https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

FiLMeR (v1.0): a FiLM-conditioned machine-learning emulator for multi-domain regional WRF-scale surface weather fields

Balbir Prasad[1,★], Deena Lad[1,★], Mohammad Rafiuddin[2], Udit Bhatia[1], and Nipun Batra[1]

[1]Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, India
[2]Council on Energy, Environment and Water (CEEW), Delhi, India
★These authors contributed equally to this work.

Correspondence: Balbir Prasad (balbir.prasad@iitgn.ac.in)

Abstract. High-resolution regional weather forecasting supports applications such as early warning, hydrological planning, and regional risk assessment, but traditional numerical models such as the Weather Research and Forecasting (WRF) system remain computationally demanding for rapid ensemble or sensitivity experiments. Machine-learning emulators can substantially lower this cost, but many are tied to a single grid or depend on future boundary fields that are not always available. We present **FiLMeR (v1.0)**, a context-conditioned machine-learning emulator for WRF-scale near-surface regional weather fields. FiLMeR encodes dynamic atmospheric forcing and static physiographic information in two separate encoders and conditions the decoder with Feature-wise Linear Modulation (FiLM) using grid spacing and cyclic time-of-day/day-of-year metadata. The model predicts six near-surface variables (temperature, humidity, surface pressure, wind components, and precipitation) from historical Global Forecasting System (GFS) states, using a hurdle-style precipitation head. A single shared-weight model is trained and evaluated across four nested WRF domains over the Indian subcontinent, at two grid spacings (27 km and 9 km).

We test whether FiLMeR's two central design choices, separating inputs into two encoders and conditioning with FiLM, are each responsible for its accuracy. This is done using a controlled comparison across seven combinations of these two choices, together with a diagnostic test on the trained model. Any form of conditioning substantially improves on an unconditioned model, but the dual-encoder design on its own performs worse than a simpler single encoder, and needs FiLM to become competitive again. Once FiLM is added, this configuration ties with a simpler single-encoder design that uses a more basic conditioning method, under a comparison rule fixed before the results were seen; neither is clearly better.

The diagnostic test further shows that FiLM's measured contribution is mainly temporal (time-of-day and day-of-year), not domain-adaptive, even though the model is trained across four geographically distinct domains. On the independent 2025 test period, FiLMeR achieves a 2 m temperature RMSE of 1.50 K and a surface-pressure RMSE of 1.37 hPa, and produces the corresponding 80-step, 3-hourly output sequence (equivalent to a 10-day WRF forecast) in a fraction of a second on a single GPU, about 38,000 times faster than the WRF dynamical integration it emulates. FiLMeR is therefore a computationally efficient emulator for these WRF-scale surface fields: conditioning is necessary for its accuracy, and FiLM's benefit is primarily a temporal-conditioning mechanism.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

1 Introduction

High-resolution regional weather forecasts are central to early warning, infrastructure planning, agriculture, water management, energy-system operation, and disaster-risk reduction (Bauer et al., 2015). Accurate weather forecasts are particularly vital in topographically complex regions, such as the Indian subcontinent, to mitigate the impacts of severe monsoonal systems and localized convective storms (Hassan et al., 2026; Navale and Singh, 2020). Numerical Weather Prediction (NWP) remains the standard approach for producing such forecasts. The Weather Research and Forecasting (WRF) model is one such widely used NWP system for regional atmospheric simulations and forecasts, providing flexible nesting, several physics parameterizations, and terrain-aware limited-area modeling (Skamarock et al., 2008). In practice, however, each high-resolution WRF forecast requires preprocessing of global analyses or forecasts, static geophysical data preparation, lateral-boundary handling, and time-stepping of a coupled dynamical system. This computational burden limits rapid sensitivity experiments, large ensemble generation, repeated deployment over many regional domains, and fast iteration during extreme-event response.

Data-driven atmospheric prediction has recently offered a complementary route to faster forecasting. Early machine-learning efforts in meteorology often focused on post-processing NWP output, correcting systematic biases, or emulating individual physical parameterizations (Rasp and Lerch, 2018; Tian et al., 2024). More recent global models have moved closer to direct weather prediction by learning the evolution of gridded atmospheric states from large historical archives. FourCastNet, GraphCast, Pangu-Weather, GenCast, and FengWu show that neural networks can represent large-scale atmospheric evolution while reducing inference cost by orders of magnitude (Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023; Price et al., 2025; Chen et al., 2025). These advances have changed expectations for operational weather prediction and for model components that can be evaluated repeatedly at low cost.

At the same time, most global learned models operate on fixed coarse grids and are not designed to resolve the fine-scale terrain effects, land-surface heterogeneity, and boundary-layer processes that govern near-surface weather at regional scales (Xu et al., 2025; Lam et al., 2023). Limited-area models such as WRF remain necessary for applications where local physiography, mountain ranges, coastlines, and land-use patterns strongly modulate surface temperature, wind, and precipitation (Skamarock et al., 2008; Navale and Singh, 2020).

To address the spatial limitations of global models, recent research has considered learning-based limited-area models (LAMs) and statistical downscaling operators. Frameworks such as YingLong (Xu et al., 2025) and STCast (Chen et al., 2026) follow the limited-area modelling idea by training neural networks on regional grids, while generative approaches such as CorrDiff (Mardani et al., 2023) and StormCast (Pathak et al., 2026) use diffusion-based residual correction to represent localized variability conditioned on coarser input fields. These systems indicate the potential of learned regional modelling, but several practical constraints remain for reusable emulator components. Many architectures are designed for a fixed grid or resolution, so transfer to a different nested region may require retraining or architectural changes; for example, YingLong (Xu et al., 2025) is trained and evaluated on a fixed East Asia domain, and STCast (Chen et al., 2026) operates on a single predefined regional grid without explicit mechanisms for adapting to different domain extents or grid spacings. FiLMeR itself still emits predictions on a common resampled output grid for every domain (Section 2.1); the distinction drawn here is that its

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

weights are shared and conditioned across different native grid spacings and geographic extents. Several regional systems also retain a dependence on future lateral-boundary information from upstream global forecasts. In addition, static physiographic fields, such as terrain elevation and land-use classification, are often concatenated directly with dynamic atmospheric states (Salman et al., 2024; Lee et al., 2024). This early-fusion approach is simple, but it may not provide a distinct pathway for learning the influence of lower-boundary information on near-surface fields.

A related modelling question is how to combine heterogeneous atmospheric, physiographic, and metadata inputs. Traditional numerical systems include surface and land-atmosphere parameterizations that link lower-boundary conditions to near-surface weather (Skamarock et al., 2008). In data-driven models, this link must be learned from data. One approach is to concatenate physiographic fields directly with atmospheric inputs and let the network learn their joint influence, but this treats all input modalities equally and provides no explicit mechanism for the model to adapt its behaviour to different domain geometries or grid spacings. FiLM (Perez et al., 2018) offers an alternative: rather than merging all inputs at the start, a small conditioning network reads domain and time metadata and adjusts the internal responses of the main network layer by layer. A related idea, spatially-adaptive normalization (Park et al., 2019), conditions network activations directly on a spatially-varying map rather than a compact vector; it is architecturally the closest prior approach to how FiLMeR's static physiographic fields could in principle be used to condition the network, though FiLMeR instead conditions on a compact spatiotemporal vector and treats the physiographic fields as a separate encoder input. In a regional emulator, this allows a single shared model to produce domain-aware surface predictions without requiring separate weights for each target region.

We address these issues with **FiLMeR (v1.0)**, a FiLM-based regional emulator for multi-resolution surface weather prediction. FiLMeR predicts the state of weather over a region at time $t + \Delta t$ from only the current and previous GFS states at t and $t - \Delta t$, together with static WRF geogrid fields and a compact spatiotemporal context vector. The model separates dynamic atmospheric and static physiographic inputs into two independent encoders and conditions the decoder with FiLM (Perez et al., 2018) on grid spacing and cyclic time-of-day/day-of-year metadata, while absolute domain position is carried separately by coordinate channels appended to the static input. A controlled comparison against single-encoder and unconditioned alternatives (Section 4) shows that conditioning, rather than the dual-encoder topology on its own, accounts for most of the resulting accuracy, and a context counterfactual further shows that FiLM's measured contribution is mainly temporal. Both design choices are therefore retained in FiLMeR v1.0, with FiLM conditioning serving primarily to offset the accuracy that would otherwise be lost by adopting the dual-encoder topology alone, rather than providing a further gain beyond the single-encoder baseline. For the mixed set of target variables, FiLMeR combines continuous regression heads for thermodynamic and wind fields with a hurdle-style precipitation head and spectral constraints motivated by neural-operator forecasting methods (Li et al., 2020). Figure 1 summarizes the intended operational contrast between the conventional WRF workflow and the proposed emulator.

The contributions of this paper are therefore threefold:

- **Model documentation.** FiLMeR is presented as a documented model component for regional weather emulation, with a clear definition of the input data streams, target variables, conditioning mechanism, and training objective.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

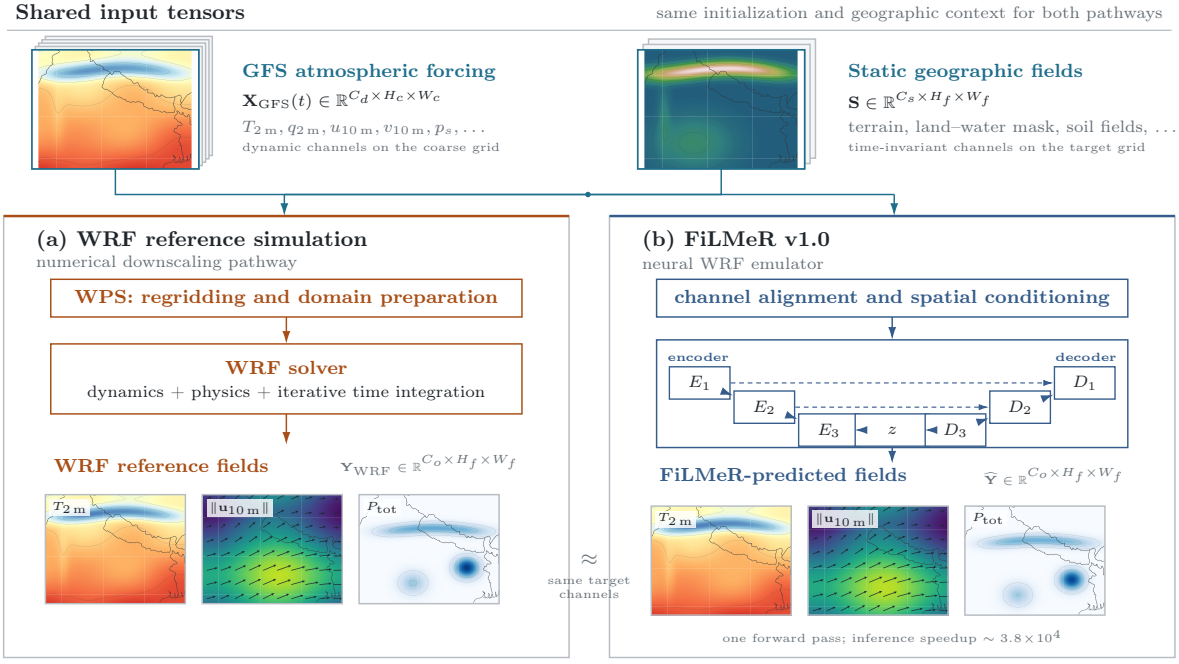


Figure 1. Traditional WRF downscaling versus the FiLMeR v1.0 emulator. WPS and WRF numerically integrate GFS and static fields to produce high-resolution regional forecasts (left); FiLMeR v1.0 learns the same mapping in a single forward pass (right).

– **Shared-weight conditioning across multiple domains and resolutions.** A single history-conditioned emulator is trained to retain useful regional structure across nested WRF domains of different geographic extents and grid spacings, without domain-specific retraining or future lateral-boundary updates at inference time. Domain identity and geographic position are carried by coordinate channels in the static input rather than by the conditioning vector itself (Section 2.1). A diagnostic test on the trained model (Section 4) shows that this shared-weight behaviour is supported mainly by the conditioning vector's temporal component rather than its resolution term, which the model makes little measurable use of on its own. All four domains used for evaluation are also present during training; extending this approach to domains or resolutions not seen during training is left for future work.

– **A controlled test of dual-encoder and FiLM contributions.** We evaluate a topology × conditioning design space spanning single- and dual-encoder topologies crossed with no, concatenation, and FiLM conditioning, using a pre-registered, seed-consistent decision rule, to isolate the separate contributions of modality separation and FiLM conditioning to the model's accuracy. A diagnostic test on the trained model further characterizes what FiLM's contribution actually consists of, not only whether it helps.

FiLMeR is not intended to replace WRF or any other numerical weather model. It learns to reproduce WRF-scale near-surface fields from historical GFS inputs, and should be understood as a fast surrogate for selected surface variables rather than

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

a complete atmospheric modelling system. At each inference step, the model requires only the two most recent GFS forecast states together with static physiographic fields and domain metadata; no future boundary information is needed. Each inference step is therefore independent, and sequential predictions over multiple time steps are obtained by repeating this procedure with updated GFS inputs rather than by feeding model outputs back as forcing. The remainder of the paper is organized as follows. Section 2 defines the emulation problem, data processing, and FiLMeR architecture. Section 3 describes the domain configuration, training setup, baselines, and evaluation protocol. Section 4 presents emulation skill, seasonal, spatial, and ablation results. Section 5 discusses model behaviour and limitations, and Section 6 concludes the study.

2 Methodology

This section describes the mathematical formulation and architecture of **FiLMeR v1.0**, illustrated in Figure 2. FiLMeR maps aligned GFS-derived atmospheric predictors, static WRF geogrid fields, and spatiotemporal metadata to WRF-scale near-surface target fields over the Indian subcontinent. The model is trained across multiple nested WRF domains with different spatial extents and grid spacings, allowing the same set of network weights to be evaluated across the four domains used in this study.

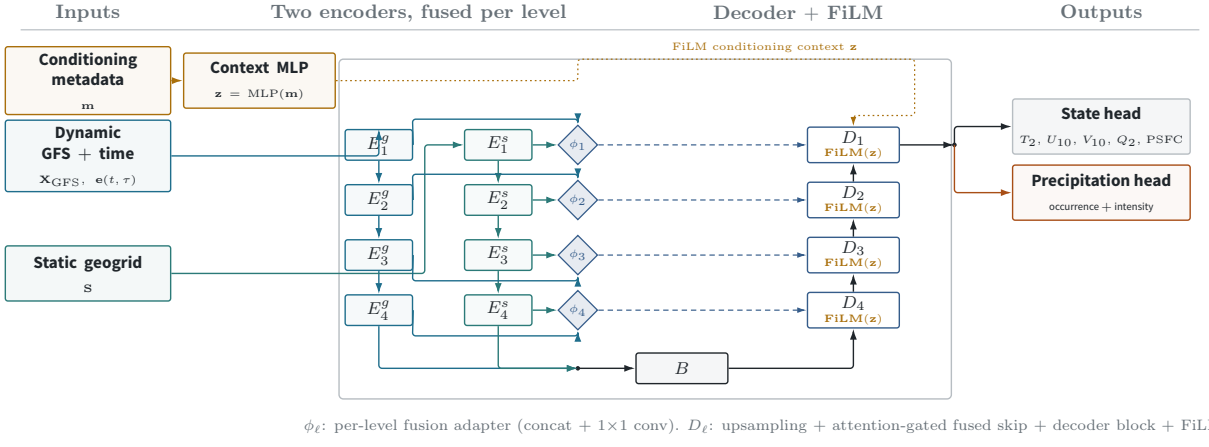


Figure 2. FiLMeR v1.0 architecture. Dynamic GFS predictors and static WRF geogrid fields are encoded by two independent four-level encoders and fused at every level before decoding. At decoder level ℓ, the corresponding fused skip feature is filtered by an attention gate and combined with the upsampled decoder representation. A 5-dimensional context vector derived from grid spacing and cyclic time-of-day/day-of-year encodings generates level-specific FiLM scale and shift parameters. Separate output heads predict the residual state correction for continuous near-surface variables and precipitation occurrence and intensity.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

2.1 Problem Formulation

Let $\Omega_g \in \mathbb{R}^{H_g \times W_g}$ represent the coarse horizontal grid of the GFS data, where H_g and W_g denote the number of latitude and longitude grid points respectively over the region of interest. Let $\Omega \in \mathbb{R}^{H \times W}$ denote a common high-resolution target grid used for every regional domain. Where a domain's native mass grid differs in size from Ω, it is bilinearly resampled to match, so the model produces a fixed output size regardless of which domain it is evaluated on; the grid size and spacing used for each domain are given in Table 2. The global atmospheric state at any given time t is represented by a tensor $\mathbf{G}_t \in \mathbb{R}^{C_g \times H_g \times W_g}$, where C_g is the number of dynamic GFS channels used at each input time; the corresponding variables and pressure levels are listed in Appendix Table B1. The global fields are sampled at uniform temporal intervals of Δt, chosen to match the native GFS forecast output frequency and the WRF output interval used for the reference targets. To capture local atmospheric trends without relying on future boundary conditions, FiLMeR acts as a history-conditioned model. At each time t, the network ingests the current and preceding global states concatenated along the channel dimension:

$$\mathbf{X}_t = [\mathbf{G}_{t-\Delta t}, \mathbf{G}_t] \in \mathbb{R}^{2C_g \times H_g \times W_g}, \tag{1}$$

where brackets denote channel-wise concatenation.

Before feature extraction, the global input $\mathbf{X}_t$ is aligned onto the target grid via a per-domain operator $\mathcal{I}_d : \Omega_g \to \Omega$ that bilinearly resamples the GFS fields using domain d's true latitude/longitude coordinates, yielding $\mathcal{I}_d(\mathbf{X}_t) \in \mathbb{R}^{2C_g \times H \times W}$; this coordinate-aware, per-domain resampling ensures that grid cell (i, j) denotes the same geographic location across the GFS, static, and target tensors for every domain.

Static surface features for domain d are provided via a geographic tensor $\mathbf{S}^{(d)} \in \mathbb{R}^{C_s \times H \times W}$ containing C_s channels: physiographic fields extracted from the WRF geogrid files (Appendix Table B2), plus two channels giving each grid cell's absolute latitude and longitude, normalized to $[-1, 1]$. These two coordinate channels carry the domain's absolute position and identity, which leaves the spatiotemporal projection vector $\mathbf{p}^{(d,t)} \in \mathbb{R}^{d_p}$ free to encode only quantities that are not tied to a specific training domain: the log-ratio of the domain's grid spacing to a reference spacing, $\log(\Delta x_d / \Delta x_{\text{ref}})$, and cyclic sine/cosine encodings of hour of day and day of year, where Δx_d is each domain's native WRF grid spacing. Geographic bounding coordinates, domain center, and nested-domain position are deliberately omitted from $\mathbf{p}^{(d,t)}$: with only four training domains, a vector encoding bounding-box identity would reduce easily to a per-domain lookup table rather than a generalizable representation of resolution and time. The full component list and the values used in this study are given in Appendix Table B3.

For the five continuous surface-state variables (T2, U10, V10, Q2, PSFC), FiLMeR predicts a residual correction to a persistence baseline. The baseline $\mathbf{B}_t^{(d)} \in \mathbb{R}^{5 \times H \times W}$ is the most recent aligned GFS surface state at those same five variables, taken directly from $\mathcal{I}_d(\mathbf{X}_t)$. The network f_θ, parameterized by weights θ, learns a mapping to the residual state correction and the precipitation fields:

$$f_\theta : \left(\mathcal{I}_d(\mathbf{X}_t), \mathbf{S}^{(d)}, \mathbf{p}^{(d,t)}\right) \mapsto \left(\Delta\widehat{\mathbf{Y}}_{t+\Delta t}^{(d)}, \widehat{o}_{\text{rain}}^{(d)}, \widehat{L}_{\text{rain}}^{(d)}\right), \tag{2}$$

where $\Delta\widehat{\mathbf{Y}}_{t+\Delta t}^{(d)} \in \mathbb{R}^{5 \times H \times W}$ is the predicted residual and the final continuous-state prediction is recovered as $\widehat{\mathbf{Y}}_{\text{state},\, t+\Delta t}^{(d)} = \mathbf{B}_t^{(d)} + \Delta\widehat{\mathbf{Y}}_{t+\Delta t}^{(d)}$. Predicting a correction to the aligned-GFS baseline conditions the learning problem on the systematic dif-

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

ference between the coarse global forcing and the regional WRF response. Precipitation, which has no natural GFS-aligned persistence analogue at this accumulation definition, is predicted directly (Section 2.2).

2.2 Model Architecture

FiLMeR v1.0 is organized around three input groups, described in Table B1, Table B2, and Table B3, which are treated separately before being combined in the decoder:

(i) **Time-dependent atmospheric predictors:** GFS-derived dynamic fields from the current and previous time steps, used to represent large-scale atmospheric state and recent temporal change.

(ii) **Static physiographic fields:** WRF geogrid variables such as terrain, land-use, soil, and vegetation-related fields, used to represent fixed lower-boundary information for each target domain.

(iii) **Spatiotemporal metadata:** A non-gridded context vector encoding grid spacing and cyclic temporal scalars (hour of day, day of year); geographic coordinates are carried separately as per-pixel channels in the static input (Section 2.1), not in this vector.

Table 1. FiLMeR v1.0 predicts six near-surface variables spanning three physical modalities: thermodynamic state, wind vectors, and precipitation.

Physical Modality	Target Variable Description	Notation	Standard Unit
Thermodynamic Fields	2 m Air Temperature	$T2$	K
	2 m Water Vapour Mixing Ratio	$Q2$	kg kg^{-1}
	Surface Pressure	$PSFC$	Pa
Wind Vectors	10 m Zonal Wind Component	$U10$	m s^{-1}
	10 m Meridional Wind Component	$V10$	m s^{-1}
Precipitation	Total Accumulated Surface Precipitation	Precipitation	mm

2.2.1 Dynamic Encoder

The dynamic encoder processes the time-dependent atmospheric predictors derived from GFS. Its input is the aligned tensor $\mathcal{I}_d(\mathbf{X}_t)$, which contains the current and previous atmospheric states after per-domain alignment to the common target grid Ω. The encoder applies a sequence of $L = 4$ convolutional blocks, each followed by 2×2 max-pooling, to extract features at multiple spatial scales:

$$\mathbf{F}_g^0 = E_g^0\left(\mathcal{I}_d(\mathbf{X}_t)\right), \quad \mathbf{F}_g^\ell = E_g^\ell\left(\text{pool}(\mathbf{F}_g^{\ell-1})\right), \quad \ell = 1, \ldots, L-1. \tag{3}$$

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Here, $\mathbf{F}_g^\ell$ denotes the pre-pooling dynamic feature map at encoder level ℓ (channel widths 32, 64, 128, 256 for $\ell = 0, \ldots, 3$), and E_g^ℓ denotes the corresponding encoder block, consisting of two 3×3 convolutions with group normalization and SiLU activation. Each $\mathbf{F}_g^\ell$ is retained as a skip connection for the decoder; the pooled output of the final level is passed to the bottleneck.

2.2.2 Static Encoder

The static encoder extracts features from the static tensor $\mathbf{S}^{(d)}$, which combines the time-invariant WRF geogrid fields (terrain, land-use, soil, and vegetation-related variables) with the two absolute-position coordinate channels described in Section 2.1. It uses an architecturally identical 4-level stack, $\mathbf{F}_s^\ell = E_s^\ell(\cdot)$, applied independently of the dynamic branch and with its own weights. Processing the two modalities in separate encoders represents fixed lower-boundary information on its own terms before it is combined with the time-dependent atmospheric predictors.

The two encoders are combined at every level through a per-level fusion adapter ϕ_{fuse}^ℓ, which concatenates the co-located dynamic and static feature maps along the channel dimension and applies a 1×1 convolution to project them to the fused skip width used by the decoder:

$$\mathbf{F}^\ell = \phi_{\text{fuse}}^\ell \left(\left[\mathbf{F}_g^\ell, \mathbf{F}_s^\ell \right] \right), \quad \ell = 0, \ldots, L-1. \tag{4}$$

The fused feature maps $\mathbf{F}^\ell$ (channel widths 64, 128, 256, 512) serve as the decoder's skip connections at every scale. The two encoders' deepest pooled outputs are concatenated directly, without a learned adapter, and passed to a bottleneck convolution block before decoding begins.

2.2.3 Spatiotemporal Projection Embedding

FiLMeR also uses the non-gridded metadata vector $\mathbf{p}^{(d,t)} \in \mathbb{R}^5$ introduced in Section 2.1 and listed in Appendix Table B3. A two-layer multi-layer perceptron (MLP) with SiLU activations, denoted by ϕ_p, maps this metadata vector to a continuous embedding:

$$\mathbf{e}_p = \phi_p(\mathbf{p}^{(d,t)}), \tag{5}$$

where $\mathbf{e}_p \in \mathbb{R}^{d_e}$ is the spatiotemporal embedding and $d_e = 128$.

This embedding is used to condition the decoder. In this design, grid-spacing and time information enter the model through the FiLM conditioning layers, which adjust decoder feature maps accordingly; domain identity and absolute position instead enter through the coordinate channels in the static input (Section 2.1), not through FiLM. Section 4.3 shows that, in the trained model, FiLM's measurable contribution is predominantly the temporal component.

2.2.4 FiLM-Conditioned Decoder

The decoder maps the fused feature representation back to the common target grid Ω. It uses upsampling layers and skip connections from the encoder, following the general U-Net design (Ronneberger et al., 2015), to combine coarse-scale information

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

with spatial details from earlier feature maps. At each decoder level ℓ, FiLM conditioning is applied using the spatiotemporal embedding $\mathbf{e}_p$.

The embedding $\mathbf{e}_p$ is passed through level-specific linear layers to produce channel-wise scale and shift parameters:

$$\boldsymbol{\gamma}_\ell = \phi_\gamma^\ell(\mathbf{e}_p), \tag{6}$$

$$\boldsymbol{\beta}_\ell = \phi_\beta^\ell(\mathbf{e}_p), \tag{7}$$

$$\widetilde{\mathbf{F}}_\ell = \boldsymbol{\gamma}_\ell \odot \mathbf{F}_\ell + \boldsymbol{\beta}_\ell, \tag{8}$$

where $\mathbf{F}_\ell$ is the decoder feature map after upsampling and skip-connection fusion, $\odot$ denotes channel-wise multiplication broadcast over the spatial grid, and $\widetilde{\mathbf{F}}_\ell$ is the FiLM-conditioned feature map.

This operation conditions the decoder features on grid-spacing and time information while keeping the convolutional weights shared across the evaluated domains. FiLM conditioning is applied at the decoder levels only in the FiLMeR v1.0 configuration. A variant that additionally supplies the context embedding to a gated version of the per-level fusion adapter (DF2, not FiLM applied at every encoder and decoder block) is evaluated in the ablation study (Section 4.3) and does not show a consistent improvement over decoder-only conditioning across metrics; decoder-only conditioning was retained as the default because it keeps the model compact, with fewer additional affine parameter sets, for the multi-domain shared-weight setting.

2.2.5 Spatial Attention Gates

Spatial attention gates (Oktay et al., 2018) are applied to the fused encoder-to-decoder skip connections $\mathbf{F}^\ell$ (Eq. 4). These gates weight the fused skip features before they are passed to the decoder:

$$\mathbf{A}_\ell = \sigma\left(\mathbf{W}_g * \mathbf{F}_{\text{dec}}^\ell + \mathbf{W}_x * \mathbf{F}_x^\ell + \mathbf{b}\right), \tag{9}$$

$$\mathbf{F}_x^{\ell,\text{att}} = \mathbf{A}_\ell \odot \mathbf{F}_x^\ell, \tag{10}$$

where $\mathbf{F}_{\text{dec}}^\ell$ is the decoder gating feature at level ℓ (distinct from the dynamic-encoder feature $\mathbf{F}_g^\ell$ of Eq. 3), $\mathbf{F}_x^\ell = \mathbf{F}^\ell$ is the corresponding fused skip feature, $*$ denotes a 1×1 convolution, σ is the sigmoid activation function, and $\mathbf{A}_\ell$ is the spatial attention map.

The weighted skip feature $\mathbf{F}_x^{\ell,\text{att}}$ is then passed to the decoder. This allows the model to adjust the contribution of skip-connection features at each spatial location.

2.2.6 Mixed-Modality Multi-Variable Output Heads

The final decoder feature map $\widetilde{\mathbf{F}}^0$ is passed to separate output heads for continuous state variables and precipitation. The continuous state head predicts the residual correction (Section 2.1) for the five continuous surface variables:

$$\Delta\widehat{\mathbf{Y}}^{(d)} = h_{\text{state}}(\widetilde{\mathbf{F}}^0) \in \mathbb{R}^{5\times\Omega}. \tag{11}$$

Adding the aligned-GFS baseline $\mathbf{B}_t^{(d)}$ to $\Delta\widehat{\mathbf{Y}}^{(d)}$ recovers the physical-unit prediction for T2, U10, V10, Q2, and PSFC.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Precipitation is treated separately because it contains many zero values and has a skewed distribution. The precipitation head follows a hurdle-style formulation with two outputs: an occurrence logit and a log-intensity value:

$$\widehat{o}_{\text{rain}}^{(d)} = h_{\text{occ}}(\widetilde{\mathbf{F}}^0) \in \mathbb{R}^{1\times\Omega}, \tag{12}$$

$$\widehat{L}_{\text{rain}}^{(d)} = \text{softplus}\left(h_{\text{int}}(\widetilde{\mathbf{F}}^0)\right) \in \mathbb{R}_{\geq 0}^{1\times\Omega}, \tag{13}$$

where $\widehat{o}_{\text{rain}}^{(d)}$ is an occurrence logit at each grid point and $\widehat{L}_{\text{rain}}^{(d)}$ is a non-negative log-intensity value, trained to match $\log(1+I_{\text{rain}})$ directly (Eq. 17). This separation allows occurrence and intensity to be represented with different output terms. Millimetre-valued precipitation is recovered only at inference time, by combining both outputs:

$$\widehat{R}_{\text{rain}}^{(d)} = \left(\exp\left(\min\left(\widehat{L}_{\text{rain}}^{(d)}, 10\right)\right) - 1\right) \cdot \mathbb{1}\left[\sigma\left(\widehat{o}_{\text{rain}}^{(d)}\right) > 0.5\right], \tag{14}$$

where the log-intensity is clamped at 10 before inversion, bounding implausible extrapolated values, and the occurrence probability $\sigma(\widehat{o}_{\text{rain}}^{(d)})$ is thresholded at 0.5 to zero out grid cells predicted dry. All precipitation results reported in Section 4 are computed from $\widehat{R}_{\text{rain}}^{(d)}$.

3 Experimental Setup

This section describes the experimental configuration used to train and evaluate FiLMeR v1.0. We first define the WRF domain configuration and geographic coverage, summarized in Table 2. We then describe the temporal data split, training setup, baseline models, and evaluation metrics used to compare FiLMeR with alternative neural architectures.

3.1 Domains and Dataset Splits

The experiments use four WRF domain configurations over the Indian subcontinent: one parent domain and three nested regional domains, all described in Table 2. The parent domain (`d01`) covers the broader Indian subcontinent at a horizontal grid spacing of 27 km with a 130×130 staggered grid (129×129 on the unstaggered mass grid used for the surface fields modeled here; Section 2.1), while the three nested domains (`d02`-`d04`) each use a 9 km grid spacing with 100×100 staggered (99×99 unstaggered) grid points, covering northern, southern, and eastern India respectively. Because d01's native 129×129 grid is bilinearly resampled down to the common 99×99 target grid used for every domain (Section 2.1), the tensor actually delivered to the network for d01 has an effective spacing of approximately 35.3 km, coarser than its native 27 km spacing. Dynamic GFS inputs use $C_g = 20$ atmospheric channels at 0.25° resolution, interpolated to each target domain grid via bilinear resampling. The target tensor encompasses the $Nv = 6$ surface weather variables described in Table 1. The GFS fields are sampled at $\Delta t = 3$ h, which matches the native GFS forecast output frequency and the WRF output interval used for reference targets. The spatiotemporal projection vector has dimension $d_p = 5$, with a single spatial component (log grid-spacing ratio) and four temporal components (cyclic hour-of-day and day-of-year encodings), as detailed in Appendix Table B3. Static physiographic inputs comprise 30 channels extracted from the WRF geogrid files, to which two per-pixel absolute-position channels are appended (Section 2.1), giving $C_s = 32$ static input channels in total. Training and evaluating the same FiLMeR v1.0 model

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

across these domains, spanning two different resolutions and four distinct geographic extents, tests whether one set of network weights can be applied across all four domains.

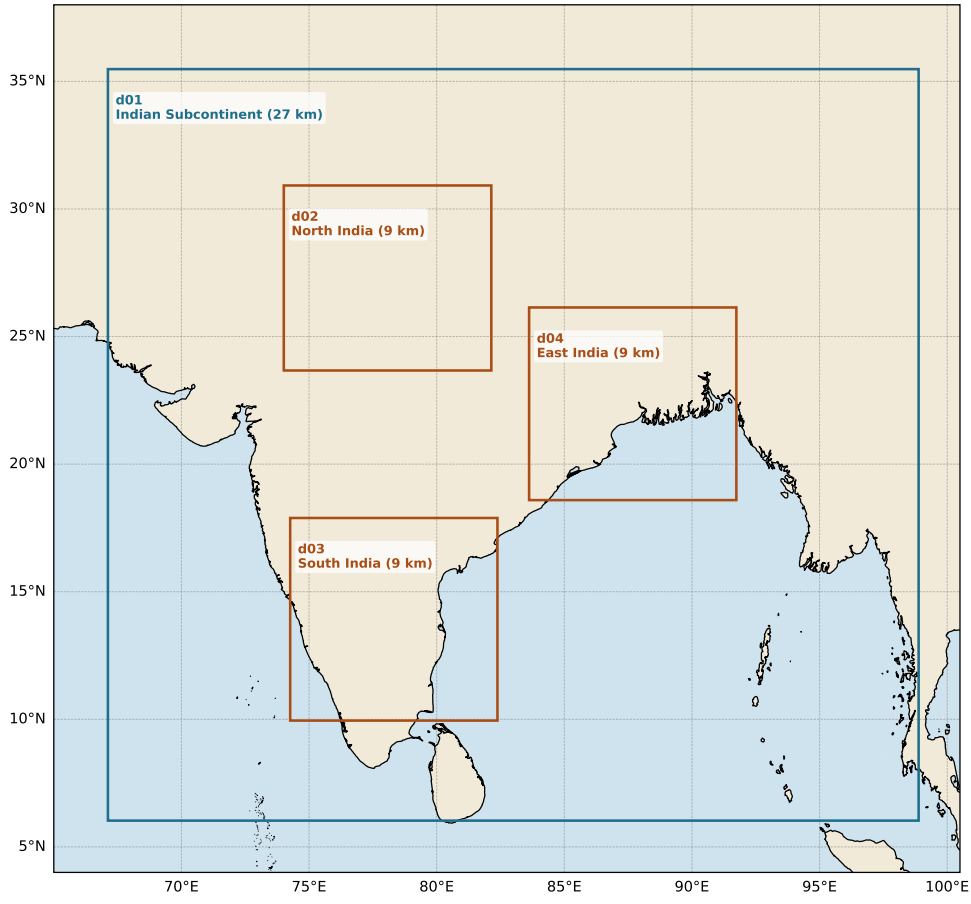


Figure 3. A single shared-weight FiLMeR model is evaluated across four WRF domains spanning two resolutions and three contrasting geographic regimes: the parent domain (**d01**, 27 km) covers the broader Indian subcontinent, while the three nested domains (**d02-d04**, 9 km) cover Himalayan terrain in northern India, peninsular coastlines in southern India, and humid eastern plains, providing a diverse multi-domain test bed for the shared-weight conditioning approach. Basemap shows physical coastline and land/ocean colouring only (Natural Earth), with no administrative or political boundaries; domain boxes generated by the authors from the coordinates in Table 2.

The dataset is split by time, using a temporally blocked (not random) split so that autocorrelated, adjacent 3-hourly samples cannot appear on both sides of the split. The training block runs from 2023-01-01 06:00 to 2024-10-17 06:00 (20,864 samples, split evenly across the four domains at 5,216 samples each, since every domain is sampled at every training timestamp). A 72-hour embargo separates the training block from the validation block, which runs from 2024-10-20 06:00 to 2025-01-01 00:00 (2,327 samples, approximately 581-582 per domain). These two years include a range of seasonal conditions over the study region, including pre-monsoon, monsoon, post-monsoon, and winter periods. Channel-wise normalization statistics are

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table 2. Four WRF-ARW domains spanning two resolutions (27 km and 9 km) provide the multi-domain training and evaluation configuration for FiLMeR v1.0, with nested domains targeting three distinct subregional geographic settings across the Indian subcontinent.

Domain	Target Region	Lat. min. (°)	Lon. min. (°)	Lat. max. (°)	Lon. max. (°)	Grid Resolution (km)
d01	Indian Subcontinent	6.03	67.12	35.48	98.88	27
d02	North India	23.67	74.01	30.92	82.14	9
d03	South India	9.95	74.26	17.89	82.38	9
d04	East India	18.59	83.62	26.14	91.74	9

computed only from the training block and then reused for validation and testing to avoid data leakage. The validation block is used for checkpoint selection and early stopping (Appendix Table C1); it is not used for weight optimization. The year 2025 is reserved as a fully independent test period, held out from both weight optimization and validation-based model selection.

3.2 Data Sources and Preprocessing

The framework relies on a multi-modal dataset comprising global atmospheric forcing, static physiographic boundary constraints, and high-resolution numerical reference targets. Detailed configurations for the WRF-ARW reference simulations, including physics parameterizations, vertical discretization, and domain settings, are provided in Appendix A1. Dynamic inputs are drawn from the NCEP GFS operational forecast archive at 0.25° resolution, not from a reanalysis product. GFS is initialized three times each month, at 00 UTC on the 1st, 11th, and 21st, and each initialization is used out to a forecast lead of 240 hours in 3-hourly steps. The paired WRF-ARW simulations are driven forward from these same GFS initializations and lead times, so the full dataset used for training, validation, and testing is built from a series of 10-day, GFS-driven WRF integrations. Static lower-boundary forcing is extracted from the WPS, including topography, land-use, soil, and seasonally-indexed vegetation metrics; a detailed list is provided in Appendix B2. All continuous inputs and states are standard-scaled via channel-wise z-scores based on training-set statistics. Given the intermittent, zero-inflated nature of precipitation, we apply a $\log(1+x)$ transformation. The global GFS fields are stored on a fixed 127×137 crop of the 0.25° operational grid (Appendix B) and are resampled onto each domain's own 99×99 target grid using a bilinear sampling grid built from that domain's true latitude/longitude coordinates (Section 2.1). For the three 9 km nested domains (d02-d04), 99×99 is exactly the native unstaggered mass-grid size; for the 27 km parent domain (d01), whose native mass grid is 129×129, the same 99×99 target size is reached by bilinear resampling, so every domain is processed at a common output resolution.

3.3 Training and Optimization Details

All models are trained using the same optimization framework so that differences in performance can be attributed mainly to the model architecture. The main training hyperparameters, optimizer settings, learning-rate schedule, batch size, and stopping criteria are summarized in Appendix Table C1.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

The training objective combines losses for the continuous state variables and precipitation through a weighted summation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{state}} + \lambda_{\text{precip}}\,\mathcal{L}_{\text{precip}}, \tag{15}$$

where $\lambda_{\text{precip}} = 1.0$ in all experiments reported here, making the two top-level terms equally weighted.

3.3.1 Continuous State Loss

$\mathcal{L}_{\text{state}}$ is applied to the normalized residual correction for the five continuous output variables (T2, U10, V10, Q2, and PSFC; Section 2.1) and combines pointwise, spatial-gradient, and spectral terms:

$$\mathcal{L}_{\text{state}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{grad}}\,\mathcal{L}_{\text{grad}} + \lambda_{\text{fft}}\,\mathcal{L}_{\text{fft}}, \tag{16}$$

with $\lambda_{\text{grad}} = 0.1$ and $\lambda_{\text{fft}} = 0.05$. The three terms serve distinct roles:

- $\mathcal{L}_{\text{MSE}} = \frac{1}{N}\|\Delta\widehat{\mathbf{Y}} - \Delta\mathbf{Y}\|_2^2$, the mean squared error over the N elements of the residual tensor, penalizes pointwise amplitude errors between the predicted and reference residuals.
- $\mathcal{L}_{\text{grad}}$ compares horizontal and vertical spatial finite differences of the predicted and reference fields. This term discourages spatially smooth predictions that minimize MSE while losing local gradients near coastlines, terrain boundaries, and convective edges.
- $\mathcal{L}_{\text{fft}}$ compares the amplitudes of the two-dimensional discrete Fourier transforms of the predicted and reference fields. This term penalizes spectral energy misrepresentation across spatial scales and follows the spectral-loss formulation introduced for neural operators (Li et al., 2020).

The weights λ_{grad} and λ_{fft} were selected so that the gradient and spectral terms each contribute roughly one order of magnitude less than the MSE term at the start of training, preventing them from dominating early gradient updates while still exerting a regularizing influence on spatial structure throughout training. These values were fixed after preliminary sensitivity runs and were not tuned per domain or per variable.

3.3.2 Precipitation Loss

Precipitation is treated with a separate hurdle-style formulation because the rainfall field contains a large fraction of zero values and has a strongly right-skewed intensity distribution. A single loss term over the full field would be dominated by the dry-cell majority and would poorly represent the intensity errors that matter for regional applications. The precipitation loss decomposes into an occurrence term and an intensity term:

$$\mathcal{L}_{\text{precip}} = w_{\text{bce}}\,\text{BCE}(\widehat{o}_{\text{rain}}, \mathbf{P}_{\text{occ}}) + w_{\text{int}}\,\text{Huber}\Big(\widehat{L}_{\text{rain}}, \log(1 + \mathbf{I}_{\text{rain}})\Big), \tag{17}$$

with $w_{\text{bce}} = w_{\text{int}} = 0.5$, giving equal weight to the two terms. Here $\widehat{o}_{\text{rain}}$ and $\mathbf{P}_{\text{occ}}$ are the predicted occurrence logit and the binary reference wet-cell mask respectively, and $\widehat{L}_{\text{rain}}$ and $\mathbf{I}_{\text{rain}}$ are the predicted log-intensity and the reference precipitation

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

intensity in millimetres. The binary cross-entropy term BCE$(\cdot)$ operates directly on the occurrence logit and trains the occurrence head to distinguish wet from dry grid cells. The intensity term is evaluated only over reference wet cells and compares the predicted log-intensity directly against $\log(1+\mathbf{I}_{\text{rain}})$, so no further log transform is applied to the prediction; a Huber loss is used in place of a squared-error term to limit the influence of a small number of extreme wet-cell outliers on the gradient. The equal weighting $w_{\text{bce}} = w_{\text{int}} = 0.5$ balances the gradient contributions of the occurrence and intensity heads, which otherwise differ substantially in magnitude because the cross-entropy operates on logits while the log-intensity term can span a wider numerical range.

3.4 Computational Efficiency Benchmark

To quantify the inference cost of FiLMeR relative to the reference WRF workflow, we conducted a controlled timing comparison under fixed hardware conditions. The WRF-ARW dynamical simulation was executed on a high-performance compute node using AMD EPYC 7452 32-Core Processor. Under this configuration, generating a 10-day regional forecast over a 9 km nested domain required approximately 4.5 h (∼16,200 s) of wall-clock time. FiLMeR inference was benchmarked on a single NVIDIA A100-SXM4 (80 GB) GPU at a batch size of 1. Producing the corresponding 80-step sequence (10 days at $\Delta t = 3$ h intervals) required 0.422 s of wall-clock time, yielding an effective speedup of ∼38,300× relative to the dynamical WRF integration under this setup. Both timings isolate the core computational step: disk I/O, preprocessing, and netCDF output are excluded from both measurements, as these are contingent on storage infrastructure. The reported speedup therefore reflects the ratio of dynamical time-stepping cost to neural-network forward-pass cost and should be interpreted in that context.

3.5 Baseline Configurations

FiLMeR's dual-encoder, decoder-only-FiLM design (topology = dual, conditioning = film, denoted **DF** below) is one point in a small topology × conditioning design space built on the shared backbone of Section 2.2: topology is either a single early-fusion encoder over the concatenated dynamic and static inputs (widths $64 \to 128 \to 256 \to 512$) or the two independent encoders described in Section 2.2 (widths $32 \to 64 \to 128 \to 256$ each); conditioning is either absent, injected by concatenating $\mathbf{p}^{(d,t)}$ into every decoder skip connection through a 1×1 convolutional adapter, or injected through the FiLM mechanism of Section 2.2. We use the single-encoder, unconditioned configuration (**S0**) and the single-encoder, concat-conditioned configuration (**SC**) as baselines against which **DF** is compared: S0 establishes the error level attainable with neither dual-encoder modality separation nor conditioning, and SC establishes the error level attainable with conditioning present but without modality separation or FiLM, isolating what each of DF's two design choices contributes on top of a matched backbone. The remaining three points in the design space (single-encoder with FiLM, SF; dual-encoder with concat conditioning, DC; and dual-encoder with FiLM applied additionally at the fusion adapters, DF2) are examined together with S0, SC, and DF in the ablation study (Section 4.3). All seven configurations are trained with the same loss function, optimization setup, and normalization statistics, and the unconditioned and concatenation-based conditioning adapters are matched in capacity, so that performance differences reflect the information available to the model rather than differences in model size. Implementation details are given in Ap-

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

pendix D; trainable parameter counts range from 36.45M (D0, DC) to 38.86M (SF), a spread of under 7% across the full design space.

3.6 Evaluation Methodology

The independent test set is based on WRF-ARW reference simulations from 2025 for all four domains ($d01$-$d04$). To sample the annual cycle while limiting the cost of generating reference fields, we evaluate the first 10 days of each month, which corresponds to one complete GFS-initialized WRF forecast cycle (00 UTC on the 1st), giving 12 cycles in total across the year and 3,788 test samples overall (947 per domain). The evaluation follows metric conventions established in regional NWP emulation and data-driven downscaling studies (Lam et al., 2023; Pathak et al., 2026; Xu et al., 2025; Rasp et al., 2020), using separate metric groups for continuous thermodynamic variables, wind components, and precipitation, as summarized in Table 3.

Table 3. Evaluation metrics are grouped by physical modality, combining pointwise error, structural similarity, and categorical precipitation skill scores to assess complementary aspects of emulation quality.

Variable Class	Target Channels	Evaluation Metrics
Thermodynamic States	T2, Q2, PSFC	Latitude-weighted RMSE, SSIM
Wind Components	U_{10}, V_{10}	Component RMSE, RMSVE
Precipitation	RAIN_{tot}	Masked intensity MAE/RMSE
		FAR, CSI, Frequency Bias

3.6.1 Thermodynamic Evaluation Metrics

For T2, Q2, and PSFC, we report latitude-weighted RMSE, following the area-weighted error convention used in global and regional learned forecasting systems (Lam et al., 2023; Bi et al., 2023). The latitude weighting accounts for differences in grid-cell area with latitude:

$$\text{RMSE}_w = \sqrt{\frac{\sum_{i,j} \cos(\phi_{i,j}) \left(\widehat{\mathbf{Y}}_{i,j} - \mathbf{Y}_{i,j}\right)^2}{\sum_{i,j} \cos(\phi_{i,j})}}, \tag{18}$$

where $\phi_{i,j}$ is the latitude of grid point (i, j). We additionally compute the Structural Similarity Index Measure (SSIM) (Wang et al., 2004) to assess spatial pattern fidelity beyond pointwise errors, consistent with its use in regional weather emulation evaluation (Chen et al., 2026). SSIM's dynamic range is set, per channel, to that channel's own minimum-to-maximum value within the evaluated sample population (the full pooled test set for the pooled tables, or a single domain's test set for the per-domain tables).

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

3.6.2 Wind Vector Evaluation Metrics

For 10 m wind, we report component-wise RMSE alongside the Root Mean Square Vector Error (RMSVE), which jointly measures zonal and meridional errors and is standard in NWP verification (Skamarock et al., 2008):

$$\mathrm{RMSVE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[\left(\widehat{U}_{10,n} - U_{10,n}\right)^2 + \left(\widehat{V}_{10,n} - V_{10,n}\right)^2\right]}. \tag{19}$$

3.6.3 Hydrological and Intermittent Precipitation Metrics

For precipitation, we evaluate categorical scores at accumulation thresholds of $\tau \in \{0.1, 1.0, 5.0, 10.0\}$ mm, following the threshold-based verification framework of Schaefer (1990) and Roberts and Lean (2008). At each threshold, grid cells are classified as hits, false alarms, misses, or correct negatives, from which we derive False Alarm Ratio (FAR), Critical Success Index (CSI), and Frequency Bias. These three metrics are jointly interpreted because CSI conflates spatial coverage bias with detection skill; a model with Frequency Bias > 1 will mechanically obtain higher CSI when probability of detection is similar across models (Schaefer, 1990). Masked intensity errors (MAE and RMSE over reference wet cells only) complement the categorical scores by isolating conditional intensity performance from occurrence skill.

4 Results

This section presents the evaluation of FiLMeR v1.0 on the independent test period. The results focus on two main questions: whether separate dynamic and static encoders improve the representation of regional surface fields, and whether spatiotemporal conditioning allows the same model weights to be applied across the four WRF domains.

4.1 Baseline Comparison and Emulation Skill

We compare **DF** (FiLMeR v1.0: dual-encoder, decoder-only FiLM) against the single-encoder baselines **S0** (unconditioned) and **SC** (concat-conditioned) introduced in Section 3.5, pooled across all four domains over the independent 2025 test period and reported as mean $\pm$ standard deviation across three training seeds (42, 43, 44). The full seven-configuration matrix, including the remaining design-space points (SF, D0, DC, DF2) and the pre-registered primary decision score, is given in Section 4.3.

4.1.1 Continuous States: Thermodynamics and Wind Vectors

Tables 4 and 5 summarize performance for the continuous near-surface variables. All three configurations substantially outperform the trivial reference points established in Section 4.3 (domain climatology and aligned-GFS persistence, both computed once on the DF seed-42 test predictions), confirming that the model as a whole, regardless of conditioning choice, is learning a genuine regional correction. Within the design space, the largest single step is from S0 to any conditioned configuration: PSFC RMSE falls from 1.55 to 1.37-1.38 hPa and T2 RMSE from 1.81 to 1.49-1.50 K when conditioning is added, in either form.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table 4. Adding conditioning (SC, DF) improves substantially over the unconditioned single-encoder baseline (S0) for surface pressure and temperature. SC and DF are close to each other on every continuous thermodynamic metric. Q2 is reported in g/kg here, converted from its native kg/kg WRF units (Table 1) for readability.

Model	**PSFC (hPa)**		**T2 (K)**		**Q2 (g/kg)**	
	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑
S0 (unconditioned)	1.55 ± 0.08	1.00 ± 0.00	1.81 ± 0.13	0.96 ± 0.00	1.56 ± 0.01	$\mathbf{0.85 \pm 0.00}$
SC (concat)	1.38 ± 0.04	1.00 ± 0.00	$\mathbf{1.49 \pm 0.02}$	$\mathbf{0.97 \pm 0.00}$	$\mathbf{1.55 \pm 0.02}$	$\mathbf{0.85 \pm 0.01}$
DF (FiLMeR v1.0)	$\mathbf{1.37 \pm 0.04}$	1.00 ± 0.00	1.50 ± 0.01	0.96 ± 0.00	1.56 ± 0.01	0.84 ± 0.00

Table 5. Wind-vector error follows the same pattern as the thermodynamic variables: conditioning (SC, DF) gives a small, consistent improvement over the unconditioned baseline (S0), and DF is marginally ahead of SC rather than clearly separated from it.

Model	**U10 RMSE (m/s) ↓**	**V10 RMSE (m/s) ↓**	**Vector RMSVE (m/s) ↓**
S0 (unconditioned)	1.78 ± 0.05	1.75 ± 0.02	2.50 ± 0.04
SC (concat)	1.77 ± 0.02	$\mathbf{1.73 \pm 0.00}$	2.48 ± 0.01
DF (FiLMeR v1.0)	$\mathbf{1.75 \pm 0.01}$	$\mathbf{1.73 \pm 0.01}$	$\mathbf{2.46 \pm 0.01}$

Between the two conditioned configurations, DF is slightly ahead of SC on PSFC RMSE (1.37 vs. 1.38 hPa) and RMSVE (2.46 vs. 2.48 m s^{-1}), marginally behind SC on T2 RMSE (1.50 vs. 1.49 K) and Q2 RMSE (1.56 vs. 1.55 g kg^{-1}). These differences fall within cross-seed noise; Section 4.3 evaluates whether dual-encoder modality separation or FiLM conditioning improves on a simpler concat-conditioned single encoder directly, using a pre-registered decision score across the full design space. PSFC SSIM is saturated at 1.00 for all three configurations in Table 4: PSFC's spatial pattern over these domains is dominated by terrain elevation and is nearly deterministic given the static geogrid fields, so all configurations reproduce it well enough that SSIM no longer discriminates between them at the precision reported here; RMSE remains informative for PSFC because it is sensitive to the small residual differences that SSIM saturates over.

4.1.2 Intermittent Hydrological Extremes

Table 6 reports categorical precipitation performance at light (≥ 0.1 mm) and moderate (≥ 10.0 mm) accumulation thresholds, while Table 7 reports masked precipitation-intensity errors over reference wet grid cells only. Unlike the continuous thermodynamic variables, precipitation shows no consistent ordering among S0, SC, and DF: S0 and SC give the lowest moderate-rain bias deviation from unity, S0 gives the highest light- and moderate-rain CSI, and DF gives the lowest moderate-rain FAR and the lowest masked MAE but not the lowest masked RMSE. Differences across all three configurations are within, or close to, one cross-seed standard deviation at every threshold. This indicates that conditioning and dual-encoder modality separation

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table 6. Categorical precipitation skill is nearly identical across S0, SC, and DF at both the light- and moderate-rain thresholds; none of the three configurations dominates on FAR, CSI, and frequency bias simultaneously.

Model	**Light Rain** (≥ 0.1 **mm**)			**Moderate Rain** (≥ 10.0 **mm**)		
	FAR ↓	CSI ↑	Bias (→1)	FAR ↓	CSI ↑	Bias (→1)
S0 (unconditioned)	**0.22 ± 0.03**	**0.75 ± 0.01**	1.24 ± 0.06	0.29 ± 0.03	**0.54 ± 0.02**	**0.96 ± 0.10**
SC (concat)	**0.22 ± 0.01**	**0.75 ± 0.01**	**1.22 ± 0.01**	0.29 ± 0.01	**0.54 ± 0.01**	**0.96 ± 0.04**
DF (FiLMeR v1.0)	0.23 ± 0.01	0.74 ± 0.00	1.25 ± 0.03	**0.27 ± 0.03**	0.52 ± 0.04	0.88 ± 0.15

Table 7. Masked intensity errors, evaluated over reference wet cells only ($\geq$0.1 mm), show DF with the lowest MAE but the highest RMSE among the three configurations, indicating that its intensity errors are, on average, smaller but more occasionally large than SC's or S0's: occurrence calibration and spatial placement, not conditional intensity, remain the larger source of the differences seen in Table 6.

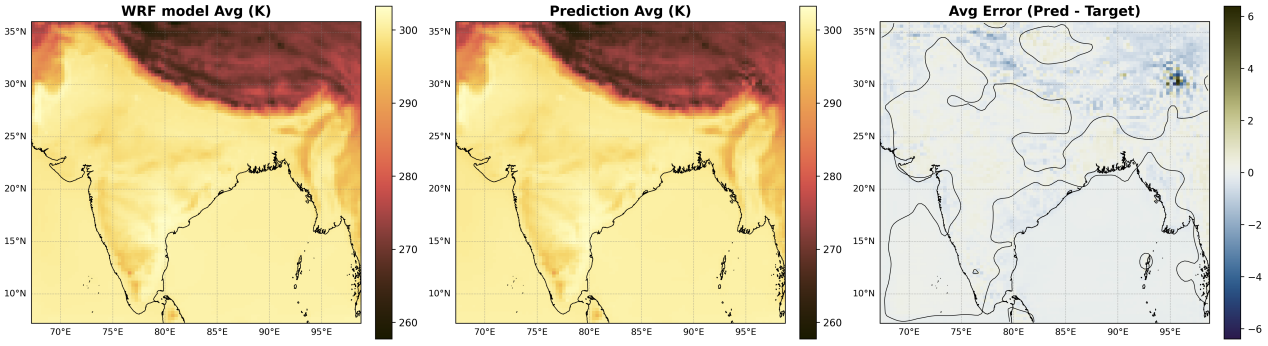
Model	MAE (mm) ↓	RMSE (mm) ↓
S0 (unconditioned)	24.73 ± 0.58	**51.80 ± 0.56**
SC (concat)	24.63 ± 0.20	52.15 ± 0.30
DF (FiLMeR v1.0)	**24.21 ± 0.35**	52.96 ± 0.62

contribute far less to precipitation skill than to the continuous thermodynamic and wind fields. Precipitation is discussed further as part of the full design-space comparison in Section 4.3.

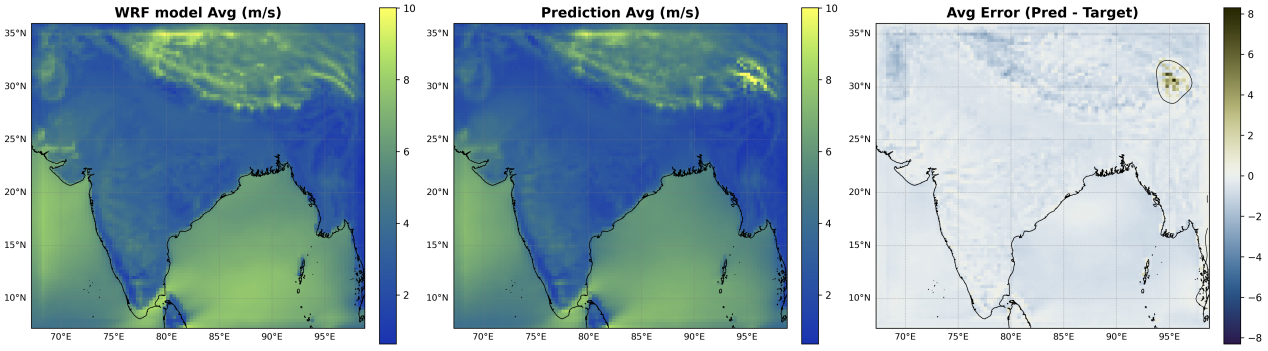
4.1.3 Geographic Fidelity and Spatial Error Characterization

Figure 4 shows time-averaged DF (FiLMeR v1.0) fields and associated spatial error patterns over domain d01 for 2 m temperature, 10 m wind speed, and precipitation, providing spatial context that domain-mean scores alone do not describe. The temperature field in Fig. 4a retains broad physiographic structure, including cooler values over elevated terrain and warmer conditions over lower-elevation regions, with residual error small and spatially diffuse over most of the domain aside from a localized hotspot in the northeastern high-terrain corner. The wind-speed field in Fig. 4b preserves large-scale land-sea contrasts and regional circulation features, with the same northeastern hotspot but otherwise small, diffuse errors elsewhere. The precipitation field in Fig. 4c shows spatially intermittent rainfall maxima and dry regions, and the error map reveals a systematic pattern: the model underpredicts accumulated precipitation along the western coastal margin and the northeastern high-rainfall belt, the two regions with the highest target-field accumulation. This is consistent with the below-unity moderate-rain frequency bias already noted in Table 6, and indicates that the model's precipitation errors are concentrated in exactly the high-accumulation regions that matter most for regional hydrological applications. Overall, the figure complements the aggregate statistics by showing that DF reproduces several geographically organized structures while revealing that its largest, most systematic errors are spatially concentrated.

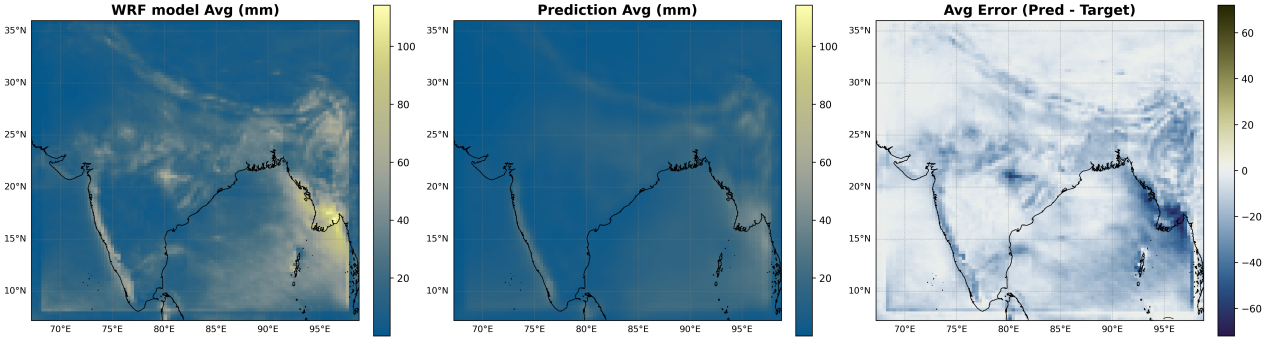
https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026



(a) 2 m Temperature



(b) 10 m Wind Speed



(c) Precipitation

Figure 4. DF (FiLMeR v1.0), averaged across 3 seeds, reproduces the broad spatial climatology of temperature, wind speed, and precipitation across the Indian subcontinent (domain d01). Residual errors for T2 and wind speed are small and spatially diffuse over most of the domain, with a shared localized error hotspot in the northeastern high-terrain corner. Precipitation shows a more systematic pattern: the model underpredicts accumulated precipitation (negative Pred−Target bias) along the western coastal margin and the northeastern high-rainfall belt, consistent with the below-unity moderate-rain frequency bias reported in Table 6. Maps were generated by the authors using `matplotlib` and `cartopy` from model output data.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table 8. Seasonal breakdown of the S0/SC/DF comparison from Section 4.3. Conditioning (SC, DF) improves over the unconditioned baseline (S0) in every season for T2, PSFC, and precipitation CSI, while DF and SC remain close to each other throughout the annual cycle, mirroring the pooled-annual result in Tables 4-5.

Season	Metric / Variable	S0	SC	DF
Winter (DJF)	T2 (K) ↓	1.88 ± 0.09	1.52 ± 0.06	**1.54 ± 0.04**
	PSFC (hPa) ↓	1.18 ± 0.13	**1.02 ± 0.11**	1.01 ± 0.09
	Wind Vector (m/s) ↓	1.85 ± 0.01	1.87 ± 0.04	**1.85 ± 0.03**
	Precip 1.0mm CSI ↑	0.39 ± 0.01	**0.43 ± 0.01**	**0.43 ± 0.02**
Pre-monsoon (MAM)	T2 (K) ↓	2.17 ± 0.39	**1.66 ± 0.04**	1.65 ± 0.05
	PSFC (hPa) ↓	1.44 ± 0.10	**1.27 ± 0.05**	1.29 ± 0.03
	Wind Vector (m/s) ↓	**2.66 ± 0.01**	2.64 ± 0.02	**2.66 ± 0.01**
	Precip 1.0mm CSI ↑	**0.53 ± 0.03**	**0.55 ± 0.01**	0.52 ± 0.06
Monsoon (JJAS)	T2 (K) ↓	1.56 ± 0.04	1.39 ± 0.01	**1.40 ± 0.01**
	PSFC (hPa) ↓	1.72 ± 0.05	**1.54 ± 0.02**	1.52 ± 0.02
	Wind Vector (m/s) ↓	**2.66 ± 0.05**	**2.66 ± 0.06**	2.61 ± 0.02
	Precip 1.0mm CSI ↑	**0.79 ± 0.00**	**0.80 ± 0.01**	0.79 ± 0.01
Post-monsoon (ON)	T2 (K) ↓	1.68 ± 0.02	**1.44 ± 0.05**	1.42 ± 0.02
	PSFC (hPa) ↓	1.85 ± 0.06	1.68 ± 0.05	**1.65 ± 0.03**
	Wind Vector (m/s) ↓	2.73 ± 0.11	**2.64 ± 0.09**	2.66 ± 0.02
	Precip 1.0mm CSI ↑	0.72 ± 0.01	**0.72 ± 0.01**	**0.73 ± 0.02**

4.2 Seasonal Robustness Across the Sampled 2025 Test Period

Seasonal evaluation provides a check on whether model errors remain stable as the regional circulation and near-surface thermodynamic state change through the year. For the Indian subcontinent, this includes winter conditions, pre-monsoon heating, the summer monsoon, and the post-monsoon transition. We therefore summarize performance across four seasonal groups in the sampled 2025 test period: winter (December-January-February; DJF), pre-monsoon (March-April-May; MAM), monsoon (June-July-August-September; JJAS), and post-monsoon (October-November; ON), for S0, SC, and DF (mean ± std across 3 seeds).

Table 8 shows that the pooled-annual pattern of Section 4.1 holds within every season individually: conditioning (SC or DF) reduces T2 and PSFC error relative to the unconditioned S0 baseline in all four seasons, with the largest relative gain in MAM, where S0's T2 RMSE (2.17 K) is both higher and more variable across seeds than SC's or DF's (1.65-1.66 K). SC and DF remain close to each other throughout the annual cycle: neither configuration is uniformly ahead of the other in any season, and the differences between them (typically 0.01-0.03 K for T2, 0.01-0.03 hPa for PSFC) are within one cross-seed

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table 9. Continuous-variable RMSE (mean ± std across 3 seeds) across the full topology × conditioning design space. Every conditioned configuration (SC, SF, DC, DF, DF2) clearly outperforms its unconditioned counterpart (S0, D0); within the conditioned configurations, differences are small, and DC (dual encoder, concat conditioning, no FiLM) is consistently among the weakest.

Variable	S0	SC	SF	D0	DC	DF	DF2
PSFC RMSE (hPa)	1.55±0.08	1.38±0.04	1.40±0.07	1.60±0.02	1.41±0.02	1.37±0.04	**1.36±0.04**
T2 RMSE (K)	1.81±0.13	**1.49±0.02**	1.52±0.04	1.94±0.03	1.53±0.03	1.50±0.01	1.51±0.06
Q2 RMSE (g/kg)	1.56±0.01	1.55±0.02	**1.54±0.02**	1.58±0.01	1.56±0.01	1.56±0.01	**1.54±0.02**
U10 RMSE (m/s)	1.78±0.05	1.77±0.02	1.76±0.02	1.79±0.01	1.79±0.02	**1.75±0.01**	**1.75±0.01**
V10 RMSE (m/s)	1.75±0.02	**1.73±0.00**	**1.73±0.03**	1.76±0.00	1.75±0.00	**1.73±0.01**	1.74±0.02
Wind RMSVE (m/s)	2.50±0.04	2.48±0.01	2.47±0.03	2.51±0.01	2.50±0.01	**2.46±0.01**	2.47±0.02

Table 10. Precipitation skill (mean ± std across 3 seeds) across the full design space. CSI at all three thresholds and the masked-intensity errors are close across every configuration, and no single configuration dominates on all four metrics simultaneously.

Metric	S0	SC	SF	D0	DC	DF	DF2
CSI@0.1mm ↑	**0.75±0.01**	**0.75±0.01**	0.74±0.01	0.74±0.01	**0.75±0.01**	0.74±0.00	**0.75±0.01**
CSI@1mm ↑	0.69±0.01	**0.70±0.00**	0.69±0.02	**0.70±0.00**	**0.70±0.01**	0.69±0.02	**0.70±0.00**
CSI@10mm ↑	**0.54±0.02**	**0.54±0.01**	0.53±0.02	0.52±0.01	**0.54±0.00**	0.52±0.04	0.51±0.02
Masked MAE (mm) ↓	24.73±0.58	24.63±0.20	24.20±0.15	**24.07±0.31**	24.62±0.04	24.21±0.35	24.29±0.43
Masked RMSE (mm) ↓	**51.80±0.56**	52.15±0.30	52.73±0.25	52.14±0.59	51.75±0.66	52.96±0.62	52.05±0.93

standard deviation. Wind-vector error and precipitation CSI show the same seasonal pattern: JJAS (monsoon) gives the highest precipitation CSI for all three configurations ($\geq$0.79), consistent with the more spatially extensive and less intermittent rainfall of the monsoon season, while DJF gives the lowest CSI ($\leq$0.43) and the largest relative S0-vs-conditioned gap. This indicates that seasonal regime change does not disproportionately affect any one configuration, and the S0-vs-conditioned/SC-vs-DF pattern established for the pooled-annual comparison is not an artefact of a particular season.

4.3 Ablation Studies

This section tests the full topology × conditioning design space introduced in Section 3.5 (S0, SC, SF single-encoder; D0, DC, DF, DF2 dual-encoder) to separate the contributions of dual-encoder modality separation and FiLM conditioning. All seven configurations are trained for 3 seeds (42, 43, 44) on the same data pipeline described in Section 2.1. With only 3 seeds per configuration, individual per-seed values are not precise enough to report on their own. Each comparison is summarized by its mean, and the sign of the improvement in each seed is used only to check direction-consistency (the decision rule below), not to characterize seed-to-seed variability in detail.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Tables 9 and 10 show the same pattern already visible in the S0/SC/DF comparison of Section 4.1: any conditioning mechanism substantially improves on the unconditioned baselines (S0, D0), while the five conditioned configurations (SC, SF, DC, DF, DF2) cluster tightly, with no configuration dominating across all metrics. Comparing seven configurations across five variables and four domains at once, using a 7-column table, makes it easy to find some metric on which any given configuration looks best, without that ranking meaning much. To avoid this, we reduce the comparison to a single primary score: the RMSE for each of the five continuous variables (T2, U10, V10, Q2, PSFC), in each of the four domains, is divided by that variable's own standard deviation in the training data, so that variables with very different physical units and scales contribute on a comparable footing; these 20 normalized values are then averaged into one number per configuration (Table 11). Both this score and the rule for declaring one configuration better than another (below) were fixed before the comparison was run, so the definition of a "win" could not be chosen after seeing which one favoured a particular configuration. The decision rule itself is simple: a configuration is judged to beat another only if the sign of the per-seed improvement is consistent across all three paired seeds (42, 43, 44). With only three seeds, this direction-consistency check is a weak screen: under the null hypothesis of no true difference between two configurations, three independent paired comparisons agree in sign by chance alone 25% of the time, so a "win" under this rule should be read as a stable, easily reproduced direction of improvement. The four domains are also not independent samples of separate weather: d02-d04 are nested inside d01 and all four domains see the same large-scale weather at every timestamp, so the 20-value average behind the primary score reflects four spatially correlated views of the same synoptic history.

Table 11. Pre-registered primary decision score (lower is better; mean ± std across 3 seeds). All seven configurations fall within a 0.014-point band.

Configuration	Primary score
S0	0.2914 ± 0.0011
SC	0.2829 ± 0.0017
SF	0.2825 ± 0.0032
D0	0.2958 ± 0.0001
DC	0.2861 ± 0.0016
DF (FiLMeR v1.0)	0.2825 ± 0.0005
DF2	0.2822 ± 0.0033

Tables 11 and 12 give a clear picture of the separate contributions of dual-encoder modality separation and FiLM conditioning. Dual-encoder modality separation, on its own, does not help: DC (dual encoder, concat conditioning) loses to SC (single encoder, concat conditioning) in every seed, by a mean of 1.16% on the primary score. FiLM conditioning fixes this: DF beats DC in every seed, by a mean of 1.25%. However, DF does *not* consistently beat SC, the simplest single-encoder, concat-conditioned configuration: the direction of improvement is not the same across the 3 seeds, so this comparison is a tie despite a small positive mean (0.11%). The same tie pattern holds for DF2 vs. SC, DF vs. DF2, and SF vs. SC. In short:

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table 12. Pairwise decision-rule outcomes on the primary score: percent improvement of the first-named configuration (A) over the second (B) in each of the 3 seeds, and their mean. A comparison is called a win for A only when the sign of the improvement is the same in all three seeds (the pre-registered decision rule); otherwise it is a tie, regardless of the mean. This is a weak, three-sample screen, not a significance test; see the discussion in the text.

Comparison (A vs. B)	Seed 42	Seed 43	Seed 44	Mean	Outcome
Dual encoder vs. single (DC vs. SC)	−1.43%	−0.82%	−1.22%	−1.16%	DC loses to SC
FiLM vs. concat, dual encoder (DF vs. DC)	+0.88%	+0.77%	+2.11%	+1.25%	DF beats DC
Gated FiLM vs. concat, dual encoder (DF2 vs. DC)	+2.15%	−0.15%	+2.15%	+1.38%	Tie
FiLM vs. concat, single encoder (SF vs. SC)	−0.86%	−0.73%	+1.98%	+0.13%	Tie
Proposed model vs. simplest baseline (DF vs. SC)	−0.54%	−0.05%	+0.91%	+0.11%	Tie
Gated FiLM vs. simplest baseline (DF2 vs. SC)	+0.75%	−0.97%	+0.95%	+0.24%	Tie
Decoder-only vs. gated-fusion FiLM (DF vs. DF2)	−1.30%	+0.91%	−0.04%	−0.14%	Tie

FiLM conditioning is necessary to make the dual-encoder topology competitive with a simple single-encoder, concat-conditioned configuration, but neither the dual-encoder topology nor FiLM conditioning individually outperforms that simpler configuration.

4.3.1 What Is FiLM Actually Conditioning On?

FiLM is necessary within the dual-encoder topology (DF beats DC), but it does not clearly outperform the simplest configuration (DF vs. SC is a tie). This makes it useful to ask what FiLM's contribution actually consists of, not just whether it helps. FiLM's context vector (Section 2.1) has 5 values: a single log grid-spacing term, plus cyclic hour-of-day and day-of-year encodings. Grid spacing only takes two values across the whole dataset (27 km for d01, 9 km for d02-d04), so this vector cannot represent a 4-way per-domain lookup table even in principle. We tested this directly with a context counterfactual on the trained DF (seed 42) checkpoint. At inference time, we replaced the true context vector with four alternatives: the correct context; a time-shuffled context (hour/day-of-year values randomly permuted across the batch, grid spacing kept correct); a spacing-swapped context (grid spacing toggled between the 27 km and 9 km values, time kept correct); and a fully zeroed context.

Table 13 shows a clear asymmetry. Swapping the grid-spacing component leaves T2 and PSFC RMSE essentially unchanged from the correct context (1.50 vs. 1.49 K; 1.42 vs. 1.41 hPa). Shuffling only the temporal component degrades T2 RMSE to 3.97 K, worse than zeroing the context vector entirely (3.21 K), while PSFC RMSE degrades to 2.36 hPa, somewhat better than zeroing (2.56 hPa). A wrong time signal is therefore at least as damaging as no time signal at all, and for T2 specifically more damaging, consistent with the model having learned to rely on the temporal component closely enough that misleading it costs more than removing it outright. FiLM's measurable contribution in this model is therefore mainly *temporal* conditioning (making use of hour-of-day and day-of-year information), not *domain* conditioning, even though the model is trained across

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table 13. Context counterfactual on the trained DF (seed 42) checkpoint, full 2025 test set. Shuffling the temporal component is at least as damaging as removing the context entirely (more damaging for T2, similar for PSFC), while swapping the grid-spacing component leaves performance essentially unchanged from the correct context.

Context condition	T2 RMSE (K)	PSFC RMSE (hPa)	Wind RMSVE (m/s)
Correct (as trained)	**1.49**	**1.41**	**2.47**
Spacing swapped	1.50	1.42	2.47
Time shuffled	3.97	2.36	2.77
Context zeroed	3.21	2.56	2.69

four geographically distinct domains. This also explains the design-space result above: a mechanism that mainly encodes time of day and season is a plausible reason why FiLM narrows the gap opened by DC's dual-encoder fusion (Table 12) without adding a further advantage over SC, which already receives the same temporal information through simple concatenation.

For context, all seven configurations clearly outperform two simple reference points evaluated on the same DF checkpoint and test set: an aligned-GFS persistence baseline (predicting no correction to the input GFS state; T2 RMSE 3.41 K, PSFC RMSE 5.02 hPa) and a domain-climatology baseline (predicting each domain's training-period mean field; T2 RMSE 7.97 K, PSFC RMSE 77.53 hPa). Every configuration in Table 9 is at least 2.2 times better than persistence on T2 and 3.5 times better on PSFC, confirming that the model is learning a real regional correction. The open question is which of its two design choices, dual-encoder modality separation and FiLM conditioning, is responsible, and the evidence above credits conditioning in general, and FiLM's temporal component specifically, not modality separation.

5 Discussion

Comparing FiLMeR's dual-encoder, FiLM-conditioned design against matched single-encoder baselines within a controlled topology $\times$ conditioning design space (Section 4.3) supports a specific, testable reading of those results: conditioning of some kind is necessary, dual-encoder modality separation is only useful when paired with FiLM, and FiLM's own advantage over the simplest concat-conditioned baseline is not established under the decision rule used here. A plausible architectural reason for the dual-encoder result is available and worth stating explicitly. The per-level fusion adapter that combines the two encoders' feature maps (Eq. 4) is a 1×1 convolution, so it can only mix the dynamic and static modalities channel-wise at each spatial location; the single encoder, in contrast, mixes them within a full 3×3 receptive field from its very first convolutional layer. Without FiLM, the dual-encoder topology therefore has strictly less capacity to let one modality's local spatial context inform the other at fusion time, consistent with DC (dual encoder, concat conditioning) losing to SC (single encoder, concat conditioning) in every seed (Table 12). FiLM's decoder-level affine modulation is not itself a spatial-mixing operation, but it adds a channel-wise degree of freedom, conditioned on grid spacing and time, that can partially compensate for the fusion adapter's restricted receptive field, consistent with DF recovering DC's deficit (DF beats DC in every seed) without, on its own,

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

exceeding what the single encoder already achieves (DF does not consistently beat SC; the same tie pattern holds for DF2 vs. SC, DF vs. DF2, and SF vs. SC). On the evidence available here, SC, SF, DF, and DF2 are a statistical tie, and DC is the only configuration that can be said to lose outright.

This also reframes how FiLM's mechanism should be interpreted. The context counterfactual in Section 4.3 shows that FiLM's measurable contribution in the trained DF model is predominantly temporal: shuffling the hour-of-day/day-of-year components of the context vector degrades T2 error more than removing the context vector entirely (and PSFC error nearly as much), while swapping the grid-spacing component (the only channel through which domain identity could influence FiLM; Section 2.1) leaves performance essentially unchanged. FiLM in this model is functioning primarily as a diurnal- and seasonal-cycle conditioning mechanism. This is consistent with why it repairs DC's dual-encoder fusion deficit (DC and DF receive the same context vector; only DF can use it to modulate decoder features conditionally on time) without providing a further advantage over SC, which already receives the same temporal information through simple concatenation. We regard this as a genuine, useful finding about how FiLM operates in a shared-weight, multi-domain emulator, even though it does not support the stronger claim that a dual-encoder, FiLM-conditioned architecture is the best available choice for this task.

This finding carries an important scope limitation. The training corpus spans four WRF domains at only two distinct grid spacings (27 km and 9 km; Table 2), and all four domains are seen during training (Section 2.1). This finding is scoped to the two resolutions and four fixed geographic extents evaluated here; a larger or more geographically diverse domain set may reveal a domain-adaptive FiLM benefit that this corpus cannot expose. The tie between SC and the dual-encoder/FiLM configurations observed here is a property of this specific multi-domain corpus, not a general statement about dual-encoder or FiLM architectures for regional weather emulation.

Precipitation shows a different pattern from the continuous variables. No configuration in Table 10 dominates across categorical thresholds and the masked-intensity regression metric simultaneously: S0 gives the highest light- and moderate-rain CSI, DC gives the highest moderate-rain CSI among the conditioned configurations, and DF gives the lowest masked MAE but not the lowest masked RMSE. As with the continuous variables, CSI and frequency bias should be interpreted jointly, since CSI conflates spatial coverage bias with detection skill and a configuration with a frequency bias above unity can obtain a higher CSI at similar probability of detection without better event placement (Schaefer, 1990; Roberts and Lean, 2008). Unlike the continuous fields, where conditioning gives a clear, seed-consistent improvement over the unconditioned baseline, precipitation skill in this study shows no comparably clear benefit from either dual-encoder separation or FiLM conditioning; this may reflect genuine model behaviour, or it may reflect that precipitation's larger event-to-event variability requires more than three seeds to resolve small ordering differences with the confidence achieved for the smoother continuous fields.

The spatial diagnostics (Figure 4) provide additional context for interpreting the domain-mean scores above. Time-averaged maps show that the model retains several geographically organized patterns in temperature, wind speed, and precipitation, while also revealing localized residual errors that are not visible in the aggregate metrics. This distinction matters for regional modelling applications, where a low domain-mean error may still be accompanied by systematic errors near coastlines, steep topography, or regions of intermittent precipitation. The present evaluation therefore frames the model as an efficient emulator of selected WRF-scale surface fields.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Several limitations follow from this framing. The model is trained to reproduce WRF-derived targets and therefore inherits uncertainties and biases from that reference workflow; independent evaluations of WRF-ARW over the Indian subcontinent and Himalayan terrain (Navale and Singh, 2020; Hassan et al., 2026), though not of this exact physics configuration, give an external point of reference for the kinds of systematic biases (e.g. topographic rainfall sensitivity) such a reference simulation can carry. It does not enforce mass, energy, or moisture conservation, and the current evaluation does not quantify error growth under long autoregressive rollouts. Extreme precipitation and convective displacement errors require further analysis, particularly for rare events where deterministic grid-point metrics can be sensitive to small spatial offsets. As noted above, the four-domain, two-resolution training corpus limits how confidently the FiLM mechanism findings generalize to a larger or more diverse domain set; testing on a substantially larger and more geographically varied set of domains is a natural next step for distinguishing a genuine domain-conditioning benefit from the temporal-conditioning benefit established here. Future model development can examine conservation-aware losses, probabilistic output formulations for precipitation, and evaluation on a substantially larger and more diverse set of training domains, all under the same constraints of reproducibility and computational cost considered here.

6 Conclusions

This study presents FiLMeR (v1.0), a context-conditioned neural emulator for WRF-scale near-surface regional weather fields over the Indian subcontinent, and a controlled evaluation of the two design choices motivating it: dual-encoder modality separation and FiLM conditioning. The model uses current and previous GFS states, static WRF geogrid information, and a compact spatiotemporal context vector to predict six surface variables without requiring future lateral-boundary fields at inference time. FiLMeR is evaluated as an emulator of WRF-derived surface states. The main findings from the sampled 2025 evaluation are as follows:

- **Conditioning, not architecture, drives the largest gain.** Every conditioned configuration (concat or FiLM, single- or dual-encoder) substantially outperforms its unconditioned counterpart on the continuous thermodynamic and wind variables, and this pattern persists across all four sampled seasons (DJF, MAM, JJAS, ON).
- **Dual-encoder modality separation is only useful when paired with FiLM.** Under a pre-registered, seed-consistent decision rule, the dual-encoder, concat-conditioned configuration (DC) consistently loses to the simpler single-encoder, concat-conditioned baseline (SC); FiLM consistently repairs this deficit (DF beats DC in every seed), but does not consistently outperform SC itself. SC, SF, DF, and DF2 are a statistical tie on the pre-registered primary score; DC is the only configuration that loses outright.
- **FiLM's contribution is predominantly temporal, not domain-adaptive.** A context counterfactual on the trained DF checkpoint shows that shuffling the temporal component of the context vector is at least as damaging to continuous-variable error as removing the context vector entirely (more damaging for T2), while swapping the grid-spacing compo-

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

nent leaves performance essentially unchanged. FiLM in this model functions mainly as a diurnal- and seasonal-cycle conditioning mechanism, a finding scoped to the four-domain, two-resolution training corpus used here.

- **Precipitation shows no consistent winner.** Categorical skill and masked wet-cell intensity errors are close across all seven configurations, with no single configuration dominating simultaneously on FAR, CSI, and intensity error; CSI and frequency bias should be interpreted jointly.
- **Spatial structure and limitations:** Spatial diagnostics show that the model retains several geographically organized patterns in temperature, wind speed, and precipitation, while localized residual errors remain. The model does not enforce conservation of mass, energy, or moisture, and additional work is needed to assess long autoregressive rollouts, rare extremes, displacement errors in convective precipitation, and behaviour on a larger and more geographically diverse set of training domains than the four evaluated here.

Taken together, these results support a clear, specific claim: conditioning of some form is necessary for accurate multi-domain, shared-weight emulation. Dual-encoder modality separation needs FiLM to be competitive with a simpler concat-conditioned single encoder. FiLM's demonstrated value in this study is a temporal-conditioning mechanism, not a domain-adaptive one. FiLMeR (v1.0, configuration DF) remains a reasonable default choice: it is statistically indistinguishable from SC in accuracy while using about 4% fewer parameters (36.85M vs. 38.46M; Appendix Table D1), and its temporal-conditioning mechanism is the more clearly characterized of the two. It is not, however, uniquely superior to the simpler SC configuration on accuracy alone. Future work should test a larger, more geographically diverse set of training domains to check whether a genuine domain-conditioning benefit appears at larger scale, examine conservation-aware training objectives, explore probabilistic or ensemble formulations for precipitation, and evaluate multi-step forecast degradation.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Appendix A: Data Access

The dynamic atmospheric forcing data used in this study are obtained from the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) operational forecast archive. These data are distributed through the National Center for Atmospheric Research (NCAR) Geoscience Data Exchange (GDEX) under the archive identifier `d084001`, the same underlying archive registered with the NSF NCAR Research Data Archive as dataset ds084.1 (see the code and data availability statement below). The regional reference fields were generated with the Weather Research and Forecasting (WRF) model, version 4.7.1, using the nested-domain configuration described in the main text. Static geographic inputs were prepared with the WRF Preprocessing System (WPS), version 4.6.0, using the domain-specific `geo_em` files to extract terrain, soil, and land-use information. The derived paired GFS-WRF spatial tensors, optimized model weights for FiLMeR (v1.0), and corresponding data preprocessing pipelines constitute core research products of this study. The trained model weights, configuration scripts, and a representative 3-month sample of the paired GFS-WRF tensors are permanently archived on Zenodo (see the code and data availability statement below). The full multi-year, four-domain paired corpus underlying training and evaluation is not itself archived in repository form: the raw WRF-ARW output (full three-dimensional fields across all vertical levels and all four domains, at 3-hourly output over roughly two years of GFS-driven 10-day cycles) is substantially larger than is practical to host in a general-purpose repository. It can be regenerated in full from the publicly archived GFS $0.25°$ forecast fields (dataset ds084.1, described above) combined with the WRF-ARW namelist and physics configuration documented in Appendix A1 and the domain-specific `geo_em` files used for the WPS preprocessing step.

A1 Reference Simulation Configuration (WRF-ARW)

Regional reference fields for training and validation were generated with the Advanced Research Weather Research and Forecasting (WRF-ARW) model. The domain configuration includes one parent domain (d01) at 27 km horizontal grid spacing with 130×130 grid points and three nested domains (d02, d03, and d04) at 9 km horizontal grid spacing with 100×100 grid points. The vertical grid uses 48 terrain-following eta levels and a model top of 50 hPa. The integration time step is 90 s. All four domains are integrated concurrently within a single two-way nested run (feedback enabled). Each 10-day cycle begins from the GFS initialization time itself, with no spin-up period discarded before the first archived output; this is consistent with the near-zero accumulated precipitation recorded at the first target timestamp of every cycle.

The WRF physics configuration used for the reference simulations is summarized below:

- **Microphysics:** Thompson aerosol-aware microphysics scheme (`mp_physics = 8`).
- **Cumulus Parameterization:** Kain-Fritsch mass-flux scheme (`cu_physics = 1`).
- **Radiation Options:** Rapid Radiative Transfer Model (RRTM) scheme for longwave radiation (`ra_lw_physics = 1`) and the Dudhia scheme for shortwave radiation (`ra_sw_physics = 1`).
- **Planetary Boundary Layer (PBL):** Yonsei University (YSU) non-local vertical diffusion scheme (`bl_pbl_physics = 1`), coupled with the MM5 Monin-Obukhov surface layer scheme (`sf_sfclay_physics = 1`).

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

– **Land Surface Model:** Noah Multiparameterization (Noah-MP) land surface model (`sf_surface_physics = 4`) running with 4 prognostic soil layers.

The 9 km nested domains (d02-d04) fall within the 3-10 km "grey zone" in which cumulus convection is only partially resolved by the model grid, and a scale-aware cumulus scheme (e.g., Multi-scale Kain-Fritsch or Grell) is generally preferable to a conventional one in this range. We retained the standard Kain-Fritsch scheme at 9 km. This choice is defensible in the present context because the goal of this study is to demonstrate that a deep-learning model can emulate a given, fixed WRF-ARW configuration, not to identify the single most physically appropriate parameterization set for operational forecasting; the 9 km domains sit at the coarse edge of the 3-10 km grey zone, and given the geographic diversity of the Indian subcontinent, no single parameterization set (scale-aware or otherwise) is likely to be uniformly optimal across all four domains regardless of this choice. The contribution being tested here is therefore the fidelity of the emulator to its WRF reference, not the physical optimality of that reference. Once emulation fidelity is established, the same approach can be retrained on WRF data generated with a more regionally appropriate physics configuration, including a scale-aware cumulus scheme, for forecasting or hindcasting applications.

The paired dataset combines GFS atmospheric inputs with the corresponding WRF reference targets. The 2023 and 2024 annual cycles are used for model development, and 2025 is held out for independent testing; the evaluation reported here uses the first 10 days of each month from that held-out year.

Appendix B: Dataset Preprocessing

The preprocessing pipeline maps GFS forcing, WRF static geographic fields, and WRF target variables to a common tensor format. It preserves the one-step-ahead temporal ordering of each sample, applies normalization based on the training period, and appends domain and time metadata used by the conditioning module. Dynamic predictors are extracted from 0.25° GFS fields. The selected variables include near-surface, cloud, precipitation, and pressure-level thermodynamic and wind fields (Table B1). For each sample, fields from two consecutive 3-hourly input times are stacked channel-wise before being passed to the dynamic encoder. Samples are represented as ordered triplets,

$$\{\mathbf{X}_{t-3\mathrm{h}}, \mathbf{X}_t, \mathbf{Y}_{t+3\mathrm{h}}\}, \tag{B1}$$

where $\mathbf{X}$ is the GFS input state and $\mathbf{Y}$ is the WRF target state. Timestamps are matched before sample construction, and incomplete triplets are excluded. The WRF targets are 2 m temperature (T2), 10 m zonal wind (U10), 10 m meridional wind (V10), accumulated precipitation from RAINC+RAINNC, 2 m water-vapor mixing ratio (Q2), and surface pressure (PSFC).

Static and monthly lower-boundary predictors are extracted from WPS `geo_em` files (Table B2). Monthly fields are selected using the sample's calendar month, and the processed static tensors are reused by domain and month during training and inference.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table B1. GFS dynamic atmospheric variables

Category	Variable (Short Name)	Description	Unit
Surface / Near-Surface	`2t`	2 m air temperature	K
	`2sh`	2 m Water Vapour Mixing Ratio	kg kg^{-1}
	`sp`	Surface pressure	Pa
	`10u`	10 m zonal wind component	m s^{-1}
	`10v`	10 m meridional wind component	m s^{-1}
	`tcc`	Total cloud cover	%
	`prate`	Precipitation rate	kg m^{-2} s^{-1}
Upper-Air (Pressure Levels)	`t`	Temperature (1000, 850, 700, 500 hPa)	K
	`u`	Zonal wind component (850, 700, 500 hPa)	m s^{-1}
	`v`	Meridional wind component (850, 700, 500 hPa)	m s^{-1}
	`q`	Specific humidity (850, 700, 500 hPa)	kg kg^{-1}

Table B2. Static predictors used by the static encoder: 30 physiographic channels extracted from the WPS `geo_em` files (monthly fields indexed by calendar month) plus two per-pixel absolute-position channels appended during preprocessing, giving $C_s = 32$ channels in total.

Variable	Description	Treatment
`HGT_M`	Terrain height	Continuous
`LANDMASK`	Land water mask	Binary
`SOILTEMP`	Deep soil temperature	Continuous
`SNOALB`	Maximum snow albedo	Continuous
`CON` / `VAR`	Soil and terrain descriptors	Continuous
`ALBEDO12M`	Monthly surface albedo	Monthly
`LAI12M`	Monthly leaf area index	Monthly
`GREENFRAC`	Monthly green vegetation fraction	Monthly
`LANDUSEF`	Land-use fractional categories	Categorical/fractional
`coord_lat`	Absolute latitude at each grid cell	Continuous, appended
`coord_lon`	Absolute longitude at each grid cell	Continuous, appended

Each sample is assigned a 5-dimensional metadata vector, $\mathbf{p}^{(d,t)} \in \mathbb{R}^5$, composed of the components listed in Table B3. Geographic bounding coordinates, domain center, spatial extent, and nested-domain position are omitted from $\mathbf{p}^{(d,t)}$ (Section 2.1); absolute position is carried instead by the two per-pixel coordinate channels in the static tensor (Table B2).

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table B3. Spatiotemporal metadata components used for FiLM conditioning. Hour-of-day and day-of-year use cyclic sine/cosine encodings, avoiding a discontinuity at day- and year-end boundaries.

Component	Description	Encoding / treatment
`log_spacing`	$\log(\Delta x_d/9\text{km})$, grid-spacing ratio	Spatial resolution
`hour_sin`	$\sin(2\pi \cdot \text{hour}/24)$	Cyclic temporal
`hour_cos`	$\cos(2\pi \cdot \text{hour}/24)$	Cyclic temporal
`doy_sin`	$\sin(2\pi \cdot \text{day\_of\_year}/365)$	Cyclic temporal
`doy_cos`	$\cos(2\pi \cdot \text{day\_of\_year}/365)$	Cyclic temporal

All normalization statistics are computed from the 2023-2024 training period and then fixed for validation and 2025 testing. Dynamic GFS channels and WRF target channels use channel-wise z-score normalization,

$$\hat{x}_{c,i,j} = \frac{x_{c,i,j} - \mu_c}{\sigma_c + \epsilon}, \tag{B2}$$

where μ_c and σ_c are the training-set mean and standard deviation for channel c, and $\epsilon = 10^{-8}$. Continuous static predictors are standardized, while binary masks, categorical fractions, bounded monthly fields, and metadata are retained in a bounded range or min-max scaled to $[0, 1]$. Model outputs are transformed back to physical units before evaluation.

Appendix C: Training Hyperparameters Details

Table C1. Hyperparameter settings and optimization schedules.

Configuration Parameter	Value / Operational Policy
Hardware Baseline	Single NVIDIA A100 GPU
Optimization Algorithm	AdamW
Base Learning Rate (η)	3×10^{-4}
Weight Decay (λ)	10^{-4} (L_2 regularization; 5×10^{-4} on precipitation-head parameters)
Total Training Budget	Maximum 50 epochs
Linear Warm-up Phase	First 5 epochs (scales η from 0 to 3×10^{-4})
Post-Warmup Schedule	Cosine decay from 3×10^{-4} to a minimum of 10^{-5}
Early Stopping Threshold	10 epochs without validation-loss improvement

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Appendix D: Design-Space Configuration Details

This appendix gives implementation details for the seven topology × conditioning configurations introduced in Section 2.2 (S0, SC, SF, D0, DC, DF, DF2), shared by the main-text baseline comparison and the ablation study (Section 4.3).

D1 Single-Encoder Topology (S0, SC, SF)

The single-encoder configurations concatenate the dynamic and static tensors along the channel dimension before encoding, giving a $2C_g + C_s = 72$-channel input to one 4-level encoder with widths $64 \to 128 \to 256 \to 512$; each level applies two 3×3 convolutions with group normalization and SiLU activation, followed by 2×2 max-pooling.

D2 Dual-Encoder Topology (D0, DC, DF, DF2)

The dual-encoder configurations, including **DF** (FiLMeR v1.0), use two independent 4-level encoders of the same form (one over the $2C_g = 40$ dynamic channels, one over the $C_s = 32$ static channels), each with widths $32 \to 64 \to 128 \to 256$, fused at every level to the shared $64, 128, 256, 512$ skip widths used by the decoder (Eq. 4).

D3 Conditioning Mechanisms (none, concat, film)

For **none** and **concat**, every decoder skip connection passes through the same 1×1 convolutional adapter; for **none** the adapter receives a zeroed context vector, and for **concat** it receives the true $\mathbf{p}^{(d,t)}$, so the two are matched in adapter capacity and differ only in the information available to it. For **film**, no skip-connection adapter is used; instead $\mathbf{p}^{(d,t)}$ is mapped by the context MLP ϕ_p (Section 2.2) to $\mathbf{e}_p \in \mathbb{R}^{128}$, and each decoder level applies its own affine transform (Eq. 8) to the post-fusion decoder feature map. **DF2** additionally supplies $\mathbf{e}_p$ to a gated variant of the per-level fusion adapter, so FiLM also controls how much of each modality's skip feature is retained at fusion time.

D4 Parameter Counts

Table D1 lists the trainable parameter count for each configuration. The 'none' and 'concat' conditioning adapters are architecturally identical, so S0/SC and D0/DC share exactly the same parameter count within a topology; FiLM adds a small number of additional per-level affine parameters (SF, DF), and the gated fusion adapters in DF2 add a further small increment.

Appendix E: Domainwise Spatial Plots and Extended Metrics

Detailed spatial error maps and comprehensive seasonal metrics broken down individually for all four target domains (**d01**, **d02**, **d03**, and **d04**) across all model variants and ablation studies are provided in the accompanying Supplementary Material.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Table D1. Trainable parameter counts for the seven topology × conditioning configurations. The single-encoder configurations (S0, SC, SF) are consistently larger than the dual-encoder configurations (D0, DC, DF, DF2) because one 4-level encoder at widths 64-512 has more parameters than two 4-level encoders at widths 32-256 plus four small fusion adapters, despite both processing the same total number of input channels.

Configuration	Topology	Conditioning	Parameters
S0	Single	None	38.46M
SC	Single	Concat	38.46M
SF	Single	FiLM	38.86M
D0	Dual	None	36.45M
DC	Dual	Concat	36.45M
DF (FiLMeR v1.0)	Dual	FiLM	36.85M
DF2	Dual	FiLM + gated fusion	37.10M

GenAI Disclosure

During the preparation of this manuscript, the authors utilized generative artificial intelligence (AI) tools solely for the purposes of text refinement and grammatical correction. All scientific concepts, model architecture design choices, data processing workflows, interpretations of results, and final conclusions were conceived and executed entirely by the human authors, who maintain full accountability for the integrity, originality, and accuracy of the presented work.

Acknowledgments

The authors acknowledge the open-access data and modelling resources used in this study. The Global Forecast System (GFS) data provided by the National Oceanic and Atmospheric Administration (NOAA) were used for model initialization and large-scale atmospheric forcing. The Weather Research and Forecasting (WRF) model, developed and maintained by the National Center for Atmospheric Research (NCAR) and the broader WRF community, was used to generate the high-resolution reference simulations.

Code and data availability. The exact version of FiLMeR (v1.0) used to produce the results in this paper, along with the trained model weights, configuration scripts, and a 3-month sample dataset of input global atmospheric fields and regional targets for testing and replication, is permanently archived on Zenodo under https://doi.org/10.5281/zenodo.21962186 (Prasad et al., 2026). The complete, large-scale historical Global Forecast System (GFS) 0.25° data used for full model training are openly available from the NSF NCAR Research Data Archive (dataset ds084.1) under https://doi.org/10.5065/D65D8PWK (National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce, 2015).

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Author contributions. BP designed the study, implemented the experiments, processed the GFS and WRF datasets, performed the evaluation, and prepared the initial manuscript draft. DL contributed to data processing, development of the FiLMeR model, experiment organization, result analysis, and manuscript revision. MR contributed to the WRF simulation configuration, meteorological interpretation, and evaluation of regional weather behaviour. UB contributed to the study design, interpretation of hydrometeorological results, and manuscript revision. NB supervised the machine-learning methodology, contributed to the model design and evaluation strategy, and revised the manuscript. All authors reviewed and approved the final manuscript.

Competing interests. The authors declare that they have no competing interests.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

References

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/10.1038/nature14956, 2015.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Chen, H., Han, T., Zhang, J., Guo, S., and Bai, L.: STCast: Adaptive Boundary Alignment for Global and Regional Weather Forecasting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), https://arxiv.org/abs/2509.25210, 2026.

Chen, K., Han, T., Ling, F., et al.: The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time, Communications Earth & Environment, 6, 518, https://doi.org/10.1038/s43247-025-02502-y, 2025.

Hassan, M., Abbas, S., Bilal, A., Chishtie, F. A., Ijaz, U. Z., Shi, X., Iqbal, W., Mahmood, T., Mahmood, R., and Fatima, I.: Advancing convection-permitting regional climate modeling for monsoon extremes in data-scarce, topographically complex regions of South Asia, Atmospheric Research, 329, 108 486, https://doi.org/10.1016/j.atmosres.2025.108486, 2026.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: GraphCast: Learning skillful medium-range global weather forecasting, Science, https://arxiv.org/abs/2212.12794, 2023.

Lee, K. et al.: A conditional U-Net surrogate model for annual average surface PM2.5 concentrations over South Korea, Journal of Air Quality Management, 2024.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Fourier Neural Operator for Parametric Partial Differential Equations, arXiv preprint arXiv:2010.08895, https://doi.org/10.48550/arXiv.2010.08895, 2020.

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Kashinath, K., Kautz, J., and Pritchard, M.: Residual corrective diffusion modeling for km-scale atmospheric downscaling, arXiv preprint arXiv:2309.15214, https://doi.org/10.48550/arXiv.2309.15214, 2023.

National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce: NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive, https://doi.org/10.5065/D65D8PWK, 2015.

Navale, A. and Singh, C.: Topographic sensitivity of WRF-simulated rainfall patterns over the North West Himalayan region, Atmospheric Research, 242, 105 003, https://doi.org/10.1016/j.atmosres.2020.105003, 2020.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999, https://doi.org/10.48550/arXiv.1804.03999, 2018.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2337–2346, https://doi.org/10.48550/arXiv.1903.07291, 2019.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al.: Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, https://doi.org/10.48550/arXiv.2202.11214, preprint, https://arxiv.org/abs/2202.11214, 2022.

https://doi.org/10.5194/egusphere-2026-4377
Preprint. Discussion started: 28 September 2026

Pathak, J., Cohen, Y., Garg, P., Harrington, P., Brenowitz, N., Durran, D., Mardani, M., Vahdat, A., Xu, S., Kashinath, K., and Pritchard, M.: Kilometer-scale convection-allowing model emulation using generative diffusion modeling, Science Advances, 12, eadv0423, https://doi.org/10.1126/sciadv.adv0423, 2026.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A.: Film: Visual reasoning with a general conditioning layer, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, https://doi.org/10.48550/arXiv.1709.07871, 2018.

Prasad, B., Lad, D., Rafiuddin, M., Bhatia, U., and Batra, N.: FiLMeR (v1.0): A FiLM-Conditioned Machine-Learning Emulator for Multi-Domain Regional WRF-Scale Surface Weather Fields, https://doi.org/10.5281/zenodo.21962186, 2026.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, Nature, 637, 84–90, https://doi.org/10.1038/s41586-024-08252-9, 2025.

Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, Monthly Weather Review, 146, 3885–3900, https://doi.org/10.1175/mwr-d-18-0187.1, 2018.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: a benchmark data set for data-driven weather forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, https://doi.org/10.1029/2020MS002203, 2020.

Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Monthly Weather Review, 136, 78–97, https://doi.org/10.1175/2007MWR2123.1, 2008.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241, Springer, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Salman, A. et al.: Emulating CMAQ using deep learning: A comparative study on simulating surface air quality over the CONUS, Atmospheric Environment, 2024.

Schaefer, J. T.: The critical success index as an indicator of warning skill, Weather and Forecasting, 5, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2, 1990.

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Duda, M., Huang, X.-Y., Wang, W., and Powers, J.: A Description of the Advanced Research WRF Version 3, Tech. rep., National Center for Atmospheric Research, https://doi.org/10.13140/RG.2.1.2310.6645, 2008.

Tian, Y., Ji, Y., Gao, X., Yuan, X., and Zhi, X.: Post-processing of short-term quantitative precipitation forecast with the multi-stream convolutional neural network, Atmospheric Research, 309, 107 584, https://doi.org/10.1016/j.atmosres.2024.107584, 2024.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing, 13, 600–612, https://doi.org/10.1109/TIP.2003.819861, 2004.

Xu, P., Zheng, X., Gao, T., et al.: An artificial intelligence-based limited area model for forecasting of surface meteorological variables, Communications Earth & Environment, 6, 372, https://doi.org/10.1038/s43247-025-02347-5, 2025.