Physics-Guided Multi-Task Learning for Subgrid Scale Turbulence Parameterization: A Comparative Study of Physics Integration Strategies

Sambit Kumar Panda*, Todd R. Jones*, Muhammad Shahzad*, Anna-Louise Ellis†, Bryan N. Lawrence‡§
*Department of Computer Science, University of Reading, Reading, United Kingdom
Email: s.panda@pgr.reading.ac.uk
†Met Office, Exeter, United Kingdom
‡Department of Meteorology, University of Reading, Reading, United Kingdom
§National Centre for Atmospheric Science, United Kingdom

***Abstract*—Neural network emulation of subgrid-scale (SGS) turbulence parameterizations in atmospheric models offers the promise of computational acceleration but suffers from brittle cross-regime generalization that limits operational deployment. We present a systematic evaluation of physics-guided multi-task learning strategies for SGS coefficient prediction, comparing six conditioning approaches across three neural architectures (MLP, ResMLP, TabTransformer) on contrasting atmospheric regimes: idealized tropical deep convection (Radiative Convective Equilibrium) and mid-latitude shallow convection (Atmospheric Radiation Measurement). Through evaluation of 222.6 million predictions (inference), we reveal dataset-dependent behaviour with important implications for neural network design. Simple MLPs with explicit Richardson number conditioning achieve improved cross-regime generalization on ARM data ($R^2 = 0.67$ viscosity, 0.73 diffusivity), outperforming architecturally complex alternatives by 158% in viscosity R^2 (0.432 vs 0.167) on dynamically important active turbulence regions. Residual networks that achieve strong performance on idealized RCE simulations ($R^2 > 0.78$) produce predictions indicative of numerical instability on the ARM regime, producing negative R^2 values with variance ratios exceeding $3\times$ the ground truth. Kling-Gupta Efficiency decomposition reveals systematic positive bias across all model configurations, reflecting training data sparsity in active turbulence regions rather than fundamental physics incompatibility. These findings establish design principles for atmospheric ML parameterization development: architectural simplicity with targeted physics conditioning provides improved offline predictive robustness compared to complex architectures relying on implicit pattern learning. Online coupling experiments remain essential to validate operational stability under feedback dynamics.**

***Index Terms*—Physics-guided neural networks, multi-task learning, atmospheric turbulence parameterization, cross-regime generalization, large eddy simulation, Scientific Machine Learning**

I. Introduction

Subgrid-scale (SGS) turbulence parameterization remains a fundamental computational bottleneck in cloud-resolving atmospheric simulations. Large Eddy Simulation (LES) explicitly resolves energy-containing eddies while parameterizing unresolved subgrid transport through closure schemes [1]. The fidelity of these parameterizations directly controls simulated cloud microphysics, boundary layer dynamics, and precipitation, which are processes that dominate uncertainty in numerical weather prediction and climate projections [2].

Modern LES frameworks such as the UK Met Office NERC Cloud Model (MONC) operate on domains containing hundreds of millions of grid points [3]. MONC implements a dynamic Smagorinsky-Lilly closure [4], [5] that adapts turbulent mixing coefficients to local atmospheric stability, providing improved physical realism compared to constant-coefficient approaches. This adaptivity comes at computational cost: stability function evaluations and coefficient updates impose significant overhead relative to static schemes, motivating interest in efficient emulation strategies.

Machine learning offers potential acceleration pathways, with demonstrated speedups from modest factors in coupled configurations [6] to 10–80$\times$ in specialized offline applications [7]. However, translating these gains to operational modeling requires addressing fundamental challenges in generalization and physical consistency; as the final realized speedup depends strongly on implementation details and requires validation through online coupling.

Data-driven turbulence emulators exhibit characteristic failure modes beyond their training distributions: feedback-driven error amplification in online coupling [8], spurious correlations that collapse under distribution shift [9], and systematically smoothed predictions that miss extreme events [10]. Recent benchmarks show neural emulators trained on one regime often fail on different turbulence characteristics [11]. Whether failures stem from physics incompatibility or data coverage has direct remediation implications.

Physics-informed machine learning (PIML) addresses these limitations by embedding conservation laws and domain knowledge into neural architectures [12], constraining the hypothesis space to physically consistent solutions.

For atmospheric SGS modeling, the Richardson number (Ri) represents a particularly important physical quantity governing the relationship between atmospheric stability and turbulent mixing intensity. Intuitively, Ri determines whether turbulence is enhanced by buoyancy (unstable conditions, $Ri < 0$) or suppressed by stable stratification ($Ri > 0$), directly controlling the magnitude of eddy viscosity and diffusivity coefficients. MONC's operational parameterization explicitly computes stability functions $f_m(Ri)$ and $f_h(Ri)$ that modulate these coefficients based on Richardson number values. This known physical dependency motivates investigating whether encoding Richardson-coefficient relationships architecturally improves generalization across atmospheric regimes with different stability characteristics.

This paper presents a systematic taxonomy of six physics integration strategies, evaluated through 54 model configurations (6 strategies $\times$ 3 architectures $\times$ 3 task weightings) on 222.6 million predictions (inference) spanning contrasting atmospheric regimes. The primary contributions are threefold. First, we demonstrate that explicit Richardson conditioning combined with simple MLP architectures provides 158% improvement in viscosity R^2 on cross-regime active turbulence compared to implicit approaches, while architecturally complex alternatives exhibit destabilizing variance amplification under distribution shift. Second, we establish that uncertainty-based task weighting [13] consistently outperforms both manual tuning and dynamic weight averaging by 20–30% for physics-coupled objectives. Third, our Kling-Gupta Efficiency analysis reveals systematic positive bias ($\beta > 1$) across all models, providing diagnostic evidence that cross-regime degradation stems from training data sparsity rather than fundamental physics incompatibility, suggesting that targeted data augmentation offers a more promising remediation path than architectural modifications.

To facilitate reproducibility and future research, complete results for all 54 configurations, including evaluation metrics, statistical analyses, extracted training/test data sets, and trained ML model checkpoints, are publicly available at https://doi.org/10.5281/zenodo.19487076. The associated training/evaluation scripts are available at https://github.com/pandasambit15/mtl-ml-sgs-ieee-wcci-ijcnn2026.git. Due to space constraints, tables in this paper present representative subsets; the repository contains full results for all architecture-strategy-weighting combinations.

II. Background

A. Subgrid-Scale Turbulence Physics

Atmospheric LES applies spatial filtering to the Navier-Stokes equations, decomposing velocity and scalar fields into resolved (grid-scale) and unresolved (subgrid-scale) components. For a generic variable ϕ, the filtered value $\bar{\phi}$ represents the resolved component while $\phi' = \phi - \bar{\phi}$ represents subgrid fluctuations. This decomposition introduces unclosed terms in the filtered equations: the SGS stress tensor $\tau_{ij} = \overline{u_i u_j} - \bar{u}_i \bar{u}_j$ for momentum and SGS scalar flux for potential temperature and other scalars [1].

The Smagorinsky-Lilly closure [4], [5], employed by MONC, parameterizes these unclosed terms through eddy viscosity (K_m) and eddy diffusivity (K_h) coefficients:

$$\tau_{ij} - \frac{1}{3}\tau_{kk}\delta_{ij} = -2K_m \bar{S}_{ij}, \quad F_j^{\theta} = -K_h \frac{\partial \bar{\theta}}{\partial x_j} \tag{1}$$

where δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ when $i = j$, otherwise 0), used here to subtract the isotropic (pressure-like) stress component $\frac{1}{3}\tau_{kk}$, leaving only the deviatoric (traceless) stress tensor responsible for anisotropic momentum transport. F_j^{θ} is the SGS scalar flux for potential temperature, $\bar{\theta}$ is the filtered potential temperature, x_j denotes spatial coordinates, and $\bar{S}_{ij}$ is the resolved strain rate tensor. The coefficients depend on grid spacing Δ, resolved strain magnitude $|\bar{S}|$, and atmospheric stability through the Richardson number:

$$Ri = \frac{(g/\theta_0)(\partial\bar{\theta}/\partial z)}{(\partial\bar{u}/\partial z)^2 + (\partial\bar{v}/\partial z)^2} \tag{2}$$

where g is gravitational acceleration, θ_0 is a reference potential temperature, z is the vertical coordinate, and $\bar{u}$, $\bar{v}$ are the filtered horizontal velocity components.

The Richardson number discriminates distinct turbulence regimes with different physical characteristics. Negative values ($Ri < 0$) indicate statically unstable stratification where buoyancy forces drive convective turbulence. Small positive values ($0 < Ri < Ri_c$) represent stable stratification where mechanical shear production maintains turbulence against buoyancy suppression. Large positive values ($Ri > Ri_c \approx 0.25$) indicate strongly stable conditions where stratification suppresses turbulence entirely. MONC's stability functions $f_m(Ri)$ and $f_h(Ri)$ encode these regime-dependent behaviors, smoothly transitioning coefficient values across stability boundaries. This explicit Richardson dependence in the operational parameterization motivates our investigation of whether neural network emulators should encode this relationship architecturally.

B. Multi-Task Learning for Physics-Coupled Objectives

Multi-task learning (MTL) jointly optimizes multiple related objectives, enabling beneficial knowledge transfer between tasks through shared representations [14]. For SGS emulation, MTL naturally arises from the physical structure: predicting coefficients (K_m, K_h), the stability indicator (Ri), and regime classification represent related tasks sharing common input features and underlying atmospheric physics.

However, balancing heterogeneous objectives requires principled weighting strategies. The uncertainty weighting method [13] treats task weights as learnable parameters representing homoscedastic uncertainty. For task k with loss $\mathcal{L}_k$, the weighted contribution becomes:

$$\mathcal{L}_k^{\text{weighted}} = \frac{1}{2\sigma_k^2}\mathcal{L}_k + \log\sigma_k \tag{3}$$

where σ_k is learned during training. The $\log\sigma_k$ regularization prevents degenerate solutions where uncertainties grow unboundedly. Dynamic Weight Averaging (DWA) [15] adapts weights based on relative learning progress, increasing weight for tasks with slower improvement.

Our approach falls within the hard constraint category of physics-informed learning [12]: we explicitly wire the Richardson-to-coefficient dependency that mirrors MONC's physical causality ($K_m, K_h = f(Ri)$), sometimes termed "physics-guided" learning. This imposes structural inductive bias persisting at inference time, distinct from soft constraints that add physics-based penalty terms to loss functions [16]. For climate applications, recent work has demonstrated benefits of physics-aware architectures for stable online coupling [17] and improved accuracy through architectural constraints enforcing physical bounds [18]. Our framework also has additional soft constraints in terms of loss functions penalizing negative coefficient (K_m, K_h) predictions, mimicking the soft physics constraint approach.

III. Methodology

A. Problem Formulation and Input Features

We formulate SGS coefficient emulation as a multi-task learning problem that jointly predicts four related quantities from local atmospheric state. The primary prediction targets are eddy viscosity K_m and eddy diffusivity K_h (both in m^2/s), which govern the magnitude of SGS momentum and scalar transport. Additionally, we predict Richardson number Ri as a dimensionless stability indicator and categorical stability regime classification distinguishing unstable ($Ri < 0$), transitional ($0 \leq Ri < 0.25$), and supercritical ($Ri \geq 0.25$) conditions.

Input features comprise a 54-dimensional vector encoding local atmospheric state across four categories. Six primary variables capture the thermodynamic and dynamic state: velocity components (u, v, w), potential temperature (θ), water vapor mixing ratio (q_v), and cloud liquid water (q_c). Spatial context is provided by 36 features representing six-point stencil derivatives for each primary variable, capturing the local gradient structure essential for turbulence detection. Six grid metadata features encode resolution information through grid spacings (Δx, Δy, Δz) and absolute height, enabling resolution-aware predictions. Finally, six geometric position features provide normalized coordinates and boundary proximity indicators.

The grid metadata features are necessary for mixed-resolution training where identical physical shear produces different numerical values at different grid spacings. The six-point stencil captures differential information analogous to what MONC's native Smagorinsky scheme uses for strain rate computation.

B. Physics Integration Strategies

The fundamental design choice in physics-aware SGS emulation concerns *when* physical knowledge influences predictions: during training only (through auxiliary task gradients) or during both training and inference (through architectural conditioning). We investigate six strategies spanning this spectrum (Figure 1, Table I).

TABLE I
Physics integration configuration summary. Each row specifies the input features provided to coefficient prediction heads (Coeff Input) and regime classification head (Regime Input), where $\mathbf{h}_s$ denotes shared backbone features, Ri denotes predicted Richardson number, and Reg denotes predicted regime. Training column indicates joint (all heads trained simultaneously) or staged (sequential freezing) optimization. Aux Tasks column lists which auxiliary predictions are included beyond the primary coefficient outputs.

Config	Coeff Input	Regime Input	Training	Aux Tasks
Baseline	$\mathbf{h}_s$	$\mathbf{h}_s$	Joint	Ri, Regime
Ri-Cond	$[\mathbf{h}_s; Ri]$	$\mathbf{h}_s$	Joint	Ri, Regime
Q1	$[\mathbf{h}_s; Ri]$	$[\mathbf{h}_s; Ri]$	Joint	Ri, Regime
Q2	$[\mathbf{h}_s; \mathrm{Reg}]$	$[\mathbf{h}_s; Ri]$	Joint	Ri, Regime
Q3	$[\mathbf{h}_s; \mathrm{Reg}]$	$[\mathbf{h}_s; Ri]$	Staged	Ri, Regime
Q4	$\mathbf{h}_s$	—	Joint	None

The Baseline configuration represents purely implicit physics learning: all prediction heads receive only shared backbone features $\mathbf{h}_s$, with Richardson number and regime classification serving as auxiliary tasks providing gradient regularization during training only. The Ri-Conditioned approach introduces explicit architectural dependency by feeding concatenated features $[\mathbf{h}_s; Ri_{\text{pred}}]$ to coefficient prediction heads, mirroring MONC's computational structure where coefficients depend on Richardson number through stability functions.

Beyond these two primary approaches, we evaluate four additional configurations. Q1 extends Richardson conditioning universally to all heads including regime classification. Q2 creates a two-stage cascade where regime classification receives Richardson-augmented features while coefficient heads receive regime-augmented features. Q3 employs identical architecture to Q2 but with sequential staged training: the Richardson head is trained first and frozen, then regime, then coefficients. Q4 serves as an ablation baseline with coefficient heads only and no auxiliary tasks.

C. Neural Network Architectures

We evaluate three architectures representing different inductive biases. The standard Multi-Layer Perceptron (MLP) employs a fully-connected backbone with layer dimensions 54→256→512→256, followed by task-specific heads (128→64→output), with Rectified Linear Unit (ReLU) activations and batch normalization, totaling 453K parameters. The Residual MLP (ResMLP) extends this with four residual blocks containing skip connections (256→256 each), enabling gradient flow through deeper networks, totaling 679K parameters. The Tabular Transformer (TabTransformer) applies self-attention to tabular data through four attention layers with 4 heads operating on 64-dimensional embeddings, totaling 943K parameters. We note that TabTransformer was not specifically optimized for tabular atmospheric data used in this case; architectures explicitly designed for spatio-temporal atmospheric fields (e.g., graph neural networks, domain-specific vision transformers) may yield different generalization characteristics and warrant future investigation. All architectures share identical prediction head structures and training hyperparameters.

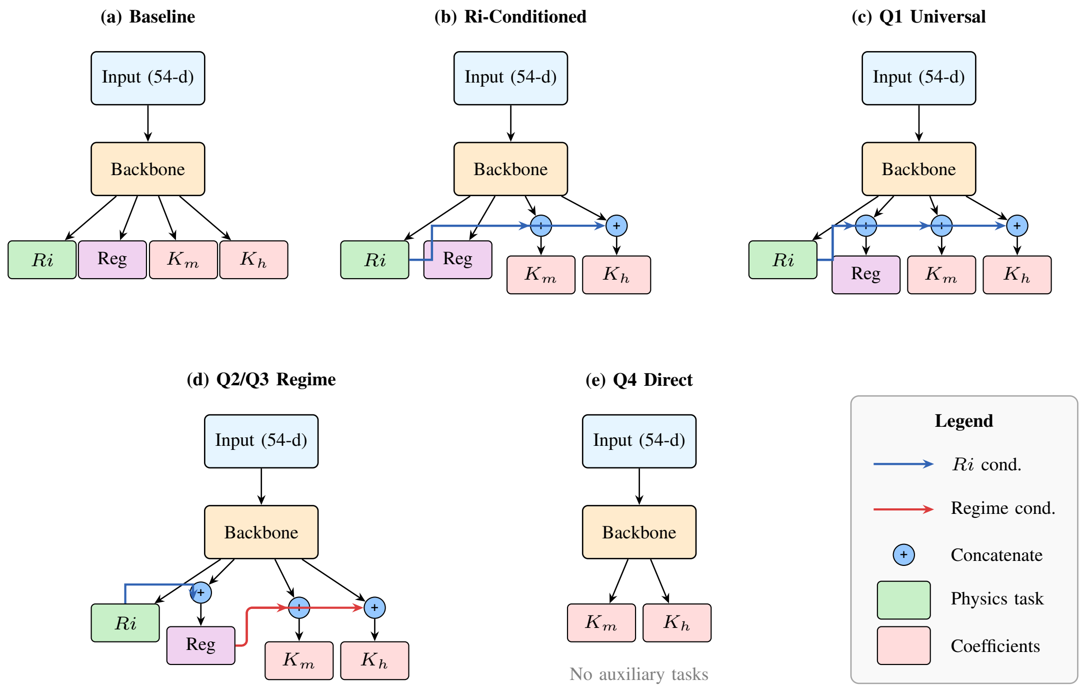

Fig. 1. Multi-task architectures for SGS coefficient prediction. (a) Baseline: Auxiliary tasks (Ri, Regime) provide gradient regularization only. (b) Ri-Conditioned: Predicted Ri concatenated before coefficient heads. (c) Q1: Ri conditioning applied universally. (d) Q2/Q3: Two-stage cascade with Ri conditioning regime, then regime conditioning coefficients; Q3 uses staged training. (e) Q4: Ablation baseline with coefficient heads only, no auxiliary tasks.

D. Training Data: Contrasting Atmospheric Regimes

Training data combines two atmospheric regimes with fundamentally different turbulence characteristics, creating a challenging mixed-distribution learning problem. Radiative Convective Equilibrium (RCE) simulations represent idealized deep tropical convection on $64 \times 64 \times 99$ grids with $\Delta \approx 1000\text{m}$, producing strong, quasi-equilibrium turbulence (mean $K_m = 8.9\ \text{m}^2/\text{s}$). Atmospheric Radiation Measurement (ARM) simulations represent shallow cumulus convection initialized from real-world observations on $192 \times 192 \times 220$ grids with $\Delta \approx 50\text{m}$, producing substantially weaker turbulence (mean $K_m = 0.48\ \text{m}^2/\text{s}$). Both datasets were generated using the MONC branch, obtained via the Met Office Science Repository Service (MOSRS), with local modifications limited to diagnostics and data extraction. The publicly available reference implementation of MONC is maintained at https://github.com/MetOffice/monc, which serves as the baseline for running MONC simulations.

The twenty-fold resolution difference poses additional challenges: identical physical gradients produce different numerical values at different grid spacings, requiring models to interpret magnitudes in conjunction with grid metadata. Regime-stratified sampling yielded 4.05 million training samples with 66% from RCE and 34% from ARM, reflecting relative data availability. This imbalance may contribute to observed bias toward RCE-calibrated predictions, though stratified sampling ensures all stability regimes are proportionally represented within each dataset. Future work should systematically investigate the effect of training data balance on cross-regime generalization.

Models were trained using AdamW optimizer ($\eta = 10^{-3}$, weight decay 0.3, cosine annealing) for up to 300 epochs with early stopping (patience: 30 epochs) on NVIDIA A100 GPUs. Three random seeds per configuration enabled statistical assessment.

The Loss function for the MTL set up (Eqn 4) was a weighted sum of individual task based losses, with equal weighted sum serving as the baseline and more advanced configurations using the different weighting methods discussed in II-B. The first 3 regression based loss components (L_{K_m}, L_{K_h} and L_{Ri}) were implemented using Huber Loss [19], which represent the loss functions for the turbulence coefficient and Ri predictions respectively. The L_{regime} was implemented using a cross-entropy loss function for the regime classification task. The last term, L_{pos}, is associated with physics-constrained loss for penalizing negative coefficient predictions, following the SGS implementation inside MONC. For the Q4 configuration of the models, the L_{regime} and L_{Ri} were set to zero. The actual implementation can be found at the GitHub repo (see I) that contains all the training and evaluation scripts used for this paper.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{K_m} + \mathcal{L}_{K_h} + \mathcal{L}_{Ri} + \mathcal{L}_{\text{regime}} + \mathcal{L}_{\text{pos}} \tag{4}$$

E. Evaluation Protocol

Full-domain inference extends beyond held-out test samples to complete spatial domains: 19.9 million predictions (49 timesteps) for RCE and 202.8 million predictions (25 timesteps) for ARM. We report coefficient of determination

(R^2), Pearson correlation (r), and variance ratio (VR $= \sigma_{\text{pred}}/\sigma_{\text{truth}}$). Kling-Gupta Efficiency (KGE) [20] decomposes performance into correlation, variability ratio (α), and bias ratio (β):

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \qquad (5)$$

Active turbulence filtering ($K > 10^{-10}$ m^2/s) isolates the 8–22% of domain volume where SGS effects matter most for simulation fidelity.

IV. Results

A. Task Weighting Comparison

Table II presents test-split viscosity R^2 across all 54 combinations (6 configurations $\times$ 3 architectures $\times$ 3 methods). The uncertainty weighting method consistently outperforms alternatives in all experiments, but the DWA fails significantly for ResMLP and TabTransformer across all configurations. Note that Baseline and Ri-Cond yield identical task weighting results because Richardson conditioning affects inference only. Q3-Manual exhibits unique collapse because staged training requires learned weighting to balance frozen upstream losses. We attribute DWA failure to oscillatory feedback between physics-coupled tasks (detailed trajectories available in our repository). Hence, all subsequent experiments use uncertainty weighting and the inference evaluations use the best-performing seed.

B. Cross-Dataset Performance Analysis

Table III presents full-domain performance across both atmospheric regimes, revealing dataset-dependent behavior patterns that constitute our central empirical finding. Q3-MLP achieves highest RCE performance ($R^2 = 0.811$ viscosity, 0.808 diffusivity) through staged training with regime classification as an intermediary. However, Ri-MLP dominates on ARM ($R^2 = 0.670$ viscosity, 0.734 diffusivity), providing 24% improvement over Baseline-MLP for viscosity. This divergence suggests explicit Richardson conditioning confers robustness benefits specifically when models encounter distribution shift.

The Q1 configuration demonstrates the limitations of naive physics integration: while achieving reasonable RCE performance ($R^2 = 0.722$), Q1 collapses on ARM with viscosity $R^2 = 0.119$. Conditioning all prediction heads amplifies error propagation: Richardson prediction errors cascade through regime classification, ultimately corrupting coefficient predictions. The selective conditioning of Ri-MLP, targeting only coefficient heads, avoids this failure while retaining physics integration benefits.

Comparing Q4-MLP (no auxiliary tasks) with Ri-MLP isolates multi-task learning's contribution: Ri-MLP achieves 4.4% higher ARM viscosity R^2 and 13.8% higher diffusivity R^2, demonstrating that auxiliary Richardson prediction provides meaningful gradient regularization even when Richardson conditioning is simultaneously applied.

C. Architectural Complexity and Generalization

Table IV and Figure 2 reveal dramatic architecture-dependent generalization patterns. ResMLP achieves competitive RCE performance ($R^2 \approx 0.78$), essentially matching MLP, yet produces deeply negative R^2 on ARM. Baseline-ResMLP degrades from $R^2 = 0.782$ on RCE to $R^2 = -1.51$ on ARM; Ri-ResMLP fares worse, dropping from 0.767 to -3.61. These negative values indicate predictions explaining less variance than simply predicting the mean.

The variance ratio illuminates the failure mechanism. ResMLP maintains VR ≈ 1.0 on RCE but produces VR $= 3.23$–5.30 on ARM, indicating overprediction by 3–5$\times$. Skip connections calibrated during training on RCE-dominated data learn to amplify strong turbulence signals characteristic of deep convection (mean $K_m = 8.9$ m^2/s). When ARM's substantially weaker signals (mean 0.48 m^2/s) pass through these same residual pathways, the preserved amplification produces predictions an order of magnitude too large.

TabTransformer exhibits similar but less severe failures. The Taylor diagrams in Figure 2 visualize these patterns compellingly: RCE shows tight clustering across all architectures near the reference point, while ARM reveals MLP variants maintaining proximity to reference as ResMLP and TabTransformer scatter toward inflated variance ratios. This finding is notable for the common practice of architecture selection based on validation metrics: in-distribution performance provides essentially no predictive signal for cross-regime generalization.

D. Active Turbulence Analysis

Full-domain metrics can obscure performance where it matters most, as most grid points contain quiescent air with negligible SGS coefficients. Additionally, the neural networks based ML SGS models with continuous activation functions inherently produce non-zero outputs, even at the locations where physical SGS coefficients are exactly zero in quiescent regions. This architectural artifact contributes to systematic over-prediction on full-domain metrics. To avoid this, table V presents results filtered to active turbulence regions (8–22% of domain volume), isolating performance where this artifact does not confound evaluation.

On ARM active regions, Ri-MLP achieves viscosity $R^2 = 0.432$ versus Baseline-MLP's 0.167, representing 158% relative improvement. Comparing Ri-MLP with Q4-MLP (which lacks auxiliary tasks) isolates the multi-task learning contribution: the 10% improvement (0.432 vs 0.392) demonstrates that auxiliary Richardson prediction provides meaningful gradient regularization beyond the conditioning mechanism itself. In active turbulence, strong velocity gradients prevent division-by-near-zero issues that can corrupt Richardson calculations in quiescent regions, providing meaningful stability information that architectural conditioning exploits.

All architectures achieve reasonable RCE active performance ($R^2 = 0.82$–0.86), confirming that complex architecture failures on ARM reflect cross-regime transfer problems rather than fundamental architectural limitations.

TABLE II

TASK WEIGHTING COMPARISON: R^2 ON HELD-OUT TEST SPLIT ACROSS ALL CONFIGURATIONS AND ARCHITECTURES (MEAN±STD, 3 SEEDS). THE UNCERTAINTY WEIGHTING METHOD WINS UNIVERSALLY. DWA FAILS SIGNIFICANTLY FOR RESMLP/TABTRANSFORMER (ITALICIZED). Q3-MANUAL COLLAPSES DUE TO STAGED TRAINING DYNAMICS. BASELINE AND RI-COND ARE IDENTICAL (CONDITIONING AFFECTS INFERENCE ONLY). THE Ri COLUMNS FOR Q4 ARE ABSENT DUE TO Q4 VARIANTS HAVING ONLY TWO PREDICTION HEADS ($Visc$ AND $Diff$)

		MLP			**ResMLP**			**TabT**		
Config	**Method**	Visc	Diff	Ri	Visc	Diff	Ri	Visc	Diff	Ri
Baseline/Ri	Manual	0.626±0.031	0.626±0.028	0.643±0.233	0.708±0.004	0.706±0.004	0.934±0.006	0.633±0.007	0.632±0.006	0.485±0.030
	Uncertainty	**0.835**±0.004	**0.831**±0.005	0.695±0.201	**0.869**±0.005	**0.867**±0.004	0.860±0.069	**0.821**±0.004	**0.818**±0.005	**0.979**±0.001
	DWA	0.601±0.047	0.593±0.052	0.723±0.074	*0.300*±0.013	*0.292*±0.011	0.724±0.057	*0.258*±0.014	*0.264*±0.013	0.850±0.100
Q1	Manual	0.626±0.026	0.624±0.028	0.642±0.230	0.710±0.003	0.706±0.004	0.936±0.009	0.617±0.022	0.617±0.018	0.487±0.019
	Uncertainty	**0.806**±0.005	**0.804**±0.006	0.478±0.193	**0.871**±0.002	**0.868**±0.002	0.894±0.020	**0.821**±0.003	**0.816**±0.004	**0.981**±0.001
	DWA	0.521±0.101	0.526±0.100	0.601±0.151	*0.304*±0.014	*0.302*±0.013	0.748±0.012	*0.249*±0.007	*0.246*±0.003	0.877±0.050
Q2	Manual	0.645±0.021	0.642±0.021	0.694±0.263	0.711±0.002	0.709±0.002	0.925±.003	0.614±0.058	0.612±0.053	0.642±0.295
	Uncertainty	**0.835**±0.004	**0.833**±0.005	**0.806**±0.057	**0.872**±0.004	**0.869**±0.005	0.880±0.009	**0.818**±0.005	**0.814**±0.004	**0.979**±0.004
	DWA	0.437±0.182	0.444±0.177	0.524±0.284	*0.325*±0.019	*0.327*±0.014	0.761±0.068	*0.326*±0.221	*0.361*±0.198	0.596±0.281
Q3	Manual	*0.352*±0.005	*0.349*±0.021	*0.099*±0.011	*0.486*±0.007	*0.484*±0.004	*0.293*±0.011	*0.544*±0.012	*0.539*±0.007	*0.274*±0.021
	Uncertainty	**0.865**±0.004	**0.863**±0.004	**0.812**±0.059	**0.877**±0.004	**0.876**±0.004	0.692±0.246	**0.789**±0.011	**0.787**±0.013	0.632±0.331
	DWA	*0.319*±0.032	*0.313*±0.014	0.655±0.032	*0.296*±0.018	*0.290*±0.018	0.593±0.065	*0.185*±0.003	*0.180*±0.007	0.538±0.343
Q4	Manual	0.675±0.010	0.671±0.011	—	0.708±0.003	0.706±0.002	—	0.654±0.007	0.651±0.006	—
	Uncertainty	**0.843**±0.008	**0.840**±0.009	—	**0.867**±0.002	**0.864**±0.002	—	**0.815**±0.010	**0.814**±0.009	—
	DWA	*0.289*±0.012	*0.302*±0.006	—	*0.259*±0.008	*0.255*±0.011	—	*0.261*±0.003	*0.255*±0.007	—

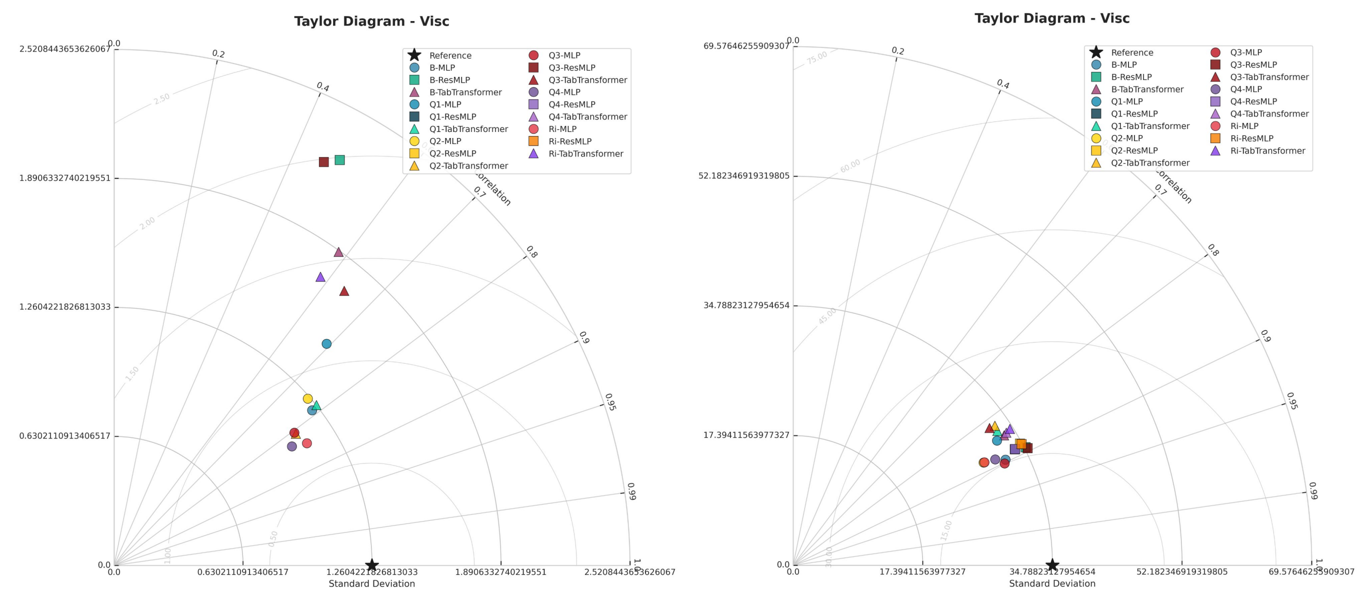

Fig. 2. Taylor diagrams [21] for viscosity coefficient prediction on ARM (left) and RCE (right), providing simultaneous visualization of correlation and variance agreement. Each point represents one model configuration; angular position indicates Pearson correlation with ground truth (perfect correlation at 0°), radial distance indicates predicted standard deviation (reference star marks ground truth standard deviation), and distance from reference indicates centered RMS error. Left (ARM): MLP variants (circles) cluster near reference, maintaining both high correlation and appropriate variance; ResMLP (triangles) and TabTransformer (squares) scatter toward inflated radial positions, indicating variance explosion under distribution shift. Right (RCE): All architectures cluster tightly near reference, demonstrating competitive in-distribution performance that fails to predict cross-regime generalization capability.

E. Diagnostic Analysis

KGE decomposition (Table VI) reveals two distinct failure modes. For MLP variants, systematic positive bias ($\beta = 1.44$–1.54) dominates the KGE penalty (78–89%), while correlation remains high ($r = 0.79$–0.85)—models capture relative patterns but fail to calibrate absolute magnitudes. In contrast, ResMLP failures are variance-dominated: $\alpha = 1.80$ on ARM versus 1.01 on RCE (contributing 61% of KGE penalty), with correlation also degrading ($r = 0.49$). Notably, Q2-TabTransformer partially escapes the TabTransformer failure pattern (ARM $R^2 = 0.62$), suggesting regime conditioning can stabilize attention mechanisms.

MLP variants maintain near-perfect physical constraint compliance (<0.03% negative predictions) across both regimes. ResMLP exhibits 3–4% violations at constant rates between regimes, indicating architectural rather than regime-specific issues.

V. DISCUSSION

A. Design Principles for Robust SGS Emulation

The systematic comparison yields coherent design principles for physics-informed atmospheric neural architectures, extending prior findings [8], [17]. The most consequential finding

TABLE III
CROSS-DATASET FULL-DOMAIN PERFORMANCE (MLP VARIANTS). Q3-MLP EXCELS ON IDEALIZED RCE; RI-MLP PROVIDES SUPERIOR CROSS-REGIME ROBUSTNESS ON ARM.

	RCE (19.9M)		**ARM (202.8M)**		**Mean R^2**	
Model	Visc	Diff	Visc	Diff	Visc	Diff
Q3-MLP	**0.811**	**0.808**	0.616	0.701	0.714	0.754
Baseline-MLP	0.801	0.798	0.540	0.644	0.671	0.721
Q4-MLP	0.784	0.782	0.642	0.645	0.713	0.713
Ri-MLP	0.773	0.770	**0.670**	**0.734**	**0.721**	**0.752**
Q2-MLP	0.772	0.769	0.485	0.647	0.629	0.708
Q1-MLP	0.722	0.721	0.119	0.424	0.421	0.572

TABLE IV
ARCHITECTURE COMPARISON: RCE VS ARM FULL-DOMAIN EVALUATION. VR = VARIANCE RATIO; ITALICS INDICATE FAILURE MODES (NEGATIVE R^2 OR VR > 2). RESMLP/TABTRANSFORMER ACHIEVE COMPETITIVE RCE PERFORMANCE BUT FAIL CATASTROPHICALLY ON ARM.

	RCE				**ARM**			
Model	R^2_V	R^2_D	r	VR	R^2_V	R^2_D	r	VR
Baseline-MLP	0.801	0.798	0.90	0.84	0.540	0.644	0.79	0.95
Baseline-ResMLP	0.782	0.778	0.89	1.01	*-1.51*	*-0.50*	0.49	*3.23*
Baseline-TabTr	0.713	0.702	0.85	0.91	*-0.60*	*-0.51*	0.58	*2.24*
Ri-MLP	0.773	0.770	0.88	0.70	**0.670**	**0.734**	**0.84**	**0.78**
Ri-ResMLP	0.767	0.763	0.88	0.99	*-3.61*	*-1.81*	0.37	*5.30*
Ri-TabTr	0.696	0.692	0.85	0.97	*-0.37*	*-0.26*	0.58	*1.89*

TABLE V
ACTIVE TURBULENCE PERFORMANCE, FILTERED TO GRID POINTS WITH $K > 10^{-10}$ m^2/s (8–22% OF DOMAIN). RI-MLP ACHIEVES 158% IMPROVEMENT OVER BASELINE-MLP ON ARM; COMPARISON WITH Q4-MLP (NO AUXILIARY TASKS) ISOLATES MULTI-TASK LEARNING CONTRIBUTION.

Data	**Model**	R^2_V	R^2_D	VR	r	%Active
RCE	Q3-MLP	**0.862**	**0.860**	0.82	0.93	8.3%
	Q4-MLP	0.832	0.830	0.77	0.89	8.3%
	Ri-MLP	0.823	0.820	0.70	0.91	8.3%
ARM	**Ri-MLP**	**0.432**	**0.579**	0.78	0.84	21.8%
	Q4-MLP	0.392	0.438	0.69	0.83	21.8%
	Baseline-MLP	0.167	0.417	0.95	0.79	21.8%

TABLE VI
KGE DECOMPOSITION FOR VISCOSITY: α = VARIABILITY RATIO, β = BIAS RATIO. MLP FAILURES ARE BIAS-DOMINATED ($\beta > 1$); RESMLP FAILURES ARE VARIANCE-DOMINATED ($\alpha \gg 1$, ITALICIZED).

Data	**Model**	KGE	r	α	β
RCE	Q3-MLP	0.809	0.901	0.905	1.13
	Baseline-MLP	0.767	0.895	0.914	1.19
	Baseline-ResMLP	0.815	0.892	1.005	1.15
ARM	Q3-MLP	**0.498**	0.805	0.868	1.44
	Ri-MLP	0.426	**0.845**	0.885	1.54
	Baseline-ResMLP	-0.025	0.487	*1.80*	1.39

concerns architectural complexity: simple MLPs with explicit Richardson conditioning achieve the best balance between accuracy and generalization, providing 158% improvement in viscosity R^2 on cross-regime active turbulence. Residual connections and attention mechanisms exhibit destabilizing variance amplification under distribution shift, consistent with recent observations [9]. Architectural simplicity with targeted physics conditioning should be prioritized over complex architectures relying on implicit pattern learning.

Conditioning strategy should match deployment conditions: regime-aware staged conditioning (Q3) maximizes in-distribution performance, while Richardson conditioning (Ri-MLP) provides improved cross-regime robustness. The stark failure of Q1 demonstrates that targeted physics integration substantially outperforms naive expansion.

B. Failure Mode Taxonomy

KGE decomposition reveals three distinct failure modes with different remediation pathways, consistent with the taxonomy developed for ML-based climate parameterizations [18]. *Bias-dominated failure* (MLP variants): high correlation ($r > 0.79$) with systematic positive bias ($\beta = 1.4$–1.6), where β contributes 78–89% of the KGE penalty. Models correctly capture relative patterns but fail to calibrate absolute magnitudes. This failure mode may respond to post-hoc calibration or targeted data augmentation focusing on underrepresented active turbulence periods.

Variance-dominated failure (ResMLP variants): variance ratio $\alpha = 1.8$–2.9 contributes 60–82% of the KGE penalty, with correlation also degrading ($r = 0.35$–0.49). Skip connections preserve signal amplification calibrated for training conditions, producing severe over-prediction under distribution shift. Remediation requires architectural changes such as constraining residual pathways or regime-specific normalization; post-hoc calibration cannot address variance instability.

Error cascading (Q1 configuration): maintained constraint compliance with severe accuracy degradation ($R^2 = 0.12$). Richardson prediction errors propagate through universal conditioning pathways, corrupting downstream coefficient predictions. Selective conditioning (Ri-MLP) limiting propagation to coefficient heads successfully avoids this failure.

C. On Active Turbulence Performance

The best active turbulence viscosity R^2 on ARM (0.432 for Ri-MLP) represents 158% improvement over baseline while maintaining high correlation ($r = 0.84$). The systematic positive bias ($\beta = 1.54$) suggests post-hoc calibration could further improve accuracy. Whether this suffices for online stability requires coupling experiments.

D. Role of Resolution in Cross-Regime Transfer

The nineteen-fold difference in mean viscosity between RCE ($8.9\ \mathrm{m}^2/\mathrm{s}$) and ARM ($0.48\ \mathrm{m}^2/\mathrm{s}$) roughly corresponds to the twenty-fold resolution difference ($\Delta_{\text{RCE}} \approx 1000$m vs $\Delta_{\text{ARM}} \approx 50$m), consistent with Smagorinsky scaling $K_m \propto \Delta^2$. However, resolution and atmospheric regime differ simultaneously; future work should evaluate generalization across resolutions within a single regime to disentangle these factors.

E. Limitations

This study evaluates generalization across two atmospheric regimes. Extending this evaluation to other diverse atmospheric regimes like marine stratocumulus, frontal systems, and stable nocturnal boundary layers remains essential future work. The TabTransformer serves as a representative attention baseline, but architectures designed for spatio-temporal atmospheric data may differ in performance and accuracy. Training data imbalance (66% RCE, 34% ARM) likely contributes to the observed ARM bias. ML SGS models inherently produce non-zero predictions even where physical coefficients are exactly zero, contributing to systematic over-prediction; mitigating this through hard output constraints or zero-penalizing loss functions could be a potential solution, to be tested in a future work. Most importantly, validating through online coupling with MONC, currently not presented here, is required to assess numerical stability, energy conservation, and the actual potential computational speedup.

VI. Conclusions

This systematic evaluation of 54 model configurations across 222.6 million predictions establishes that physics-guided neural network design determines cross-regime generalization capability in atmospheric turbulence emulation. Explicit Richardson conditioning combined with simple MLP architectures provides 158% improvement in viscosity R^2 (0.432 vs 0.167), especially on the dynamically important active turbulence regions, compared to the other implicit approaches. Architecturally complex alternatives incorporating residual connections or attention mechanisms demonstrate significant predictive instability on simulation data from different atmospheric regimes, with implications for model selection where predictive robustness matters under data distribution shifts.

Uncertainty-based task weighting consistently outperforms both manual tuning and dynamic weight averaging by 20–30%, establishing this approach as a robust option for geophysical multi-task learning problems. The failure of DWA despite success in computer vision highlights the importance of understanding physics couplings when transferring methods across domains.

The maintained physical constraint compliance despite degraded cross-regime accuracy provides diagnostic evidence that data coverage limitations, rather than fundamental physics incompatibility, drive generalization failures in this study. Models learned valid physical relationships but from insufficient samples covering the under-represented regime's active turbulence conditions. This suggests that targeted data augmentation and scaling, specifically prioritizing active convection periods and diverse atmospheric regimes, offer a viable remediation pathway without requiring fundamental architectural changes.

In the path toward operational implementation, future work will focus on validation through online coupling experiments to assess stability under feedback dynamics.

Acknowledgments

This work was supported by the University of Reading Advancing the Frontiers of Earth System Prediction (AFESP) Programme [A3720300] and the UK Met Office. We acknowledge JASMIN for computational resources and technical support.

References

[1] S. B. Pope, *Turbulent Flows*. Cambridge, UK: Cambridge University Press, 2000.

[2] T. Schneider, J. Teixeira, C. S. Bretherton, F. Brient, K. G. Pressel, C. Schär, and A. P. Siebesma, "Climate goals and computing the future of clouds," *Nature Climate Change*, vol. 7, no. 1, pp. 3–5, 2017.

[3] A. R. Brown, R. T. Cederwall, A. Chlond, P. G. Duynkerke, J.-C. Golaz, M. Khairoutdinov, D. C. Lewellen, A. P. Lock, M. K. MacVean, C.-H. Moeng, *et al.*, "Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land," *Quarterly Journal of the Royal Meteorological Society*, vol. 128, no. 582, pp. 1075–1093, 2002.

[4] J. Smagorinsky, "General circulation experiments with the primitive equations: I. The basic experiment," *Monthly Weather Review*, vol. 91, no. 3, pp. 99–164, 1963.

[5] D. K. Lilly, "On the numerical simulation of buoyant convection," *Tellus*, vol. 14, no. 2, pp. 148–172, 1962.

[6] A. Pal, "Deep learning emulation of subgrid-scale processes in turbulent shear flows," *Geophysical Research Letters*, vol. 47, no. 12, p. e2020GL087005, 2020.

[7] D. Kochkov, J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer, "Machine learning–accelerated computational fluid dynamics," *Proceedings of the National Academy of Sciences*, vol. 118, no. 21, p. e2101784118, 2021.

[8] S. Rasp, M. S. Pritchard, and P. Gentine, "Deep learning to represent subgrid processes in climate models," *Proceedings of the National Academy of Sciences*, vol. 115, no. 39, pp. 9684–9689, 2018.

[9] H. Mandler and B. Weigand, "Generalization limits of data-driven turbulence models," *Flow, Turbulence and Combustion*, 2024.

[10] R. Molinaro, S. Lanthaler, B. Raonić, T. Rohner, V. Armegioiu, S. Simonis, D. Grund, Y. Ramic, Z. Y. Wan, F. Sha, S. Mishra, and L. Zepeda-Núñez, "Generative AI for fast and accurate statistical computation of fluids," *arXiv preprint arXiv:2409.18359*, 2024.

[11] M. Sakarvadia, K. Hegazy, A. Totounferoush, K. Chard, Y. Yang, I. Foster, and M. W. Mahoney, "The false promise of zero-shot super-resolution in machine-learned operators," *arXiv preprint arXiv:2510.06646*, 2024.

[12] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[13] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

[14] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[15] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1871–1880, 2019.

[16] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[17] J. Yuval and P. A. O'Gorman, "Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions," *Nature Communications*, vol. 11, no. 1, p. 3295, 2020.

[18] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, "Enforcing analytic constraints in neural networks emulating physical systems," *Physical Review Letters*, vol. 126, no. 9, p. 098302, 2021.

[19] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[20] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling," *Journal of Hydrology*, vol. 377, no. 1-2, pp. 80–91, 2009.

[21] K. E. Taylor, "Summarizing multiple aspects of model performance in a single diagram," *Journal of Geophysical Research: Atmospheres*, vol. 106, no. D7, pp. 7183–7192, 2001.