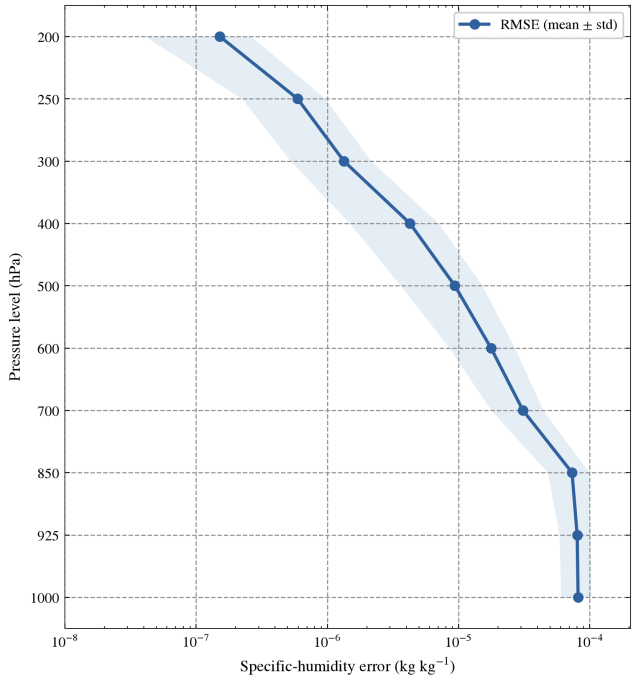


Figure 4: Round-trip regridding error for specific humidity (q) at each pressure level. Reported values are RMSE $\pm$ standard deviation.

E Specific Humidity Stratosphere Issue

Analysis of specific humidity in the HRRR dataset revealed sudden distribution shifts at upper atmospheric levels, particularly above 150 hPa. These discontinuities reflect well-documented challenges in how numerical weather prediction models represent stratospheric humidity. For over two decades (2000–2024), ECMWF's IFS did not directly assimilate observational data for humidity in the stratosphere. Although ECMWF's 4D-Var system incorporated a pseudo-relative humidity control variable from 2003 onwards, radiosonde humidity data were not assimilated above the tropopause (typically 300–100 hPa), leaving stratospheric humidity fields without direct observational constraints (ECMWF, 2025). Instead, stratospheric humidity was largely governed by the model's physics and previous forecasts. This approach was adopted after an earlier attempt to assimilate stratospheric humidity produced excessively moist values, causing systematic forecast errors. Subsequent work found that IFS specific humidity in the lowermost stratosphere reached approximately 170% of observed values, corresponding to a moist bias of about 70%, around 1 km above the tropopause in a

Table 4: Roundtrip regridding error for upper-air variables, averaged over the 10 pressure levels. The reported uncertainty is the standard deviation across pressure levels.

Variable	RMSE	MAE
T [K]	0.01000 ± 0.00105	0.00723 ± 0.00084
q [kg/kg]	0.00002 ± 0.00003	0.00001 ± 0.00001
u [m/s]	0.01086 ± 0.00061	0.00818 ± 0.00041
v [m/s]	0.01073 ± 0.00063	0.00809 ± 0.00044
z [m^2/s^2]	0.10421 ± 0.02303	0.06389 ± 0.00925

comparison with North Atlantic radiosonde observations (Bland et al., 2021). The problem stems from sharp humidity gradients near the tropopause, where small vertical errors produce large differences in humidity values and from limited observational coverage at these altitudes. Measurement uncertainty itself is substantial in the stratosphere, reaching 50–100% of mean specific humidity values more than 2 km above the tropopause. These systematic issues in the input data propagate into data-driven models trained on NWP outputs.

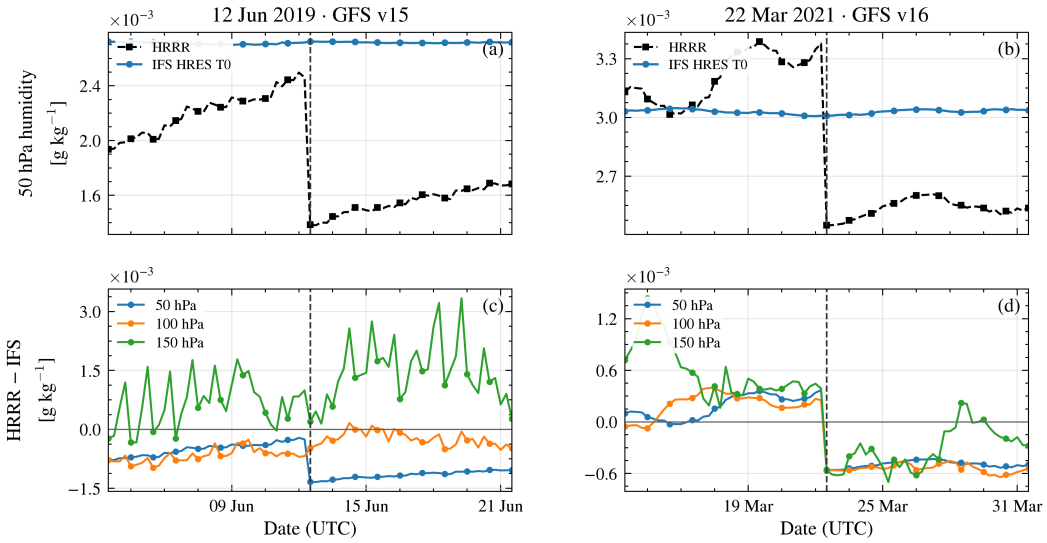


Figure 5: Specific humidity discontinuities in HRRR around the GFS v15 and v16 operational upgrades. Top panels show mean humidity at 50 hPa; bottom panels show HRRR–IFS HRES biases at 50, 100, and 150 hPa. Dashed vertical lines mark the upgrade dates. GFS supplies boundary conditions to RAP, which supplies HRRR boundary conditions.

F HRRR Great Lakes Elevation Issue

During exploratory analysis of the HRRR surface pressure field, we identified a temporally localized bias over the Great Lakes region. This artifact is consistent with the issue documented by the National Weather Service, which states that the December 2020 HRRR upgrade inadvertently specified the elevation of all Great Lakes grid points at 0m (National Weather Service & NCEP Environmental Modeling Center, 2022).

To quantify the impact of this issue, we compare HRRR surface pressure against IFS HRES respective variable. Figure 6 shows yearly mean maps of the surface pressure bias, computed as the difference between HRRR and IFS HRES, over the Great Lakes region. The bias pattern is weak in 2020, becomes clearly pronounced in 2021 and remains visible in 2022. The spatial structure aligns closely with the Great Lakes boundaries, indicating that the anomaly is not a broad regional model bias but rather a localized artifact.

The temporal behavior is shown in Figure 7, where we plot the mean surface pressure bias averaged over the full HRRR domain. An abrupt increase occurs near the end of 2020, followed by a persistently elevated bias through 2021 and into 2022. After the correction was implemented in July 2022, the bias drops significantly. This discontinuity is consistent with the timing described in Service Change Notice and strongly suggests that the affected period should not be used for pressure-based model evaluation without additional filtering.

To verify that this behavior is driven by the Great Lakes artifact rather than by a domain-wide change in HRRR quality, we repeat the same analysis after excluding the Great Lakes region. The resulting time series and distribution are shown in Figure 8. Once the Great Lakes are removed, the sharp discontinuity largely disappears and the mean bias becomes much more temporally consistent. This provides additional evidence that the anomaly is spatially confined and originates from the documented elevation error.

Based on these diagnostics, we exclude surface pressure from all modeling and physical consistency checks in this work. The identified non-physical artifact in the December 2020 through July 2022 period introduces inconsistencies that would compromise both training and the interpretation of surface-pressure-based diagnostics.

G Embedding Correlation Analysis

G.1 Embeddings similarity

Each embedding is processed independently by Aurora's decoder, therefore, it must possess all required meteorological information within the given spatial patch and does not require data from other global regions. We anticipate high similarity among:

- Embeddings spatially proximate to each other.
- Embeddings corresponding to identical seasonal periods.
- Embeddings derived from analogous geographical regions (e.g., oceanic areas exhibit higher mutual similarity than flatlands).

To enable cross-regional comparison while accounting for temporal offsets from time zone variations, embeddings from identical timestamps cannot be compared directly with each other. For analysis, we first average every 28 consecutive embeddings (representing one week) before computing similarity metrics.

Distance Similarity Figure 9 shows that the correlations between global climate and the *CONUS* region occur at *CONUS* latitudes (-25 to -45 in the northern hemisphere and 25 to 45 in the southern hemisphere). The weakest correlations appear in polar regions and equatorial oceans. These results indicate that Aurora's backbone embeddings preserve the expected distance similarity characteristics.

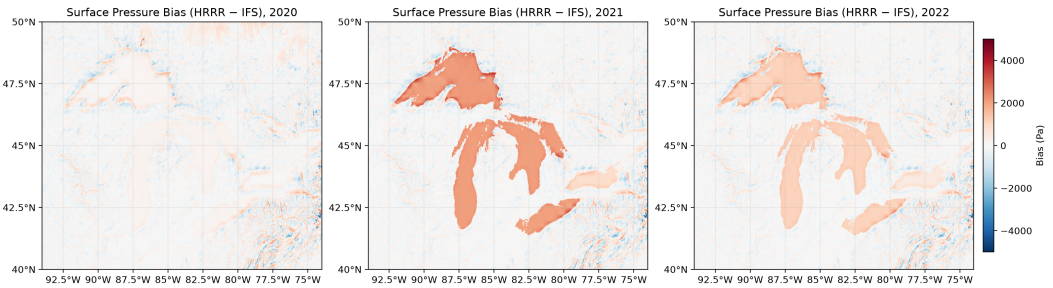


Figure 6: Yearly mean surface pressure bias over the Great Lakes region, computed as HRRR minus IFS HRES, for 2020–2022. The positive bias is strongest in 2021, consistent with the documented HRRR Great Lakes elevation issue.

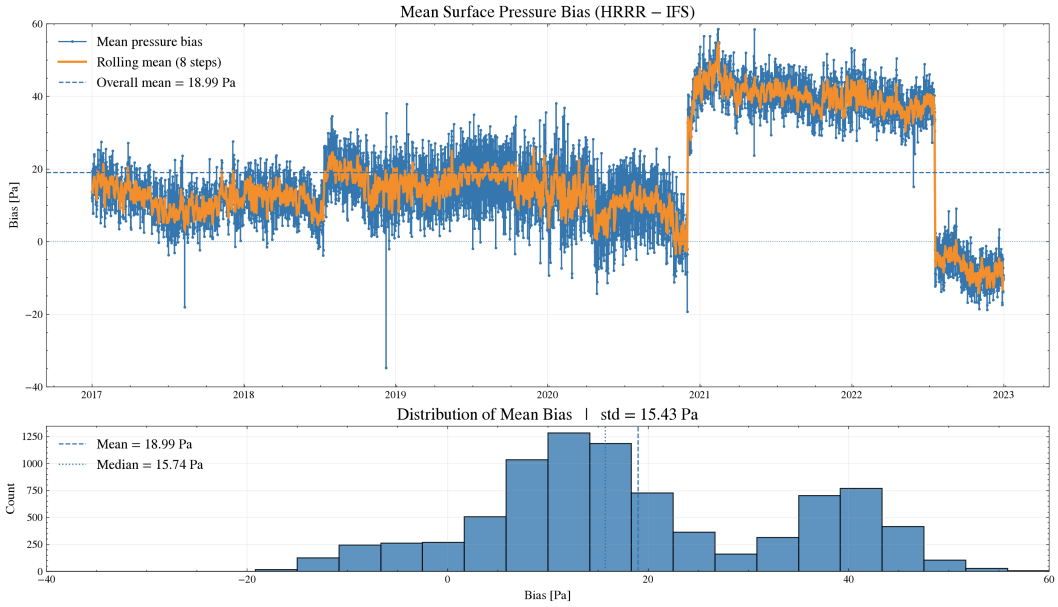


Figure 7: Full-domain time series and distribution of the spatially averaged HRRR —IFS HRES surface pressure bias. The elevated bias during 2021–2022 is consistent with the documented Great Lakes elevation issue.



Figure 8: Same as Figure 7, but with the Great Lakes region excluded from the spatial average. The sharp shift seen in the full-domain time series is strongly reduced, demonstrating that the anomaly is primarily confined to the Great Lakes and is not representative of the broader HRRR domain.

Temporal Similarity We evaluate temporal similarity of the embeddings, calculated from IFS HRES by Aurora's backbone, by first spatially averaging across the entire *CONUS* region and temporally averaging every 28 consecutive embeddings (corresponding to one week). Pairwise similarity is then computed between these weekly averaged embeddings over the 2016–2022 period. Figure 10 demonstrates distinct separation between summer and winter patterns, along with interannual seasonal consistency, for both surface-level and atmospheric variables.

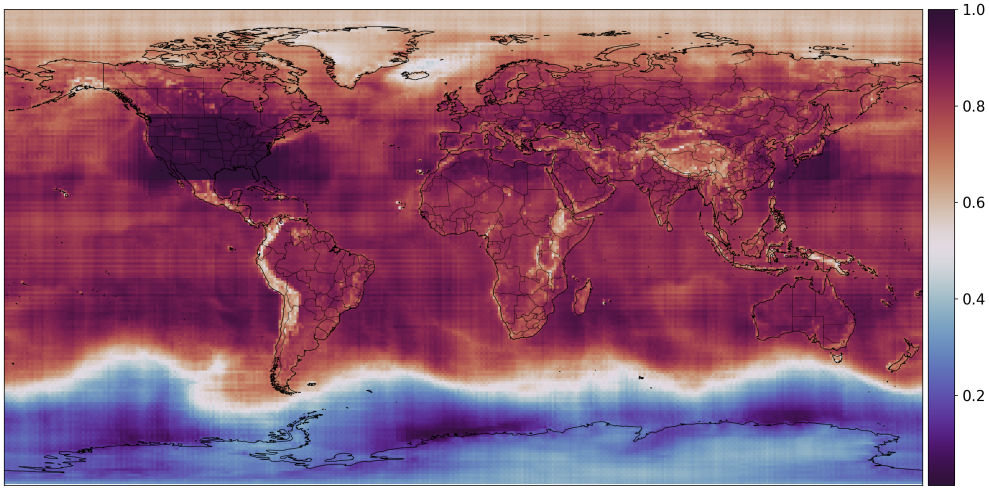
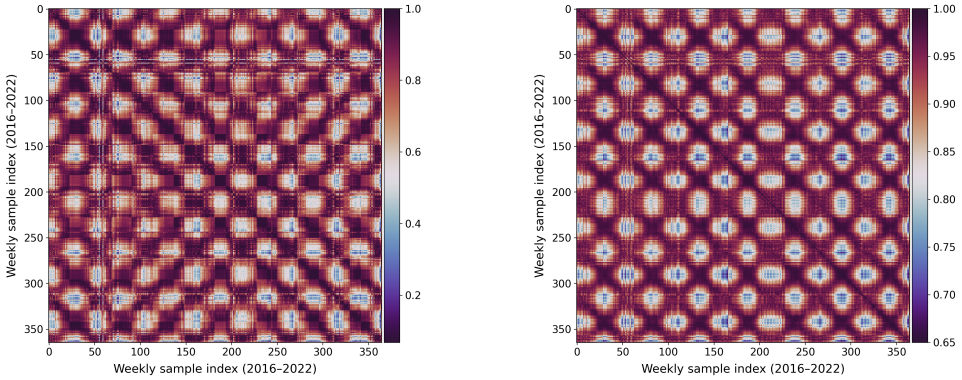


Figure 9: Embedding correlation plot. For each geographic location, we identify the embedding with the highest correlation within the CONUS region. This produces a visualization that maps, for every global position, the most similar embedding in the U.S., highlighting cross-regional embedding similarities.

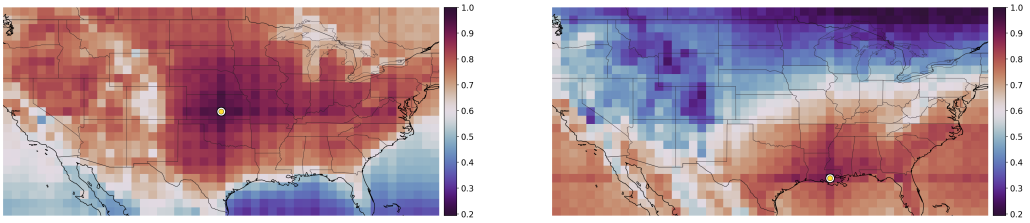


(a) Surface variables temporal correlation

(b) Atmospheric variables temporal correlation

Figure 10: Comparison of temporal correlation between surface and atmospheric channels of Aurora's embeddings, spatially averaged over the *CONUS* region.

Terrain-Dependent Spatial Similarity in the Latent Space As discussed above, embeddings are expected to exhibit similarity that depends not only on spatial proximity but also on terrain characteristics (e.g., mountains, oceans, flatlands). To investigate this hypothesis, we randomly select a single patch within the CONUS region and compare its cosine similarity to all other patches in the region, averaged over the whole week. This temporal aggregation mitigates short-term meteorological variability, the specific time period is otherwise not critical for the experiment. The resulting similarity map is shown in Figures 11a and 11b. As expected, the central patch exhibits the highest (self-)similarity. Moving away from this location, similarity generally decreases, however, the decay is markedly non-isotropic. This anisotropic pattern indicates that the relationship between spatial distance and embedding similarity is non-linear and strongly modulated by underlying terrain. In particular, sharp similarity gradients are observed along coastlines and over oceanic regions, where correlations attain their lowest values. These findings suggest that terrain topography plays a substantial role in shaping the structure of the latent space beyond mere geographic distance.



(a) Great Plains region

(b) Gulf Coast region

Figure 11: Pairwise cosine similarity between a reference patch embedding in each region and embeddings of all other patches.

H Physical Checks

To assess whether our model produces dynamically consistent atmospheric states, we evaluate several fundamental physical relationships across all pressure levels from 1000 to 200 hPa. These diagnostics do not directly measure forecast accuracy, but rather test whether the predicted fields obey basic atmospheric physics. We compare against both HRRR analysis as ground truth and WRF-ARW forecasts.

H.1 Geostrophic Balance

We decompose the wind field into geostrophic and ageostrophic components to check whether large-scale flow maintains approximate geostrophic balance. The geostrophic wind represents the theoretical balance between pressure gradient and Coriolis forces, while the ageostrophic part captures everything else – fronts, jets, convection, and other unbalanced motions. In practice, we find that the ageostrophic-to-geostrophic ratio stays small in mid-latitude synoptic systems but increases near strong gradients, which matches expected atmospheric behavior. This diagnostic reveals whether the model generates spurious unbalanced flow that would quickly distort under real dynamics.

Figure 12 shows the decomposition of wind fields into geostrophic and ageostrophic components across pressure levels. The ageostrophic-to-geostrophic ratio starts near 0.9 at surface levels and decreases to approximately 0.45 in the upper atmosphere, reflecting the expected transition from friction-dominated boundary layer flow to geostrophically balanced free-atmosphere dynamics. Our bicubic upsampling variant tracks the HRRR reference most closely, while the repeat (nearest-neighbor) variant systematically overestimates both components. This difference is consistent with the upsampling mechanism: repeat creates discontinuous patch boundaries that introduce spurious spatial gradients, which propagate into the geostrophic wind calculation through the pressure gradient term. The bicubic variant, by contrast, produces smoother fields that better preserve the gradient structure needed for balanced flow. All learned models show consistent underestimation across levels, suggesting a conservative bias rather than level-dependent failures.

H.2 Hydrostatic Balance

We verify hydrostatic balance using the hypsometric equation residual, which tests whether the geopotential thickness between pressure levels is consistent with the layer-mean virtual temperature. Unlike geostrophic balance, which is an approximation that holds only for large-scale flow, hydrostatic balance is satisfied to high accuracy in the real atmosphere except in intense convective updrafts. Learned models have no architectural guarantee of maintaining this relationship, and systematic violations would indicate unrealistic vertical accelerations that could destabilize longer forecasts.

Figure 13 shows the hydrostatic residual RMS across pressure levels from 1000 to 200 hPa. All model variants—including the repeat upsampling model that showed larger deviations in geostrophic diagnostics—track the ground truth profile closely. The residual is largest near the surface at approximately 120 m^2/s^2 at 1000 hPa, then decreases rapidly through the lower troposphere. Above 850 hPa the profile remains relatively flat around 20 m^2/s^2, with two modest peaks at 800 hPa (36 m^2/s^2) and 400 hPa (40 m^2/s^2). These peaks appear consistently across all models and the HRRR

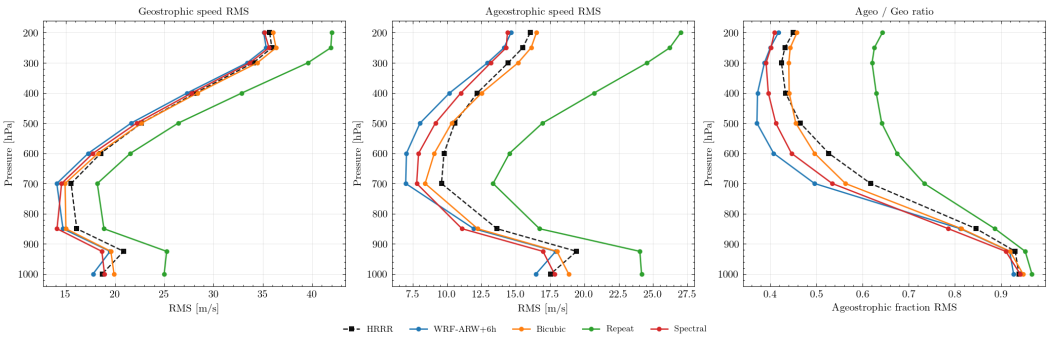


Figure 12: Root-mean-square (RMS) of geostrophic and ageostrophic wind components, along with their ratio, across pressure levels from 1000 hPa to 200 hPa. We evaluate HRRR ground truth, WRF-ARW +6h forecasts, and three variants of our best-performing model.

reference alike, suggesting they reflect genuine atmospheric structure rather than model artifacts. The surface maximum is expected given that boundary layer processes, sub-grid turbulence, and terrain effects introduce local departures from hydrostatic equilibrium that even the NWP solver must represent. Notably, the repeat variant does not diverge here as it did for geostrophic balance, because hydrostatic residuals depend on vertical column structure rather than horizontal gradients—the blocky upsampling artifacts corrupt spatial continuity but leave vertical relationships intact.

H.3 Divergence and Vorticity

We compute the kinematic ratio of horizontal divergence to relative vorticity to assess whether models capture the correct balance between rotational and divergent wind components. In mid-latitude synoptic systems, vorticity typically dominates (ratio below 1), while elevated ratios indicate regions of strong convergence, outflow, or ageostrophic motion. This diagnostic is particularly sensitive to wind field inconsistencies because both quantities are spatial derivatives that amplify small-scale errors.

Figure 13 shows the kinematic ratio across pressure levels from 1000 to 200 hPa. All learned models systematically underestimate the ratio by approximately 0.08–0.2 relative to HRRR ground truth, while WRF-ARW forecasts slightly overestimate it. The ground truth ratio decreases from 1.2 at 1000 hPa to a minimum near 0.5 at 300 hPa, then rises to 0.65 at upper levels, reflecting the expected transition from boundary-layer divergent motions to upper-tropospheric rotational flow. Among the learned variants, the bicubic upsampling model tracks ground truth most closely, whereas the spectral loss model shows the largest underestimation. This degradation likely stems from the spectral objective prioritizing global energy distribution over local gradient fidelity: divergence and vorticity are computed from spatial derivatives of the wind field, and spectral regularization can smooth precisely the fine-scale structures that contribute to divergent flow. The consistent underestimation across all learned models suggests they produce wind fields that are slightly too rotational relative to the true balance, possibly because the MAE loss penalizes extreme values in the derived quantities. Nevertheless, all variants reproduce the correct vertical profile shape.

H.4 Energy Spectra

To assess whether our model preserves multi-scale energy cascades and mitigates the spectral attenuation typical of models trained with mean absolute error (MAE), we compute horizontal isotropic power spectra $E(k)$ across synoptic ($k \sim 10^0$), mesoscale ($k \sim 10^1$), and convective scales ($k \sim 10^2$–10^3) down to our 2.7 km grid resolution. The energy spectrum quantifies how variance is distributed across spatial scales, with the slope of $E(k)$ revealing whether the model captures realistic multi-scale structure. This diagnostic is particularly sensitive to model formulation because numerical diffusion, inadequate resolution, or smoothing losses can artificially damp high-wavenumber energy, leading to overly smooth forecasts that fail to capture fine-scale structure. As shown in Figure 14, geopotential (z), temperature (t), and wind fields maintain robust, physically consistent power-law slopes across vertical levels (1000, 600 and 200 hPa). Crucially, while standard MAE training can

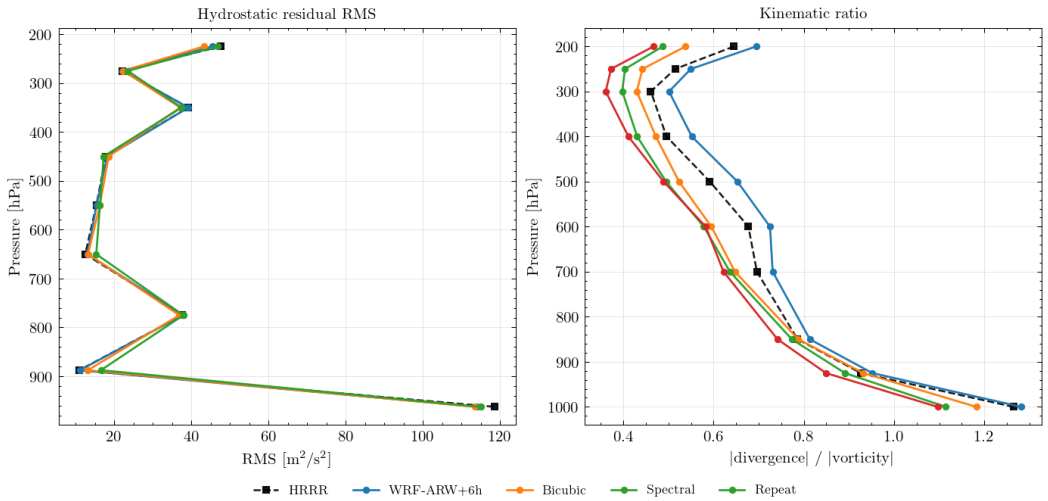


Figure 13: Root-mean-square (RMS) of hydrostatic residual, along with kinematic ratio, across pressure levels from 1000 hPa to 200 hPa. We evaluate HRRR ground truth, WRF-ARW +6h forecasts, and three variants of our best-performing model.

lead to high-frequency energy degradation or grid-scale noise artifacts caused by patch-decoding, integrating dedicated spectral loss (red lines) successfully regularizes the high-frequency tail, aligning the spectra closely with reference numerical simulations (WRF-ARW). For specific humidity (q), however, the spectral loss provides limited regularization, with high-frequency noise persisting slightly across convective scale. This reflects fundamental differences in how moisture behaves compared to other atmospheric variables: humidity exhibits sharper spatial gradients, stricter physical constraints (such as non-negativity), and conservation properties that operate globally rather than locally in frequency space. Recent work shows that these moisture-specific challenges require loss functions that explicitly enforce physical conservation laws and bounds, rather than spectral regularization alone (Sha et al., 2025). Incorporating such physics-aware constraints into the training objective represents a promising direction for future work.

H.5 Wind Vector Diagnostics

Figure 15 evaluates wind vector accuracy using four diagnostics computed from paired u and v wind components at each pressure level. For each u/v wind pair, we identify matching channels (e.g., u_{500}/v_{500} for 500 hPa level) and compute metrics over all valid spatial pixels.

Wind-pair vector RMSE measures the root-mean-square of the vector error magnitude:

$$\text{Wind-pair vector RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((u_{\text{pred},i} - u_{\text{tgt},i})^2 + (v_{\text{pred},i} - v_{\text{tgt},i})^2)},$$

where N is the number of valid pixels pooled across all samples and spatial locations. This metric captures the overall vector error combining both zonal and meridional components.

Mean vector error magnitude computes the mean of the vector error magnitudes:

$$\text{Mean vector error magnitude} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(u_{\text{pred},i} - u_{\text{tgt},i})^2 + (v_{\text{pred},i} - v_{\text{tgt},i})^2}.$$

Unlike wind-pair vector RMSE, this metric treats all errors linearly rather than emphasizing larger errors through squaring.

Wind-speed RMSE evaluates errors in wind speed (scalar magnitude) rather than vector components:

$$\text{Wind-speed RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{u_{\text{pred},i}^2 + v_{\text{pred},i}^2} - \sqrt{u_{\text{tgt},i}^2 + v_{\text{tgt},i}^2}\right)^2}.$$

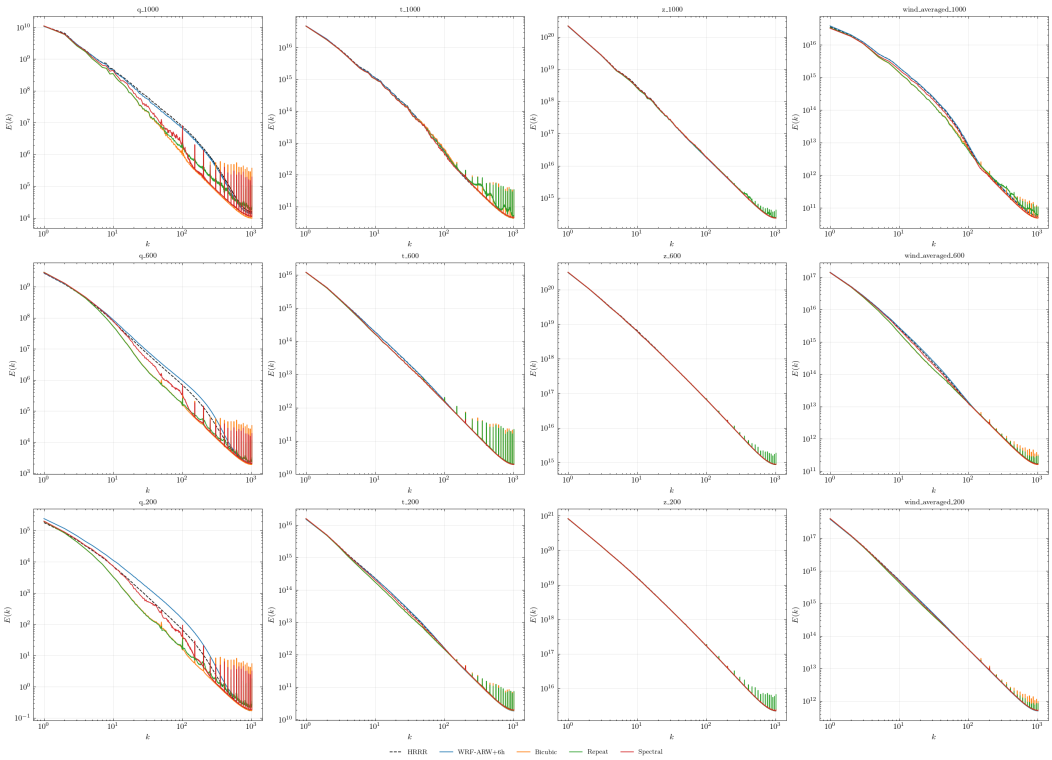


Figure 14: Spectral analysis across pressure levels (1000, 600, and 200 hPa) for specific humidity (q), temperature (t), geopotential (z) and wind speed (w, combined u and v components). We evaluate HRRR ground truth, WRF-ARW +6h forecasts, and three variants of our best-performing model.

This diagnostic is sensitive to speed biases even when wind direction is captured correctly.

Vector correlation measures the dot-product correlation between predicted and target wind vectors:

$$\text{Vector correlation} = \frac{\frac{1}{N}\sum_i (u_{\text{pred},i}u_{\text{tgt},i} + v_{\text{pred},i}v_{\text{tgt},i})}{\sqrt{\left(\frac{1}{N}\sum_i (u^2_{\text{pred},i} + v^2_{\text{pred},i})\right)\left(\frac{1}{N}\sum_i (u^2_{\text{tgt},i} + v^2_{\text{tgt},i})\right)}}.$$

This metric equals 1.0 when wind vectors are perfectly aligned in direction and magnitude, and decreases with both directional and speed errors.

In Figure 15, the WRF-ARW +6h forecast (blue) generally exhibits the largest errors across all metrics. Bicubic (orange) and Repeat (green) variants show lower errors and better correlation for almost all metrics and their results almost do not differ between each other. Spectral (red) has visibly larger error for lower levels, but still outperforms WRF-ARW on above 500hPa. Vector correlation remains above 0.95 for all learned models through 500 hPa, increasing to 0.99-1.00 at upper levels.

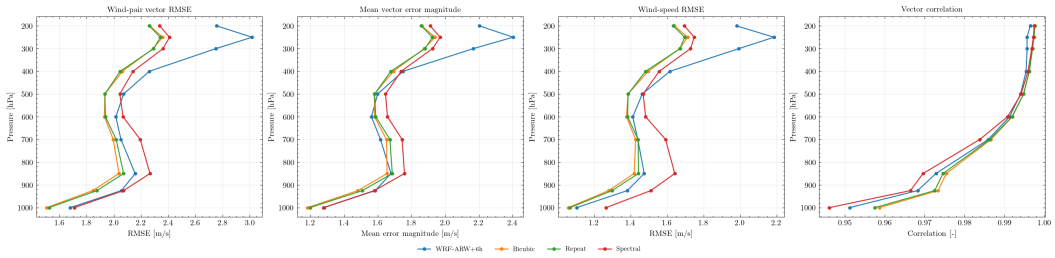


Figure 15: Wind-pair vector diagnostics across pressure levels from 1000 to 200 hPa. From left to right: vector RMSE, mean vector error magnitude, wind-speed RMSE, and vector correlation. Metrics are computed from paired u/v wind components at each level. We compare, WRF-ARW +6h forecasts, and Repeat, Bicubic and Spectral variants of our best-performing model.

H.6 Bias Profiles

Figure 16 shows vertical bias profiles for five atmospheric variables: specific humidity (q), temperature (t), u and v wind components, and geopotential (z). For each output channel, bias is computed as the mean error over all valid pixels:

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N}(y_{\text{pred},i} - y_{\text{tgt},i}).$$

Positive bias indicates systematic overestimation relative to HRRR ground truth, while negative underestimation. Bias is computed independently for each pressure level, revealing systematic errors across the vertical structure.

In general, biases are low across all variables and models. Our learned models exhibit lower absolute bias than WRF-ARW +6h forecasts, but with systematically different characteristics. While WRF-ARW generally underestimates (negative bias) across variables, our models tend to slightly overestimate (positive bias). For temperature, our variants stay within ± 0.05 K versus WRF-ARW's cold bias up to -0.15 K. Wind component biases show a similar pattern, with our models within ± 0.15 m/s compared to WRF-ARW's negative biases reaching -0.25 m/s for u and -0.15 m/s for v. The u component is notably different at middle levels (700–400 hPa), where WRF-ARW exhibits almost no bias while our models show small positive bias. For geopotential all models show bias constantly increasing/decreasing with height where learned models display positive bias up to approximately 30 m at 200 hPa, consistent across all decoder variants and -30 m for WRF-ARW.

I Distribution of error on in-situ observations

Evaluating against in situ observations from *HadISD*, we find that our model consistently outperforms *Aurora* across the *CONUS* region. This improvement is particularly pronounced in areas that are

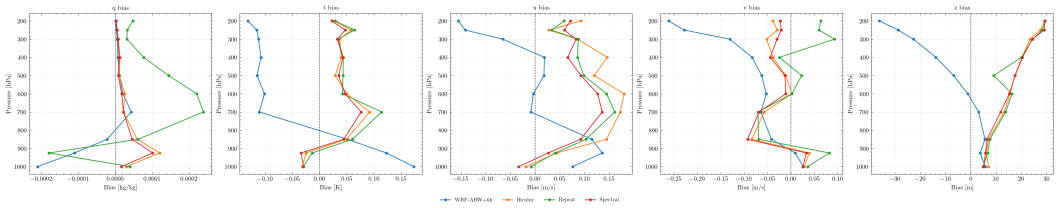


Figure 16: Bias profiles across pressure levels from 1000 to 200 hPa for specific humidity (q), temperature (t), zonal wind (u), meridional wind (v), and geopotential (z). We evaluate WRF-ARW +6h forecasts and three variants of our best-performing model.

difficult to model, such as mountainous terrain with sharply varying elevations and regions influenced by strong land–water contrasts (e.g., along the California coast and around the Great Lakes), where associated mesoscale processes render weather forecasting more challenging. In these regions, atmospheric conditions exhibit strong spatial variability; therefore, fine-resolution modelling is crucial to adequately capture physical processes that are not resolved at the coarser $0.25°$ scale. This observation is consistent for both temperature and wind.

Figures 17 and 18, which illustrate these results, use a shared colour scale for each variable to facilitate comparison between models. For each variable, the colour scale spans from the minimum *MAE* observed across both models to the larger of their 95th-percentile *MAE* values. Values above the 95th percentile are capped for visualization purposes. A background elevation map is overlaid on the visualizations to highlight error patterns associated with mountainous terrain.

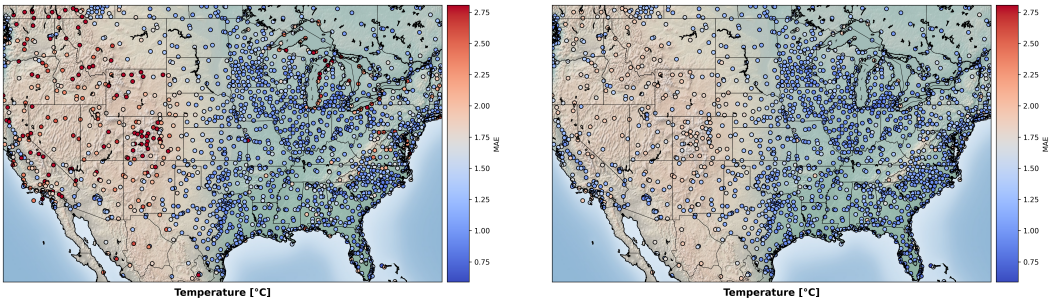


Figure 17: Distribution of temperature *MAE* between *HadISD* weather station observations and model predictions for the 2022 evaluation year: Aurora predictions (left) and our best model (right).

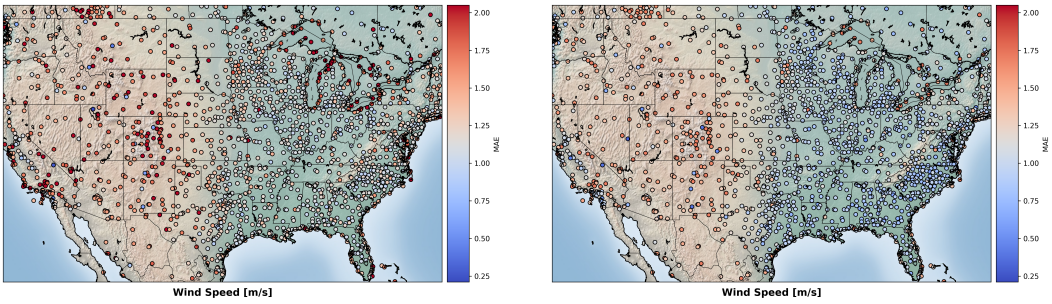


Figure 18: Distribution of wind *MAE* between *HadISD* weather station observations and model predictions for the 2022 evaluation year: Aurora predictions (left) and our best model (right).

J ADDITIONAL RESULTS

Original decoder impact To check whether the use of original pretrained Aurora's linear decoder helps, Table 5 compares three multi-scale architectures with and without this pretrained decoder head (AUR-MLP-4-0). For the linear configuration (MLP-40-0 | MLP-8-0), the pretrained decoder provides a strong warm start, reducing RMSE for all variables. For deeper configurations, however, the benefit is not obvious. For the single and two hidden layer models the differences with and without the pretrained decoder remain below 0.3%. This indicates that once a decoder is deep enough to learn the latent-to-output mapping on its own, the pretrained read-out becomes largely redundant. However, we still decided to keep the original model for our best models as it provides a better initialization and allows for faster convergence.

Table 5: Performance comparison between models with and without the pretrained Aurora linear decoder head (AUR-MLP-4-0$\uparrow_{40}$). Evaluation performed on the HRRR grid using RMSE at +6h lead time. T_2 denotes 2-m air temperature (°C), u_{10} and v_{10} wind component speeds (m/s), and $MSLP$ mean sea level pressure (hPa).

Model	Pretrained decoder	HRRR			
		T_2	u_{10}	v_{10}	$MSLP$
MLP-40-0 | MLP-8-0$\uparrow_{40}^{\text{NN}}$	×	1.8285	1.5163	1.5401	1.1433
MLP-40-0 | MLP-8-0$\uparrow_{40}^{\text{NN}}$	✓	1.4948	1.4532	1.4682	0.7579
MLP-40-1 | MLP-8-1$\uparrow_{40}^{\text{NN}}$	×	1.2029	1.2950	1.3273	0.5710
MLP-40-1 | MLP-8-1$\uparrow_{40}^{\text{NN}}$	✓	1.2055	1.2954	1.3297	0.5694
MLP-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$	×	1.1795	1.2681	1.3045	0.5700
MLP-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$	✓	1.1832	1.2731	1.3024	0.5653

Multi-region training To measure our model's ability to adapt to multiple domains via cross-regional training we train a AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$| MLP-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$ model jointly on two geographically separate datasets – DANRA and HRRR. See Appendix B.2 for a more detailed description. As a benchmark of how good prediction in the Baltic region can be, we train three types of models:

- Trained solely on DANRA dataset.
- Trained solely on HRRR dataset.
- Trained on both DANRA and HRRR datasets.

The first and second experiments are meant to provide the best possible result that can be achieved by our models each region separately. The third one shows the potential of training a model on two different regions of the world simultaneously. HRRR and DANRA are quite representative examples of regions that models might be trained on, as they are far away from each other. We restrict training time domain to 2017-2019 on both datasets to not leak any year-related meteorological information from HRRR's 2020 year to DANRA evaluation.

Multi-region training results are summarized in Table 6, which shows the performance when training exclusively on one region versus training on the combined dataset. We observe that a unified model trained on both HRRR and Danish regions successfully adapts to both domains. Performance remains comparable with only marginal increases in error. Notably, for the DANRA region, the combined training even yields small improvement for temperature at 2 meters.

This demonstrates that our lightweight decoder heads are capable of generating accurate weather predictions for two different regions located over 7000 kilometres away from each other, thereby making the achieved performance particularly meaningful.

Seeding To quantify the sensitivity of our reported results to random initialization, we trained the best-performing configuration (AUR-MLP-4-0$\uparrow_{40}^{\text{NN}}$| MLP-40-2 | MLP-8-2$\uparrow_{40}^{\text{NN}}$) ten times with different random seeds and report the per-variable RMSE on the HRRR grid at +6h lead time in

Table 6: Multi-domain joint training across two geographically distinct areas. We evaluate the effect of training on both CONUS (HRRR) and the Baltic region (DANRA). The models were trained on data from 2017 to 2019 and evaluated in 2020. We report RMSE for T_2, u_{10}, and v_{10} wind components (m/s), as well as mean sea level pressure $MSLP$ (hPa). Additionally, we report the difference in RMSE relative to single-domain training.

	HRRR				DANRA			
	T_2	u_{10}	v_{10}	$MSLP$	T_2	u_{10}	v_{10}	$MSLP$
HRRR	1.1042	1.1638	1.1661	0.4887	–	–	–	–
DANRA	–	–	–	–	0.6831	0.9028	0.8789	0.2700
HRRR + DANRA	1.1128	1.1754	1.1767	0.4925	0.6640	0.9049	0.8808	0.2915
Difference	+0.0086	+0.0116	+0.0106	+0.0038	-0.0191	+0.0020	+0.0019	+0.0216

Table 7: RMSE of the best configuration trained with 10 random seeds on the HRRR grid (+6h lead time). T_2 denotes 2-m air temperature (°C), u_{10} and v_{10} wind component speeds (m/s), and $MSLP$ mean sea level pressure (hPa).

Seed	T_2	u_{10}	v_{10}	$MSLP$
1	1.1827	1.2725	1.3019	0.5652
2	1.1840	1.2730	1.3027	0.5644
3	1.1827	1.2726	1.3038	0.5664
4	1.1834	1.2730	1.3032	0.5653
5	1.1854	1.2720	1.3036	0.5652
6	1.1827	1.2719	1.3040	0.5671
7	1.1831	1.2725	1.3030	0.5655
8	1.1834	1.2719	1.3039	0.5665
9	1.1805	1.2726	1.3013	0.5675
10	1.1839	1.2715	1.3033	0.5665
Mean ± std	1.1832 ± 0.0012	1.2724 ± 0.0005	1.3031 ± 0.0009	0.5660 ± 0.0010

Table 7. Across the ten runs results are highly consistent: the standard deviation of the RMSE is over two orders of magnitude smaller than the mean error for every variable, corresponding to a coefficient of variation below 0.2% (T_2: 1.1832 ± 0.0012 K, u_{10}: 1.2724 ± 0.0005 m/s, v_{10}: 1.3031 ± 0.0009 m/s, MSLP: 0.5660 ± 0.0010 hPa). These results confirm that our training pipeline is robust to the choice of seed, and that the reported gains over baselines, across decoder variants and other methods reflect genuine architectural improvements rather than a single favorable initialization.

Evaluation on StormCast grid To enable direct comparison with StormCast, which operates on a Lambert Conformal Conic (LCC) projection at kilometer-scale resolution, we regrid our predictions onto StormCast's native grid. Although roundtrip regridding error between LCC and latitude–longitude coordinates is small (Appendix D), this experimental setup provides the fairest possible comparison by eliminating interpolation artifacts that could differentially affect evaluation metrics.

Our best model, evaluated on the LCC grid, outperforms the StormCast 5-member ensemble across all reported variables on both HadISD in-situ observations and the HRRR grid (Table 8). On the HRRR grid, our approach reduces RMSE by 44.5% for temperature (1.247 vs. 2.236), 18.3% for u_{10} (1.275 vs. 1.562), and 18.6% for v_{10} (1.317 vs. 1.619) compared to StormCast. Similar improvements are observed on HadISD stations, where our model achieves 1.539 vs. 2.446 for temperature and comparable gains for both wind components. The consistent advantage across both evaluation grids indicates that improvements are not artifacts of coordinate system choice but reflect genuine gains in capturing regional atmospheric structure.

To enable comprehensive comparison between our approach and StormCast, we evaluate both models across atmospheric pressure levels (Figure **??**). StormCast operates on hybrid sigma-pressure

Table 8: Performance evaluation of our approach against WRF-ARW, StormCast and Aurora's $0.25°$ base model. Results are presented for HadISD in-situ observations and the HRRR grid using RMSE at +6h lead time. T_2 denotes 2-m air temperature (°C), u_{10} and v_{10} wind component speeds (m/s), and $MSLP$ mean sea level pressure (hPa).

Grid	Model	HadISD (in-situ)			HRRR Grid			
		T_2	u_{10}	v_{10}	T_2	u_{10}	v_{10}	$MSLP$
	AUR-MLP-4-0$\uparrow_{40}^{NN}$							
LCC (km)	| MLP-40-2 | MLP-8-2$\uparrow_{40}^{NN}$	1.539	1.493	1.570	1.247	1.275	1.317	0.559
Lat-lon (°)	| MLP-40-2 | MLP-8-2$\uparrow_{40}^{NN}$	**1.524**	**1.485**	**1.565**	**1.204**	**1.215**	**1.264**	**0.550**
LCC (km)	StormCast	2.446	1.828	1.882	2.236	1.562	1.619	1.377
Lat-lon (°)	StormCast	2.428	1.835	1.894	2.236	1.572	1.631	1.379

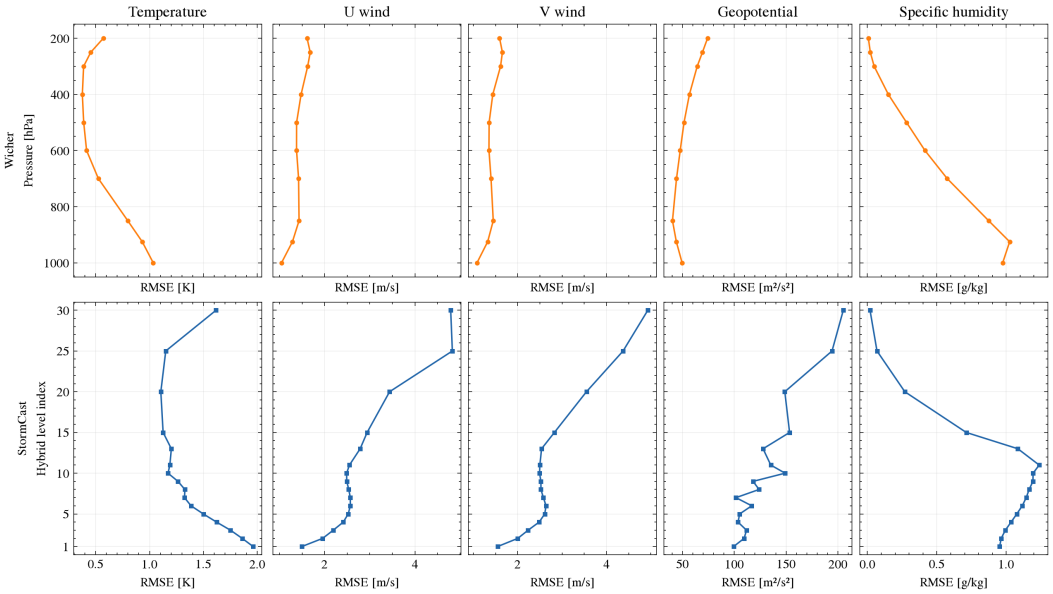


Figure 19: Vertical RMSE profiles relative to HRRR ground truth. The top row shows the bicubic version of our best model (Wicher) on isobaric pressure levels; the bottom row shows the StormCast 5-member ensemble on hybrid sigma-pressure levels, with level 1 nearest the surface. Columns show temperature (T), zonal and meridional wind (u, v), geopotential (Φ), and specific humidity (q). Both rows share an RMSE scale within each column.

coordinates, which combine terrain-following layers near the surface with constant pressure surfaces at altitude, whereas our model predicts directly on uniform isobaric levels. Since HRRR provides data in both coordinate systems, we compare each model against HRRR ground truth in its native vertical coordinate, ensuring fair evaluation without cross-coordinate interpolation artifacts. All plots show RMSE computed with respect to HRRR ground truth.

Comparison to ML-LAM(DANRA) Figure 20 shows relative difference (the lower the better) between RMSE values of our model trained on DANRA region and ML-LAM(DANRA).

K Data Scale

Data volume and preprocessing. The complete training corpus comprises three raw meteorological sources (IFS HRES, HRRR, DANRA) totaling approximately 15 TB of decompressed fields. Directly operating on this data during training would incur substantial I/O and memory overhead. To address this, we precompute and cache intermediate embeddings produced by Aurora's frozen backbone from IFS inputs. The full-resolution global embeddings occupy approximately 10TB; spatially cropping

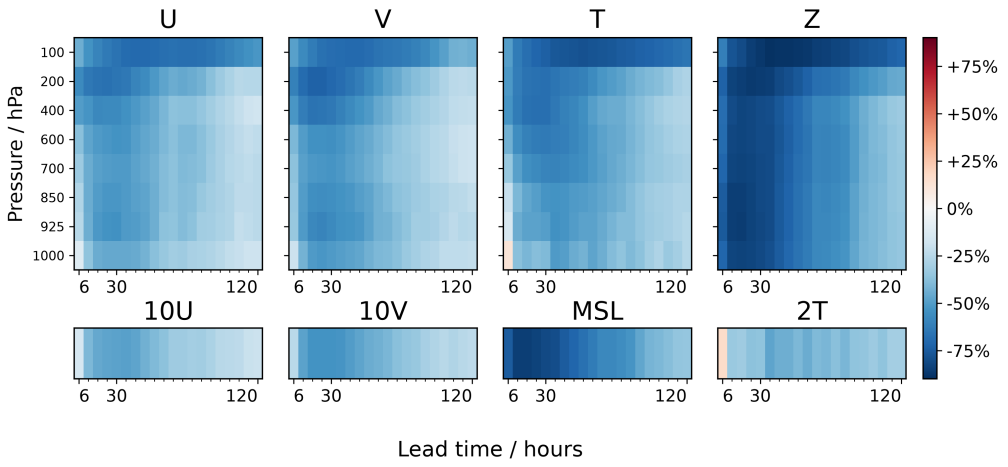


Figure 20: Performance comparison over a 120-hour horizon for our model relative to the ML-LAM(DANRA).

to the HRRR domain reduces this to approximately 200GB. These cached embeddings are reused across all experiments, enabling efficient training of multiple decoder configurations without repeated backbone inference. All datasets are loaded using memmaps, a memory-mapped data format that enables efficient random access to large spatiotemporal fields without loading them fully into memory.

L Sample Predictions

This appendix presents visual examples of model predictions for key surface variables at representative timestamps from the 2022 evaluation period. Each figure displays three panels: the model prediction (left), the HRRR target ground truth (center), and the prediction residual (right). All predictions are shown for the CONUS domain using the bicubic upsampling variant of our best-performing model at +6h lead time.

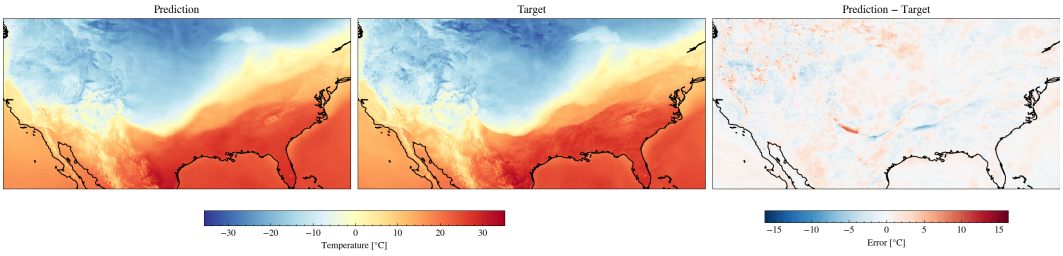


Figure 21: 2 m air temperature [°C]. Timestamp: January 1, 2022, 18:00 UTC.

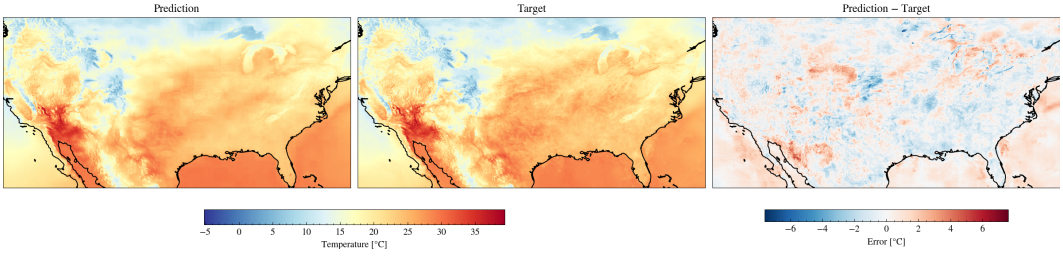


Figure 22: 2 m air temperature [°C]. Timestamp: July 1, 2022, 06:00 UTC.

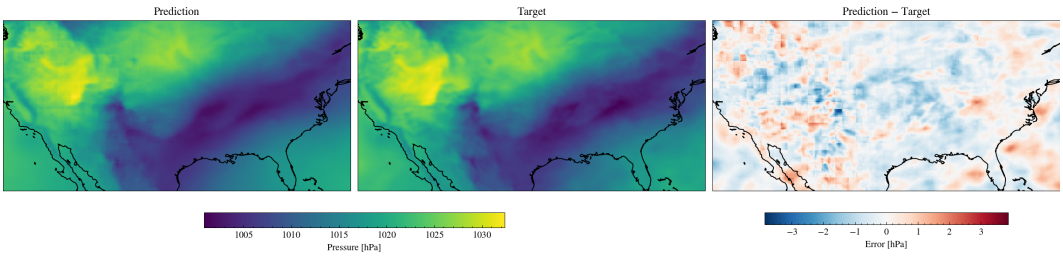


Figure 23: Mean sea level pressure [hPa]. Timestamp: January 1, 2022, 18:00 UTC.

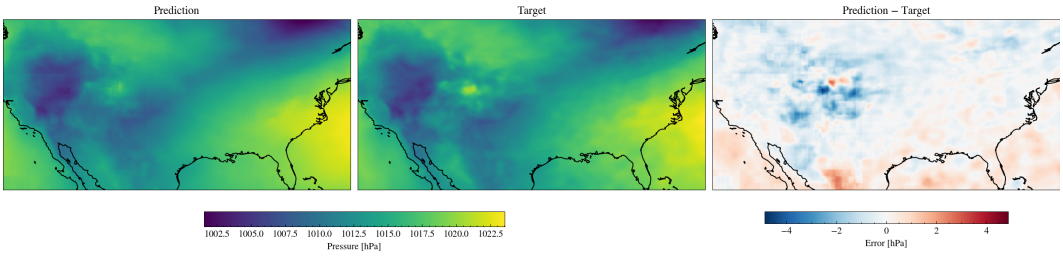


Figure 24: Mean sea level pressure [hPa]. Timestamp: July 1, 2022, 06:00 UTC.

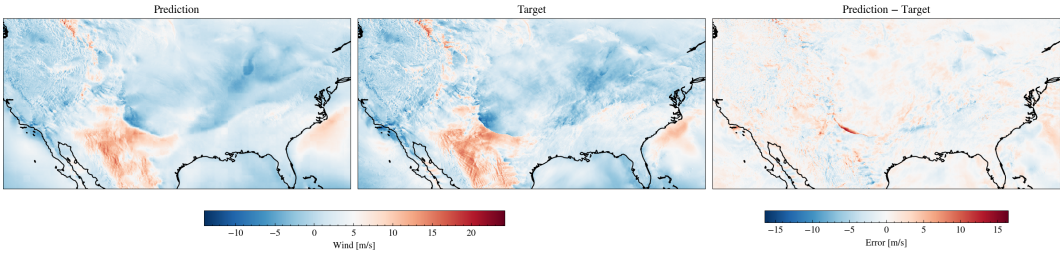


Figure 25: 10 m u-component of wind [m/s]. Timestamp: January 1, 2022, 18:00 UTC.

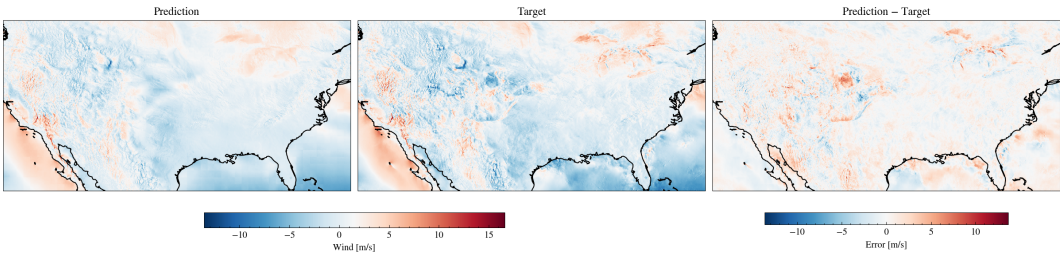


Figure 26: 10 m u-component of wind [m/s]. Timestamp: July 1, 2022, 06:00 UTC.

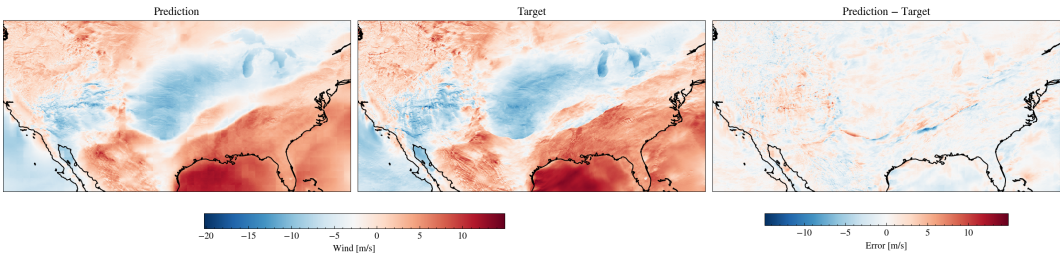


Figure 27: 10 m v-component of wind [m/s]. Timestamp: January 1, 2022, 18:00 UTC.

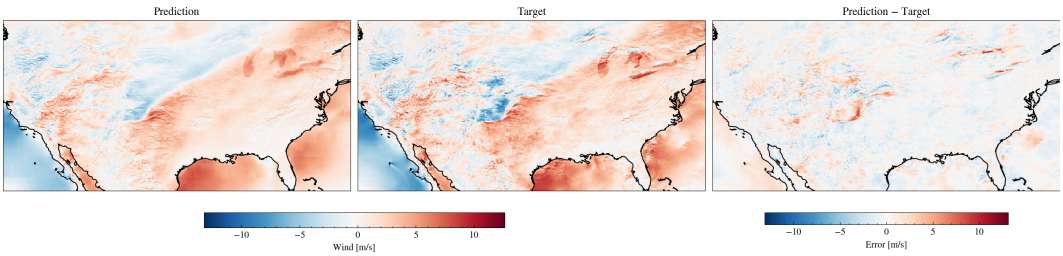


Figure 28: 10 m v-component of wind [m/s]. Timestamp: July 1, 2022, 06:00 UTC.