



Bounded Transport-Residual Closure for Solver-Coupled Graph-Based Global Atmospheric Forecasting

Chuanhu Deng^{1,2}, Chaorong Li¹, Wenjie Luo^{1,2}, Zhuo Wang^{1,2}, and Lihong Zhu¹

¹School of Computer Science and Technology (School of Artificial Intelligence), Yibin University, Yibin 644000, China

²College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China

Correspondence: Chaorong Li (lichorong88@163.com)

Abstract. Solver-coupled learning offers a structured alternative to direct state-to-state regression for global atmospheric forecasting, but its forecast quality depends on how accurately the embedded coefficient dynamics represent unresolved forcing and coarse-grid numerical effects. We represent these unresolved effects using a bounded transport-residual closure (BTRC) embedded in an explicit sub-stepping solver. Building on the reduced horizontal transport-coefficient dynamics of PASSAT, we separate a retained spherical advection–curvature operator from aggregate terms omitted, approximated, or numerically distorted by the reduced model, including pressure-gradient, Coriolis, frictional, subgrid, vertical-coupling, and discretization contributions. A graph network estimates a raw aggregate equation discrepancy, which is mapped through a scaled hyperbolic tangent to obtain the BTRC. The resulting closure is inserted into the coefficient-update pathway and modifies the transport coefficients used by subsequent atmospheric state updates. The bound limits the direct closure-induced coefficient increment per Euler sub-step to $\delta t \tau$, providing a numerical safeguard against unconstrained learned forcing; it is not presented as a proof of global dynamical stability. The BTRC is shared across prognostic channels as a compact directional correction, while an auxiliary node-level gated wavelet branch refines multi-resolution graph features. Experiments on coarse-resolution ERA5 data under WeatherBench-style protocols show the clearest gains for temperature and 500-hPa geopotential (Z500). Near-surface wind fields exhibit more consistent improvements in ACC than in RMSE, and the latter gains are small or not uniformly positive. A residual-free control, $\tilde{R}_t \equiv 0$, provides targeted evidence that the learned in-solver BTRC materially contributes to the long-lead improvements within the evaluated setting. These results support bounded transport-residual learning as a promising mechanism for improving autoregressive behavior in reduced solver-coupled atmospheric models.

1 Introduction

Global atmospheric forecasting from reanalysis-derived gridded fields is an important problem in geoscience information processing. Reanalysis products such as ERA5 integrate diverse observations and model information to provide spatially continuous estimates of atmospheric states, enabling the development and evaluation of data-driven forecasting models Hersbach et al. (2020). Standardized benchmarks, including WeatherBench Rasp et al. (2020) and WeatherBench 2 Rasp et al. (2024), have further promoted reproducible comparison of global data-driven forecasting systems. Although operational numerical weather prediction (NWP) systems have achieved remarkable progress through advances in dynamical cores, data assimilation



25 lation, and high-performance computing, they remain computationally expensive and may still suffer from errors related to imperfect initial conditions, numerical discretization, and unresolved subgrid processes Bauer et al. (2015); Dueben and Bauer (2018). These challenges motivate efficient learning-based forecasting methods that can maintain both accuracy and stability over medium-range lead times.

Recent learning-based forecasting models have progressed from early convolutional and recurrent predictors Scher (2018); Weyn et al. (2020); Arcomano et al. (2020) to graph-based models Keisler (2022); Lam et al. (2023), neural-operator and spectral architectures Pathak et al. (2022), transformer-based systems Bi et al. (2023); Chen et al. (2023a), hybrid graph–transformer forecasters Lang et al. (2024), cascade and probabilistic generative systems Chen et al. (2023b); Price et al. (2025), end-to-end forecasting systems Allen et al. (2025), and foundation models Nguyen et al. (2023); Bodnar et al. (2025). These models can produce global forecasts much faster than traditional solvers and have demonstrated strong performance on standard
35 benchmarks. However, most data-driven forecasters primarily learn a direct mapping from historical atmospheric states to future states. Such direct-regression formulations provide limited explicit control over physically meaningful intermediate quantities, such as transport coefficients, local tendencies, or unresolved correction terms. This limitation makes it difficult to diagnose and mitigate error accumulation during autoregressive rollouts, where small stepwise errors may accumulate into large-scale drift Schultz et al. (2021); Dueben and Bauer (2018).

40 Physics-guided learning provides a complementary route toward improving stability and interpretability. Representative approaches include physics-informed neural networks Raissi et al. (2019); Karniadakis et al. (2021), neural differential equation models Chen et al. (2018); Verma et al. (2024), physics-assisted graph forecasting Zheng et al. (2025), differentiable dynamical-core learning Kochkov et al. (2024), and learned physical parameterizations or residual closures Rasp et al. (2018); Brenowitz and Bretherton (2018, 2019); Bolton and Zanna (2019); Mardani et al. (2025). Some existing hybrid approaches incorporate
45 physical knowledge through loss-level constraints, external parameterizations, differentiable operators, or post-hoc residual corrections. These designs motivate tighter solver coupling, where learned equation-related terms are consumed inside the inference-time update rather than only used as training penalties or output-level corrections.

Fig. 1 illustrates the motivation and scope of the proposed formulation. PASSAT already couples a spherical advection equation with numerically evolved velocity-like transport coefficients. Our starting point is therefore not a new coefficient-
50 evolution mechanism. Instead, we ask a narrower question: how should a reduced coefficient equation represent terms that are omitted, approximated, or numerically distorted when embedded in an autoregressive learned forecaster? We represent this mismatch using a bounded transport-residual closure (BTRC) placed inside the coefficient update, thereby modifying the transport trajectory consumed by subsequent state sub-steps.

The distinction is important. A post-hoc residual predictor corrects the forecast state after the numerical update and can
55 obscure whether the error originated from transport or from local tendencies. In contrast, the proposed BTRC changes the intermediate transport trajectory consumed by subsequent state sub-steps. This explicitly localizes the closure within the solver and enables a controlled residual-free ablation in which the same update is retained but $\tilde{R}_t$ is set to zero.

In this paper, we develop a solver-coupled graph forecaster centered on BTRC. The graph encoder predicts the local interaction tendency, the initial transport coefficients, and a common directional residual estimate for the reduced coefficient

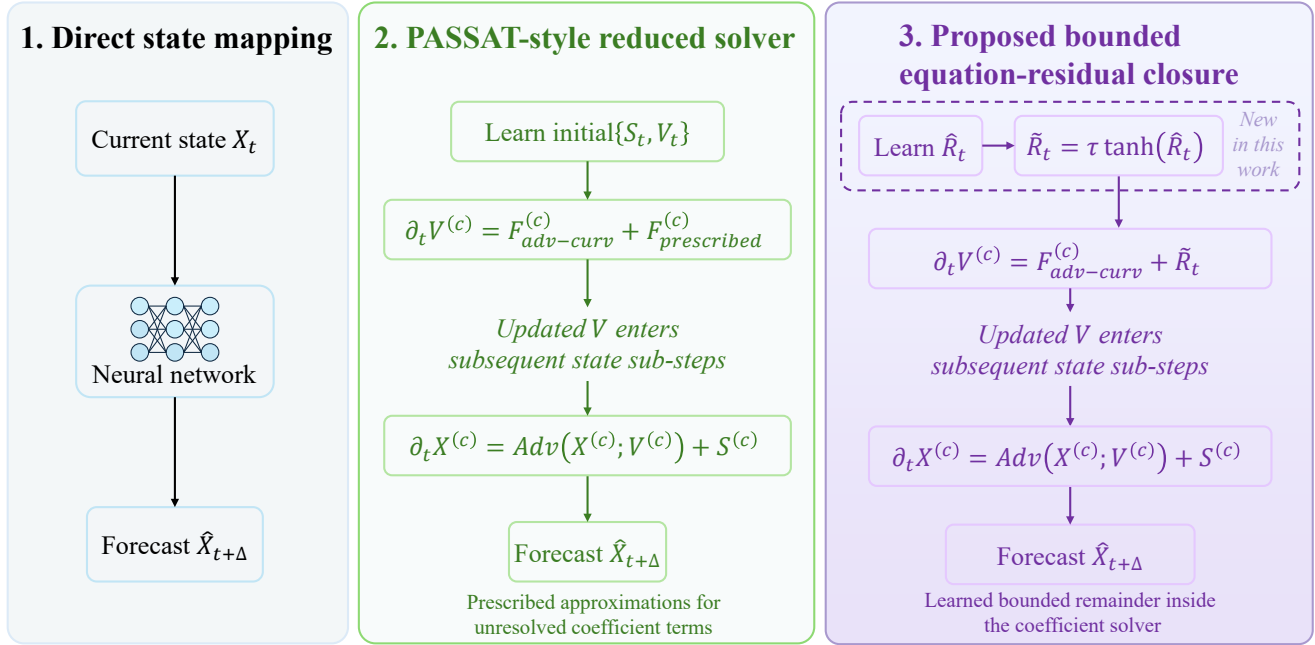


Figure 1. Conceptual positioning of the proposed bounded transport-residual closure (BTRC). Direct state mapping predicts the future atmospheric state without exposing equation-related intermediate variables. A PASSAT-style reduced solver predicts interaction tendencies and initial velocity-like transport coefficients and evolves the coefficients using a simplified spherical transport-coefficient update. The proposed method retains this solver-coupled structure but represents the unresolved coefficient-equation remainder using the BTRC, $\tilde{R}_t = \tau \tanh(\hat{R}_t)$. The closure is applied inside the coefficient update and modifies the transport coefficients consumed by subsequent state sub-steps. The highlighted BTRC pathway, rather than coefficient evolution itself, is the methodological contribution of this work.

60 equations. The raw estimate is mapped through $\tau \tanh(\cdot)$ to obtain the bounded closure used during explicit Euler integration. The saturation map is motivated by its smoothness, odd symmetry, near-linear response around zero, and finite range, which together limit the direct learned forcing applied in one sub-step. A node-level gated wavelet branch is retained as an auxiliary representation-refinement component rather than the central methodological claim.

The main contributions are summarized as follows:

- 65 – We formulate BTRC as a bounded approximation to the discrepancy between fuller horizontal transport-coefficient dynamics and a retained spherical advection–curvature operator. This formulation identifies plausible discrepancy sources—including omitted pressure-gradient, Coriolis, frictional, subgrid, vertical-coupling, and discretization effects—without claiming that the learned field uniquely recovers any individual physical force.
- We introduce the in-solver parameterization $\tilde{R}_t = \tau \tanh(\hat{R}_t)$ and derive its per-sub-step correction bound. This provides a numerical safeguard for learned forcing while explicitly avoiding the stronger claim of a formal global-stability theorem.



- We evaluate BTRC using two complementary controls. Comparison with the capacity-matched PASSAT precursor evaluates the complete reformulation, whereas the $\tilde{R}_t \equiv 0$ variant isolates the effect of the learned residual within the otherwise matched coefficient-update pathway. The node-level wavelet branch is analyzed as a secondary feature-refinement component.

2 Related Work

2.1 Data-Driven Global Atmospheric Forecasting

Data-driven global atmospheric forecasting has advanced rapidly with the availability of large-scale reanalysis products and standardized benchmarks. Early studies explored convolutional, recurrent, and hybrid neural architectures for learning atmospheric evolution from gridded fields Scher (2018); Weyn et al. (2020); Arcomano et al. (2020). Keisler further demonstrated the potential of graph neural networks for global weather forecasting Keisler (2022). Recent large-scale systems have achieved strong medium-range forecast skill, including FourCastNet Pathak et al. (2022), Pangu-Weather Bi et al. (2023), GraphCast Lam et al. (2023), FuXi Chen et al. (2023b), FengWu Chen et al. (2023a), AIFS Lang et al. (2024), GenCast Price et al. (2025), Aurora Bodnar et al. (2025), and Aardvark Weather Allen et al. (2025). Foundation models such as ClimaX further demonstrate the potential of pretraining across heterogeneous weather and climate datasets Nguyen et al. (2023).

Most of these systems primarily learn a mapping from historical atmospheric states to future states. Although this formulation is highly expressive and computationally efficient, it provides limited explicit access to equation-related intermediate quantities. In particular, forecast errors cannot be readily attributed to inaccurate transport, missing local tendencies, unresolved processes, or numerical accumulation. This limitation motivates structured forecasting approaches in which learned quantities are consumed inside an explicit state-evolution procedure rather than used only to produce the final forecast field.

2.2 Physics-Guided and Solver-Coupled Forecasting Models

Physics-guided learning introduces scientific knowledge through governing equations, physical constraints, differentiable operators, or structured latent variables. Physics-informed neural networks incorporate partial differential equation residuals into the training objective Raissi et al. (2019), while broader physics-informed machine learning frameworks combine data-driven learning with scientific models and constraints Karniadakis et al. (2021); Willard et al. (2022). Such loss-level constraints can encourage physical consistency, but they do not by themselves determine how learned terms participate in the inference-time evolution of a high-dimensional atmospheric state.

A more tightly coupled family of methods embeds learned dynamics, physical operators, or differentiable solvers directly into the forecasting process. Neural ordinary differential equations provide a continuous-time framework for learning dynamical systems Chen et al. (2018). ClimODE applies a physics-informed neural ODE formulation to climate and weather forecasting Verma et al. (2024), whereas NeuralGCM combines learned physical parameterizations with a differentiable atmospheric dynamical core Kochkov et al. (2024). PASSAT couples topology-aware graph learning with a simplified spherical



momentum solver Zheng et al. (2025). These methods motivate the use of learned equation-related terms inside numerical state evolution rather than only as training regularizers or output-level corrections.

105 The following paragraphs clarify the methodological boundary between the proposed model and the three most closely related solver-coupled frameworks.

Relationship to PASSAT: PASSAT is the direct methodological precursor of this work. It already predicts variable-specific velocity-like transport fields and local interaction tendencies using a topology-aware graph network, and it advances the transport fields through a simplified horizontal spherical momentum formulation Zheng et al. (2025). Consequently, spherical
 110 transport, transport-coefficient evolution, graph-based interaction estimation, and explicit numerical sub-stepping are inherited premises rather than new contributions of the present study. The distinction lies in the treatment of the unresolved coefficient-equation remainder. Whereas PASSAT relies on prescribed approximations for terms such as pressure-gradient, Coriolis, and frictional effects, the proposed formulation retains the spherical advection–curvature operator and represents the aggregate discrepancy of omitted, approximated, and numerically distorted terms using the proposed bounded transport-residual closure
 115 (BTRC). The BTRC modifies the transport trajectory used by subsequent atmospheric state updates.

Relationship to ClimODE: ClimODE learns continuous-time atmospheric evolution through a physics-informed neural ODE Verma et al. (2024). It does not use the same decomposition into variable-specific transport coefficients, local interaction tendencies, and a reduced horizontal coefficient equation. The learned vector field in ClimODE and the bounded coefficient-equation residual considered here therefore act on different governing formulations. The contribution of the present work is
 120 not neural continuous-time integration itself, but the placement of a bounded learned equation discrepancy inside a reduced transport-coefficient solver.

Relationship to NeuralGCM: NeuralGCM combines learned physical parameterizations with a substantially more complete differentiable atmospheric dynamical core Kochkov et al. (2024). Its learned components represent unresolved physical processes within a global circulation model, whereas the present work considers a narrower setting: a coarse-resolution reduced
 125 horizontal transport-coefficient solver derived from a PASSAT-style formulation. We do not claim to replace a dynamical core or to identify individual physical parameterizations. Instead, we study whether the aggregate equation discrepancy of a reduced coefficient model can be learned in a bounded form and consumed inside the inference-time solver. Accordingly, the novelty lies in the formulation, bounding, and in-solver application of the coefficient-equation remainder rather than in graph representation, coefficient evolution, or solver coupling individually.

130 **2.3 Geometry-Aware and Multi-Resolution Representations**

Global atmospheric fields are defined on the Earth’s spherical surface and contain spatial interactions over a broad range of scales. Forecasting architectures must therefore account for spherical geometry while representing both large-scale circulation and localized structures.

Geometry-aware representations have been developed using graphs, meshes, spherical convolutions, and spectral operators.
 135 Graph neural networks provide a flexible mechanism for representing non-Euclidean domains and relational physical interactions Battaglia et al. (2018); Gilmer et al. (2017). Mesh-based message-passing methods have also been applied to physical



simulation and neural PDE solving Pfaff et al. (2020); Brandstetter et al. (2022). In global atmospheric forecasting, Keisler’s model, GraphCast, and PASSAT use graph- or mesh-based connectivity to model spatial interactions without relying exclusively on planar latitude–longitude convolutions Keisler (2022); Lam et al. (2023); Zheng et al. (2025). Spherical convolutional and spectral methods provide alternative geometry-aware representations Cohen et al. (2018); Esteves et al. (2018); Bonev et al. (2023).

In the proposed framework, the spherical graph serves as the representation substrate for estimating the equation-related terms required by the numerical solver. The graph encoder predicts local interaction tendencies, initial transport coefficients, and a raw coefficient-equation residual. The final atmospheric forecast is then generated by the solver-coupled state and coefficient evolution described in Section 3, rather than by direct graph-based regression of the future state.

Multi-resolution modeling provides a complementary mechanism for processing atmospheric structures at different spatial scales. Fourier and neural-operator methods support efficient global mixing Li et al. (2020); Guibas et al. (2021); Pathak et al. (2022), whereas wavelet representations provide spatially localized frequency decompositions Mallat (1989). Wavelet and multi-wavelet neural operators have been applied to the learning of multi-scale physical systems Tripura and Chakraborty (2022); Gupta et al. (2021); Hu et al. (2025). In our framework, the Node-WT branch reshapes node features to the latitude–longitude grid, applies gated wavelet refinement, and then returns the corrected features to graph order. Edge features remain in the topology-aware message-passing pathway and are not transformed by the wavelet operator. This module is used as an auxiliary representation refinement mechanism rather than as the central solver-coupling contribution.

3 Methodology

The overall framework is shown in Fig. 2. The methodological focus is the BTRC pathway inside the explicit solver. The graph encoder supplies local tendencies, initial transport coefficients, and a raw estimate of the unresolved coefficient-equation remainder. The auxiliary wavelet branch modifies only the node representation used by these heads.

3.1 Solver-Coupled Forecasting With Learned Equation Terms

We formulate global atmospheric forecasting as a solver-coupled learning problem. Let X_t denote the resolved atmospheric state at time t . Instead of learning an unconstrained mapping from X_t to $X_{t+\Delta}$, we learn a set of equation-related terms that parameterize a discrete time-stepping update:

$$\hat{X}_{t+\Delta} = \mathcal{I}_{\Delta} \left(X_t; S_t, V_t, \tilde{R}_t \right), \quad (1)$$

where $\mathcal{I}_{\Delta}(\cdot)$ denotes a numerical update over lead interval Δ , S_t denotes local interaction tendencies, V_t denotes dynamic transport coefficients, and $\tilde{R}_t$ denotes the BTRC. This formulation allows the neural network to predict equation-structured intermediate components while the final state update is performed through an explicit integration scheme.

The proposed formulation is related to neural differential equation models and physics-guided forecasting frameworks Chen et al. (2018); Verma et al. (2024); Kochkov et al. (2024). However, unlike methods that use physical constraints only as

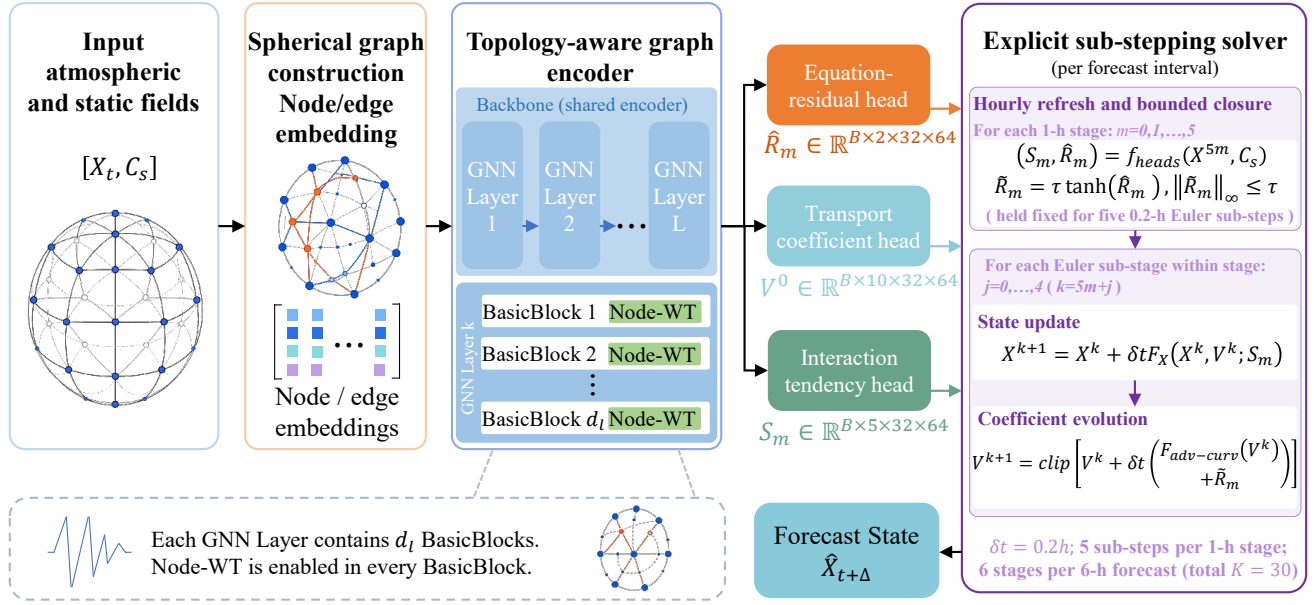


Figure 2. Overview of the proposed BTRC-centered solver-coupled forecasting framework. Atmospheric and static fields are embedded on a spherical graph and processed by a topology-aware graph encoder. Each GNN layer contains multiple BasicBlocks, and a Node-WT refinement module is enabled in every BasicBlock; The encoder produces a raw two-component equation residual $\hat{R}_m$, variable-specific transport coefficients V^0 , and interaction tendencies S_m . During each 1-h physical stage, the residual and interaction tendency are refreshed, and the BTRC $\tilde{R}_m = \tau \tanh(\hat{R}_m)$ is held fixed for five 0.2-h Euler sub-steps. Within each sub-step, the atmospheric state is advanced using the current transport coefficients and interaction tendency, after which the transport coefficients are evolved using the retained advection–curvature operator and the bounded closure. Six stages produce one 6-h forecast, corresponding to a total of $K = 30$ Euler sub-steps.

training penalties, our approach consumes learned equation-related terms inside the inference-time update. The model therefore operates as a structured forecaster that couples state evolution, transport-coefficient evolution, and residual correction during autoregressive rollouts.

3.2 Problem Setup and Spherical Graph Representation

Let $X_t \in \mathbb{R}^{C \times H \times W}$ denote the atmospheric state at time t on a global latitude–longitude grid. In this work, we use $C = 5$ prognostic variables:

$$\{t2m, t850, z500, u10, v10\}. \quad (2)$$

Let $C_s \in \mathbb{R}^{C_0 \times H \times W}$ denote static geophysical variables, such as latitude, longitude, land–sea mask, and orography. We flatten the grid into $N = HW$ nodes and construct a fixed spherical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.



The node feature at time t is defined as

$$n_t = \text{vec}([X_t, C_s]) \in \mathbb{R}^{N \times (C+C_0)}. \quad (3)$$

For each edge $(i, j) \in \mathcal{E}$, we encode geometric descriptors:

$$e_{ij} = [\Delta\theta_{ij}, \Delta\phi_{ij}, d_{ij}] \in \mathbb{R}^3, \quad (4)$$

where θ and ϕ denote latitude and longitude in radians, and d_{ij} denotes the great-circle distance between nodes i and j . For efficiency, we use sparse connectivity based on local spherical neighborhoods and a normalized sparse adjacency matrix.

Graph representations have been widely used to model relational and mesh-based physical systems Battaglia et al. (2018); Pfaff et al. (2020); Brandstetter et al. (2022). For global weather forecasting, graph-based models are particularly suitable because they can encode spherical topology and non-local spatial interactions Keisler (2022); Lam et al. (2023); Zheng et al. (2025).

3.3 Transport–Tendency–BTRC Coupled Update

The central idea of the proposed method is to predict the components required by a structured update rather than directly predicting the next state. We decompose the learned update into three parts: local interaction tendencies, dynamic transport coefficients, and the BTRC.

3.3.1 Multi-Head Estimation

The graph encoder produces a latent node representation $h_t \in \mathbb{R}^{N \times D}$. From this representation, separate prediction heads estimate the equation-related terms:

$$S_t = f_S(h_t) \in \mathbb{R}^{N \times C}, \quad (5)$$

$$[V_t, \hat{R}_t] = f_V(h_t) \in \mathbb{R}^{N \times (2C+2)}. \quad (6)$$

Here, S_t represents local interaction tendencies, $V_t \in \mathbb{R}^{N \times 2C}$ contains variable-specific transport coefficients $\{v_\theta^{(c)}, v_\phi^{(c)}\}_{c=1}^C$, and $\hat{R}_t = (\hat{R}_\theta, \hat{R}_\phi) \in \mathbb{R}^{N \times 2}$ is a common directional raw transport-residual estimate. Broadcasting this two-component field across channels imposes a low-rank inductive bias: the BTRC represents spatially coherent meridional and zonal corrections shared by the reduced coefficient solver, rather than a separately identifiable physical forcing for each prognostic variable. This compact choice reduces closure degrees of freedom, but it cannot represent all channel-specific omitted physics and is therefore treated as a modeling limitation rather than a physical identity.



3.3.2 State Update With Transport and Tendency Terms

For each prognostic variable $c \in \{1, \dots, C\}$, the state tendency is modeled as

$$\frac{\partial X^{(c)}}{\partial t} = -v_{\theta}^{(c)} \frac{\partial X^{(c)}}{\partial \theta} - v_{\phi}^{(c)} \frac{1}{\cos \theta} \frac{\partial X^{(c)}}{\partial \phi} + S^{(c)}. \quad (7)$$

The transport coefficients V_t parameterize the advection-like part of the update, while S_t captures local tendencies, source–sink effects, and cross-variable interactions not represented by the transport term. This separation makes the learned components more interpretable than direct state regression and enables targeted stabilization through coefficient evolution and bounded closure.

3.3.3 Residual Formulation of Transport-Coefficient Dynamics

PASSAT already evolves velocity-like transport coefficients through a simplified spherical momentum equation Zheng et al. (2025). Therefore, coefficient evolution, solver coupling, and explicit sub-stepping are inherited premises rather than new contributions of the present study. The methodological distinction is the introduction of BTRC inside the coefficient-evolution pathway. At the equation level, this distinction can be summarized as

$$\text{Reduced coefficient model: } \frac{\partial V^{(c)}}{\partial t} = F_{\text{ret}}^{(c)}(V), \quad (8)$$

$$\text{Proposed closed model: } \frac{\partial V^{(c)}}{\partial t} = F_{\text{ret}}^{(c)}(V) + \tilde{R}_t. \quad (9)$$

Here, $F_{\text{ret}}^{(c)}$ denotes the reduced spherical advection–curvature operator retained by the coefficient solver. Thus, the contribution of the present work is not coefficient evolution itself, but the formulation and in-solver application of the BTRC term $\tilde{R}_t$.

To clarify the origin of this term, let $F_{\text{full}}^{(c)}(V, X; \Pi)$ denote a fuller coefficient tendency, where Π represents auxiliary quantities and physical processes that are unavailable, simplified, or unresolved in the reduced forecasting system. We write

$$\frac{\partial V^{(c)}}{\partial t} = F_{\text{full}}^{(c)}(V, X; \Pi) = F_{\text{ret}}^{(c)}(V) + R_{\star}^{(c)}(V, X; \Pi), \quad (10)$$

where the idealized full-model discrepancy is defined as

$$R_{\star}^{(c)}(V, X; \Pi) = F_{\text{full}}^{(c)}(V, X; \Pi) - F_{\text{ret}}^{(c)}(V). \quad (11)$$

The discrepancy $R_{\star}^{(c)}$ may be schematically decomposed as

$$R_{\star}^{(c)} = R_{\text{forc}}^{(c)} + R_{\text{unres}}^{(c)} + \varepsilon_{\text{model}}^{(c)} + \varepsilon_{\text{disc}}^{(c)}. \quad (12)$$

Here, $R_{\text{forc}}^{(c)}$ represents forcing contributions that are omitted or only approximately represented by the reduced coefficient equation, potentially including pressure-gradient, Coriolis, and frictional effects. $R_{\text{unres}}^{(c)}$ collects unresolved processes such as vertical coupling, boundary-layer dynamics, subgrid mixing, and interactions with atmospheric variables not included in the present five-variable state. The term $\varepsilon_{\text{model}}^{(c)}$ represents errors caused by the reduced two-dimensional coefficient formulation



230 and the use of learned transport coefficients as proxy quantities, whereas $\varepsilon_{\text{disc}}^{(c)}$ denotes spatial-derivative and time-integration errors.

Equation (12) represents plausible sources of the equation discrepancy rather than a supervised decomposition available during training. Since the model is optimized only through forecast errors in normalized state space, the individual contributions cannot be identified separately. The learned BTRC should therefore be interpreted as an aggregate effective transport residual
 235 rather than as an estimate of any single physical force or parameterization.

Using the learned residual, whose bounded parameterization is introduced in the following subsection, the latitudinal coefficient evolves as

$$\begin{aligned} \frac{\partial v_{\theta}^{(c)}}{\partial t} = & -v_{\theta}^{(c)} \frac{\partial v_{\theta}^{(c)}}{\partial \theta} - v_{\phi}^{(c)} \frac{1}{\cos \theta} \frac{\partial v_{\theta}^{(c)}}{\partial \phi} \\ & - \left(v_{\phi}^{(c)}\right)^2 \tan \theta + \tilde{R}_{\theta,t}, \end{aligned} \quad (13)$$

240 and the longitudinal coefficient evolves as

$$\begin{aligned} \frac{\partial v_{\phi}^{(c)}}{\partial t} = & -v_{\theta}^{(c)} \frac{\partial v_{\phi}^{(c)}}{\partial \theta} - v_{\phi}^{(c)} \frac{1}{\cos \theta} \frac{\partial v_{\phi}^{(c)}}{\partial \phi} \\ & + v_{\theta}^{(c)} v_{\phi}^{(c)} \tan \theta + \tilde{R}_{\phi,t}. \end{aligned} \quad (14)$$

The learned BTRC acts as an effective closure for the reduced coefficient equation and modifies the transport trajectory used by subsequent atmospheric state updates. It is therefore neither a post-hoc correction to the forecast state X nor a replacement
 245 for the local interaction tendency S in Eq. (7).

3.3.4 BTRC Parameterization and Numerical Rationale

The graph encoder predicts a raw two-component residual field

$$\hat{R}_t = \left(\hat{R}_{\theta,t}, \hat{R}_{\phi,t} \right) \in \mathbb{R}^{N \times 2}. \quad (15)$$

The two directional components are shared across the variable-specific coefficient pairs. This shared formulation provides a
 250 compact low-rank inductive bias for spatially coherent directional corrections and limits the number of unconstrained closure degrees of freedom. It is an architectural regularization choice rather than a physical identity, and it does not imply that the unresolved processes are identical across all prognostic variables.

The raw network output is mapped to the solver through

$$\tilde{R}_t = \tau \tanh \left(\hat{R}_t \right), \quad (16)$$

255 where $\tau > 0$ sets the maximum BTRC magnitude in normalized coefficient-tendency units. The hyperbolic tangent is used as a smooth saturation map: it is odd, differentiable, approximately linear near the origin, and saturates for large positive or negative inputs. It therefore allows small residual predictions to pass approximately linearly while preventing arbitrarily large



learned forcing. We use this form as a numerical design choice rather than claiming that $\tanh$ is a physical constitutive law or the unique optimal bounding function.

260 Since $|\tanh(z)| \leq 1$, the closure satisfies

$$\|\tilde{R}_t\|_{\infty} \leq \tau. \quad (17)$$

For an explicit Euler sub-step of size δt , its direct contribution to the coefficient increment satisfies

$$\|\delta t \tilde{R}_t\|_{\infty} \leq \delta t \tau. \quad (18)$$

Let $\tilde{R}^{(k)} = \tilde{R}_{\lfloor k/5 \rfloor}$ denote the stage-wise BTRC applied at Euler sub-step k . Although the closure is refreshed every five sub-
 265 steps, each stage-wise field satisfies $\|\tilde{R}^{(k)}\|_{\infty} \leq \tau$. Therefore, over a lead interval $\Delta = K\delta t$, the accumulated directly injected contribution before nonlinear amplification and coefficient clipping satisfies

$$\begin{aligned} \left\| \sum_{k=0}^{K-1} \delta t \tilde{R}^{(k)} \right\|_{\infty} &\leq \sum_{k=0}^{K-1} \delta t \|\tilde{R}^{(k)}\|_{\infty} \\ &\leq K\delta t \tau = \Delta \tau. \end{aligned} \quad (19)$$

These inequalities bound only the forcing inserted directly by the stage-wise BTRC. They do not prove stability of the complete nonlinear atmosphere–coefficient system, whose behavior also depends on the retained operator, spatial discretization,
 270 state–coefficient feedback, local interaction tendencies, and repeated autoregressive network calls. We therefore use “toward stable forecasting” to describe the empirical objective and reserve “bounded” for the mathematically explicit closure property.

We additionally clip the evolved transport coefficients to a fixed range. The latitude-dependent factors in Eqs. (7)–(14) are evaluated on WeatherBench latitude points that exclude the exact poles, and a lower bound on $|\cos \theta|$ is applied when evaluating $1/\cos \theta$ and $\tan \theta$. Closure bounding, coefficient clipping, and metric regularization address different numerical
 275 pathways: the first limits the learned forcing, the second limits the coefficient magnitude, and the third mitigates high-latitude metric amplification.

3.3.5 Explicit Sub-Stepping Integration

The learned terms are integrated using an explicit Euler scheme with sub-stepping. Each 6-h forecast interval is divided into six 1-h physical stages. At the beginning of stage $m \in \{0, \dots, 5\}$, the network refreshes the interaction tendency and raw equation
 280 residual:

$$(S_m, \hat{R}_m) = f_{\text{heads}}(X^{5m}, C_s). \quad (20)$$

The bounded closure is then computed as

$$\tilde{R}_m = \tau \tanh(\hat{R}_m). \quad (21)$$

Both S_m and $\tilde{R}_m$ are held fixed during the following five Euler sub-steps.



285 For $j = 0, \dots, 4$, let $k = 5m + j$. Following the implemented update order, the atmospheric state is first advanced using the current transport coefficients:

$$X^{k+1} = X^k + \delta t F_X(X^k, V^k; S_m). \quad (22)$$

The transport coefficients are subsequently updated as

$$V^{k+1} = \text{clip} \left[V^k + \delta t \left(F_{\text{adv-curv}}(V^k) + \tilde{R}_m \right) \right]. \quad (23)$$

290 Here, $\delta t = 0.2$ h. Six stages and five sub-steps per stage yield $K = 30$ Euler updates for one 6-h forecast interval.

The closure is not added directly to the atmospheric state. Because the state update at sub-step k uses V^k , the coefficient correction producing V^{k+1} affects subsequent state increments rather than the current state increment. The final forecast is $\hat{X}_{t+\Delta} = X^K$.

3.4 Graph Encoder and Auxiliary Wavelet Refinement

295 The architecture consists of node and edge embeddings, topology-aware graph blocks, a gated wavelet refinement branch, and multi-head predictors for the solver-coupled update.

3.4.1 Node and Edge Embedding

We project node and edge features into a latent space of dimension D using multi-layer perceptrons:

$$h^{(0)} = \text{MLP}_n(n_t), \quad g^{(0)} = \text{MLP}_e(e), \quad (24)$$

300 where $h^{(0)} \in \mathbb{R}^{B \times N \times D}$ and $g^{(0)} \in \mathbb{R}^{B \times |\mathcal{E}| \times D}$ denote latent node and edge features.

3.4.2 Topology-Aware Message Passing

Each graph block updates edge and node representations using two message-passing stages. First, edge features are updated from incident node features:

$$g_{ij} \leftarrow g_{ij} + f_e(g_{ij}, h_i, h_j). \quad (25)$$

305 Then, node features aggregate messages from neighboring edges:

$$h_i \leftarrow h_i + f_n \left(h_i, \text{Agg}_{j:(i,j) \in \mathcal{E}} g_{ij} \right), \quad (26)$$

where $\text{Agg}(\cdot)$ denotes sparse edge-to-node aggregation. We further propagate spatial information through a sparse graph convolution:

$$h \leftarrow h + f_c(h, Ah), \quad (27)$$

310 where A is the normalized sparse adjacency matrix. This design combines local message passing with efficient sparse spatial propagation.

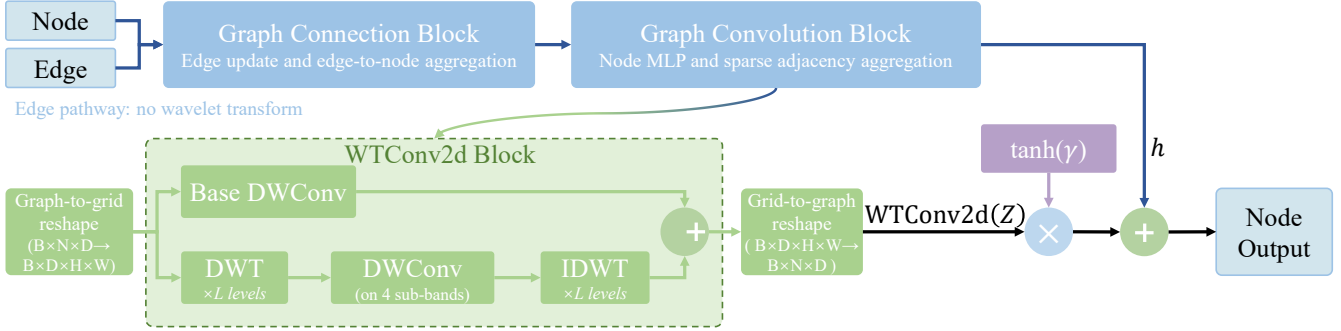


Figure 3. Structure of the Node-WT-enhanced BasicBlock. Node and edge features are first processed by the Graph Connection Block, which performs edge updating and edge-to-node aggregation. The resulting node features are then updated by the Graph Convolution Block through a node MLP and sparse adjacency aggregation. The node representation is subsequently split into an identity pathway and a gated wavelet-refinement pathway. In the wavelet pathway, node features are reshaped from graph order to the latitude–longitude grid and processed by WTConv2d, which combines a base depthwise convolution with multi-level discrete wavelet decomposition, subband depthwise convolution, and inverse wavelet reconstruction. The reconstructed correction is returned to graph order, scaled by $\tanh(\gamma)$, and added to the identity node representation. Edge features remain in the graph message-passing pathway and are not transformed by the wavelet module.

3.4.3 Auxiliary Gated Wavelet Multi-Resolution Refinement

Standard graph message passing may oversmooth fine-scale structures and offers limited explicit control over spatial–frequency components. To introduce frequency-aware residual refinement, we apply a node-level gated wavelet refinement branch, as illustrated in Fig. 3. Given node features $h \in \mathbb{R}^{B \times N \times D}$, we reshape them into a latitude–longitude feature map $Z \in \mathbb{R}^{B \times D \times H \times W}$ according to their grid positions. This operation is well-defined for node features because each node corresponds to a grid cell. Edge features are not passed through the wavelet branch; they remain in the topology-aware message-passing pathway to encode geometric relationships and support node–edge aggregation. We then apply a wavelet convolution operator:

$$\begin{aligned} \text{WTConv2d}(Z) = & \text{DWConv}(Z) \\ & + \text{IDWT}(\text{Conv}(\text{DWT}(Z))), \end{aligned} \quad (28)$$

where DWT and IDWT denote discrete wavelet transform and inverse discrete wavelet transform, respectively. The corrected signal is injected through a gated residual connection:

$$h \leftarrow h + \tanh(\gamma) \cdot \text{vec}(\text{WTConv2d}(Z)), \quad (29)$$

where the gate parameter is initialized to $\gamma = 0.1$, corresponding to $\tanh(\gamma) \approx 0.10$, to introduce multi-scale corrections gradually during training.

The wavelet branch is motivated by multi-resolution analysis Mallat (1989) and recent wavelet neural operators for scientific machine learning Tripura and Chakraborty (2022); Gupta et al. (2021); Hu et al. (2025). Compared with global spectral



transforms Li et al. (2020); Bonev et al. (2023), wavelet representations provide localized spatial–frequency corrections that are useful for residual refinement.

330 3.5 Training Objective

The model is trained end-to-end using a multi-step autoregressive forecasting loss. For a rollout horizon of T steps, the latitude-weighted prediction loss is

$$\mathcal{L}_x = \frac{1}{T} \sum_{\tau=1}^T \left\| W^{1/2} \odot \left(\hat{X}_{t+\tau\Delta} - X_{t+\tau\Delta} \right) \right\|_2^2, \quad (30)$$

where W denotes cosine-latitude weights. To encourage smooth and stable transport fields, we add regularization on the transport coefficients:

$$\mathcal{L} = \mathcal{L}_x + \lambda_v \|V\|_2^2 + \lambda_g \|\nabla_{\text{grid}} V\|_2^2, \quad (31)$$

where λ_v and λ_g control the strength of coefficient magnitude and gradient regularization. During training, gradients are backpropagated through the graph encoder, prediction heads, wavelet refinement module, and explicit sub-stepping update.

4 Experiments

340 4.1 Experimental Setup

We evaluate the BTRC-centered solver-coupled graph network on ERA5 using a WeatherBench-style experimental protocol Hersbach et al. (2020); Rasp et al. (2020, 2024). The existing experiments answer two closure-specific questions without claiming a complete physical validation: whether the full reformulation improves the capacity-matched PASSAT precursor, and whether removing the residual while retaining the coefficient pathway degrades the forecast. The node-level wavelet branch is treated as a secondary component.

4.2 Dataset, Metrics, and Baselines

All experiments use the ERA5 global reanalysis dataset Hersbach et al. (2020). ERA5 provides gridded atmospheric fields derived from the assimilation of diverse observations and model information, making it a suitable reference dataset for evaluating global data-driven atmospheric forecasting. Following the WeatherBench protocol Rasp et al. (2020), atmospheric variables are regridded to a 5.625° latitude–longitude grid, corresponding to $H \times W = 32 \times 64$, and sampled every 6 hours.

We evaluate the five prognostic variables defined in Eq. (2): t2m, t850, z500, u10, and v10. Here, t2m denotes 2-m temperature, t850 denotes 850-hPa temperature, z500 denotes 500-hPa geopotential (Z500), and u10/v10 denote 10-m zonal and meridional wind components. Following the WeatherBench convention, Z500 is retained as geopotential rather than converted to geopotential height, and its physical unit is $\text{m}^2 \text{s}^{-2}$. Static geophysical fields, including latitude, longitude, land–sea mask,



and orography, are used as auxiliary inputs. The training period is 1979–2015, the validation year is 2016, and the test period is 2017–2018. During evaluation, all models perform autoregressive inference at 6-hour intervals up to 144 hours.

We report latitude-weighted root mean square error (RMSE) and anomaly correlation coefficient (ACC), which are standard metrics in WeatherBench-style global forecasting evaluation Rasp et al. (2020, 2024). RMSE evaluates absolute forecast error, while ACC reflects the ability to preserve large-scale atmospheric patterns relative to climatology. The reported RMSE units are K for T2M and T850, $\text{m}^2 \text{s}^{-2}$ for Z500, and m s^{-1} for U10 and V10; ACC is dimensionless.

We compare with two groups of references. The primary reference is the capacity-matched PASSAT implementation Zheng et al. (2025). PASSAT already evolves velocity-like transport coefficients through a simplified spherical transport-coefficient update; our comparison therefore evaluates the complete BTRC-centered reformulation, not the introduction of coefficient evolution. The direct effect of the learned residual is instead isolated by the $\tilde{R}_t \equiv 0$ control in Table 2. The second group contains our capacity-matched reimplementations of representative architectures, denoted GraphCast-CM Lam et al. (2023), Pangu-CM Bi et al. (2023), FourCastNet-CM Pathak et al. (2022), SFNO-CM Bonev et al. (2023), FengWu-CM Chen et al. (2023a), and ClimODE-CM Verma et al. (2024). They are trained from scratch with the same data split, 5.625° grid, five variables, rollout horizon, and approximately 1.15M parameters. These identifiers do not refer to the official large-scale systems and must not be interpreted as operational-scale comparisons. IFS T42 is included only as a historical NWP reference when available.

4.3 Implementation Details

All learnable models are implemented in PyTorch and trained from scratch on 8 NVIDIA RTX 3090 GPUs. We use AdamW with a cosine-annealing learning-rate schedule. The base learning rate is 1×10^{-3} , the minimum learning rate is 3×10^{-7} , and a one-epoch warm-up is used. The batch size is 4, and each model is trained for 200 epochs. Automatic mixed precision is enabled for computational efficiency. Gradient norm clipping with a threshold of 0.3 is used to improve training stability.

For model configuration, the hidden embedding dimension is set to 48. The graph encoder contains two topology-aware graph blocks, and the prediction heads use lightweight branch-specific depths for transport and interaction estimation. The spherical graph is constructed from the latitude–longitude grid using local great-circle neighborhoods, with the same graph construction applied to all graph-based models. The node-level wavelet refinement branch uses Daubechies-1 wavelets with one decomposition level and a kernel size of 5. The wavelet gate is initialized to 0.1, allowing the multi-resolution correction to be introduced gradually during training.

During training, autoregressive rollouts use 4 forecast steps, corresponding to 24 h. During evaluation, all models are rolled out for 24 forecast steps, corresponding to 144 h. The explicit solver uses Euler sub-stepping with $\delta t = 0.2$ h, resulting in $K = 30$ Euler updates for each 6-hour forecast interval. The BTRC is defined as $\tilde{R}_t = \tau \tanh(\hat{R}_t)$ with $\tau = 10^{-3}$. The learned transport coefficients are clipped to $[-0.005, 0.005]$ after each coefficient-evolution step. The transport regularization weights in Eq. (31) are set to $\lambda_v = 10$ and $\lambda_g = 1$.

Unless otherwise specified, all learnable neural baselines use the same input variables, static geophysical fields, data split, grid resolution, rollout horizon, and evaluation metrics. Their backbone widths and depths are adjusted to obtain comparable



Table 1. Main comparison on WeatherBench in terms of RMSE↓/ACC↑. RMSE units are K for T2M and T850, m^2s^{-2} for 500-hPa geopotential (Z500), and ms^{-1} for U10 and V10; ACC is dimensionless. GraphCast-CM, Pangu-CM, FourCastNet-CM, SFNO-CM, FengWu-CM, and ClimODE-CM are our capacity-matched reimplementations under the same data split, 5.625° resolution, five variables, rollout horizon, and approximately 1.15M parameter budget. They are not the official large-scale systems. IFS T42 is reported only as a historical NWP reference when available. “–” denotes unavailable entries.

	Lead (h)	Ours	PASSAT Zheng et al. (2025)	GraphCast-CM Lam et al. (2023)	ClimODE-CM Verma et al. (2024)	Pangu-CM Bi et al. (2023)	FourCastNet-CM Pathak et al. (2022)	SFNO-CM Bonev et al. (2023)	FengWu-CM Chen et al. (2023a)	IFS T42 Rasp et al. (2020)
t2m	24	1.17 / 0.974	1.25 / 0.970	1.21 / 0.972	1.59 / 0.950	1.31 / 0.968	1.29 / 0.969	1.43 / 0.962	1.30 / 0.968	--
	48	1.45 / 0.961	1.54 / 0.955	1.52 / 0.956	2.23 / 0.910	1.63 / 0.950	1.58 / 0.953	1.75 / 0.942	1.60 / 0.952	--
	72	1.75 / 0.943	1.85 / 0.935	1.84 / 0.937	--	1.94 / 0.929	1.87 / 0.934	2.02 / 0.922	1.88 / 0.933	3.21 / 0.870
	96	2.06 / 0.921	2.16 / 0.911	2.17 / 0.912	--	2.26 / 0.903	2.16 / 0.911	2.30 / 0.899	2.17 / 0.911	--
	120	2.35 / 0.897	2.44 / 0.886	2.48 / 0.886	--	2.55 / 0.877	2.44 / 0.888	2.55 / 0.875	2.43 / 0.888	3.69 / 0.830
	144	2.60 / 0.873	2.69 / 0.862	2.75 / 0.861	--	2.80 / 0.853	2.68 / 0.865	2.76 / 0.853	2.67 / 0.866	--
t850	24	1.24 / 0.969	1.29 / 0.966	1.29 / 0.966	1.46 / 0.960	1.40 / 0.960	1.39 / 0.960	1.51 / 0.953	1.36 / 0.962	--
	48	1.61 / 0.947	1.66 / 0.943	1.69 / 0.941	2.26 / 0.900	1.83 / 0.931	1.78 / 0.934	1.91 / 0.924	1.75 / 0.936	--
	72	2.06 / 0.911	2.12 / 0.906	2.17 / 0.902	--	2.31 / 0.888	2.24 / 0.895	2.37 / 0.881	2.19 / 0.899	3.09 / 0.860
	96	2.54 / 0.864	2.59 / 0.859	2.67 / 0.851	--	2.79 / 0.835	2.71 / 0.845	2.84 / 0.828	2.65 / 0.851	--
	120	2.98 / 0.812	3.03 / 0.806	3.12 / 0.796	--	3.21 / 0.780	3.12 / 0.792	3.25 / 0.773	3.06 / 0.800	3.83 / 0.780
	144	3.35 / 0.761	3.39 / 0.756	3.49 / 0.745	--	3.55 / 0.732	3.47 / 0.743	3.59 / 0.724	3.41 / 0.753	--
z500	24	156 / 0.988	171 / 0.986	180 / 0.984	209 / 0.980	196 / 0.981	188 / 0.983	205 / 0.979	177 / 0.985	--
	48	274 / 0.963	293 / 0.958	307 / 0.954	452 / 0.910	337 / 0.944	313 / 0.951	342 / 0.942	302 / 0.955	--
	72	399 / 0.920	420 / 0.912	438 / 0.904	--	477 / 0.885	442 / 0.901	478 / 0.883	428 / 0.907	489 / 0.900
	96	525 / 0.860	543 / 0.849	565 / 0.838	--	605 / 0.813	566 / 0.835	607 / 0.809	549 / 0.844	--
	120	638 / 0.791	651 / 0.780	675 / 0.766	--	712 / 0.737	674 / 0.763	718 / 0.731	656 / 0.774	743 / 0.780
	144	733 / 0.721	738 / 0.713	767 / 0.698	--	797 / 0.668	760 / 0.695	807 / 0.658	743 / 0.708	--
u10	24	1.51 / 0.932	1.54 / 0.929	1.59 / 0.924	1.77 / 0.910	1.68 / 0.914	1.68 / 0.914	1.83 / 0.898	1.65 / 0.918	--
	48	2.19 / 0.853	2.22 / 0.848	2.29 / 0.838	2.73 / 0.760	2.41 / 0.818	2.38 / 0.824	2.50 / 0.802	2.33 / 0.832	--
	72	2.86 / 0.743	2.90 / 0.733	2.97 / 0.718	--	3.08 / 0.692	3.05 / 0.702	3.18 / 0.672	2.98 / 0.716	--
	96	3.43 / 0.623	3.46 / 0.611	3.54 / 0.593	--	3.61 / 0.565	3.60 / 0.578	3.73 / 0.541	3.51 / 0.596	--
	120	3.86 / 0.514	3.89 / 0.501	3.97 / 0.485	--	4.01 / 0.457	4.00 / 0.472	4.13 / 0.434	3.91 / 0.491	--
	144	4.19 / 0.423	4.19 / 0.414	4.27 / 0.399	--	4.29 / 0.373	4.29 / 0.389	4.41 / 0.349	4.21 / 0.406	--
v10	24	1.55 / 0.929	1.57 / 0.926	1.62 / 0.921	1.80 / 0.900	1.71 / 0.912	1.72 / 0.911	1.86 / 0.895	1.68 / 0.915	--
	48	2.23 / 0.848	2.26 / 0.842	2.32 / 0.833	2.73 / 0.770	2.44 / 0.814	2.42 / 0.818	2.53 / 0.798	2.37 / 0.825	--
	72	2.92 / 0.732	2.95 / 0.722	3.03 / 0.708	--	3.14 / 0.680	3.12 / 0.689	3.24 / 0.662	3.04 / 0.703	--
	96	3.52 / 0.602	3.55 / 0.589	3.64 / 0.572	--	3.70 / 0.542	3.70 / 0.553	3.83 / 0.519	3.62 / 0.571	--
	120	4.00 / 0.480	4.01 / 0.468	4.10 / 0.452	--	4.13 / 0.420	4.14 / 0.435	4.27 / 0.399	4.05 / 0.454	--
	144h	4.35 / 0.380	4.34 / 0.372	4.44 / 0.355	--	4.43 / 0.324	4.44 / 0.344	4.57 / 0.307	4.35 / 0.360	--

parameter budgets of approximately 1.15M. The reported comparison is therefore intended to evaluate architectural design and solver coupling under a controlled capacity-matched setting, rather than differences caused by larger model scale or external pretraining.

4.4 Main Quantitative Results

Table 1 reports RMSE and ACC under the common coarse-resolution protocol. The clearest and most consistent gains over PASSAT occur for temperature and geopotential variables. For near-surface winds, ACC improves more consistently than RMSE; the 144-h U10 RMSE is essentially unchanged and the V10 RMSE is slightly worse. We therefore interpret the wind results as improved large-scale pattern correlation without a corresponding uniform improvement in amplitude error.

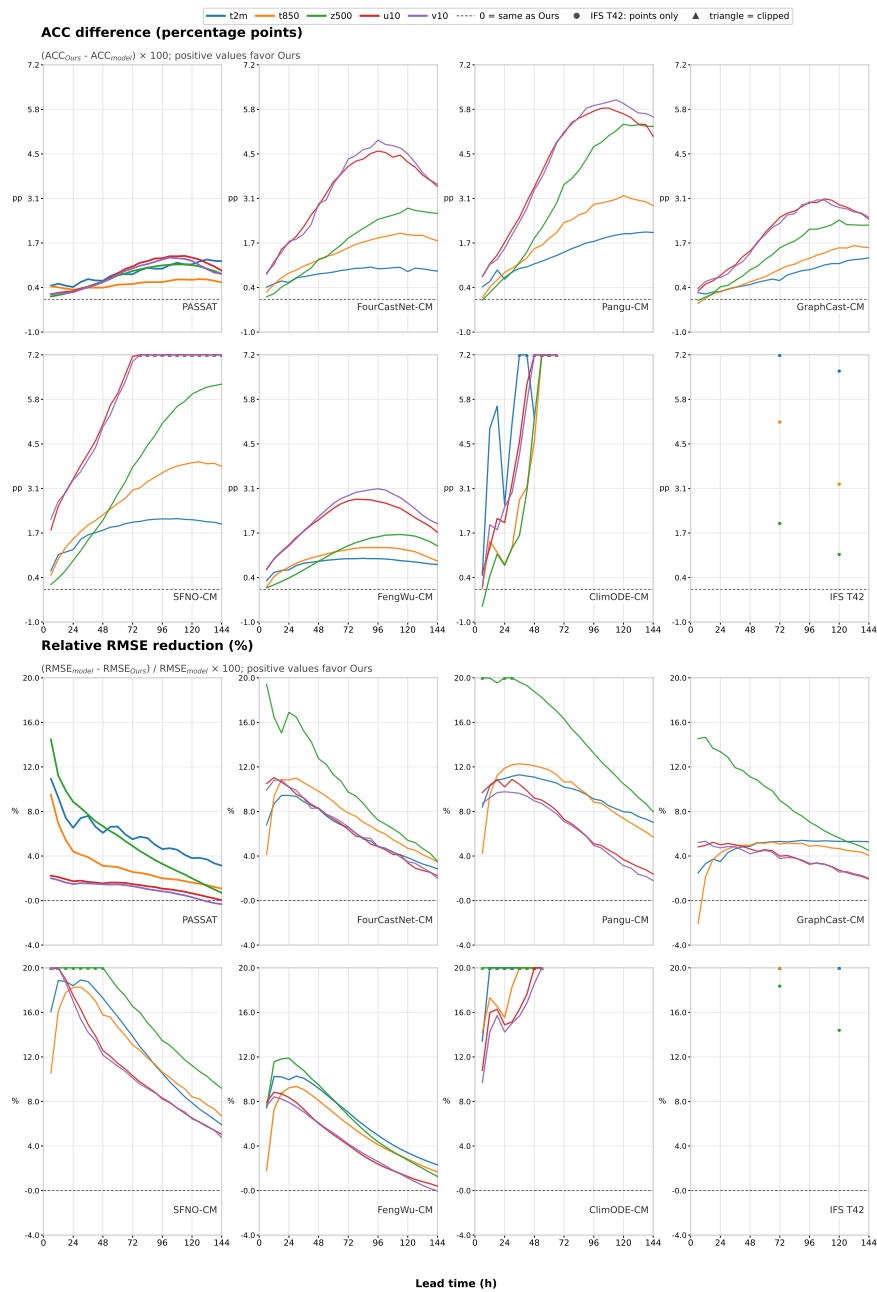


Figure 4. Forecast differences between the proposed model and the capacity-matched references over lead times up to 144 h. The upper panels report the ACC difference, $(ACC_{Ours} - ACC_{model}) \times 100$, in percentage points; this is an absolute ACC difference rather than a relative percentage. The lower panels report relative RMSE reduction, $(RMSE_{model} - RMSE_{Ours}) / RMSE_{model} \times 100$. Positive values favor the proposed model.



The improvements are most evident for temperature and geopotential variables. For 2-m temperature, the proposed method reduces RMSE from 1.25 to 1.17 at 24 h and from 2.69 to 2.60 at 144 h compared with PASSAT. For 850-hPa temperature, it improves RMSE from 1.29 to 1.24 at 24 h and from 3.39 to 3.35 at 144 h. For 500-hPa geopotential (Z500), the proposed model achieves a 144-h RMSE of $733 \text{ m}^2 \text{ s}^{-2}$, compared with $738 \text{ m}^2 \text{ s}^{-2}$ for PASSAT and $767 \text{ m}^2 \text{ s}^{-2}$ for GraphCast-CM. These results are consistent with the hypothesis that learning a bounded discrepancy inside the reduced coefficient equation can improve the forecast trajectory in this setting; they do not establish resolution-independent or operational-scale superiority.

Wind variables remain more challenging because near-surface winds contain stronger local variability and are more sensitive to unresolved processes. The consistent ACC improvement indicates better preservation of broad spatial organization, whereas the nearly neutral or negative 144-h RMSE changes indicate that wind amplitude calibration remains unresolved. Both observations are reported because ACC and RMSE measure different forecast properties.

4.5 Relative Gap Analysis

Fig. 4 further analyzes the performance difference between the proposed model and each capacity-matched reference over lead times up to 144 h. For ACC, we report the absolute difference $(\text{ACC}_{\text{Ours}} - \text{ACC}_{\text{model}}) \times 100$ in percentage points. For RMSE, we report the relative reduction $(\text{RMSE}_{\text{model}} - \text{RMSE}_{\text{Ours}}) / \text{RMSE}_{\text{model}} \times 100\%$. In both cases, positive values favor the proposed model.

The ACC differences show that the proposed method maintains higher spatial-pattern correlation than the capacity-matched PASSAT baseline across all five variables. The generally increasing gap is consistent with reduced autoregressive degradation under the evaluated protocol. Similar trends are observed against FourCastNet-CM, Pangu-CM, GraphCast-CM, SFNO-CM, and FengWu-CM, although the magnitude varies across variables and models.

The relative RMSE reductions provide complementary evidence. The proposed method achieves lower RMSE than PASSAT throughout most lead times, with particularly stable improvements for T2M and Z500. For stronger parameter-matched neural baselines, the gaps are smaller but remain mostly positive at medium and long lead times. These controlled differences are compatible with improved autoregressive behavior under a common parameter budget. Because the evaluation ends at 144 h and does not test perturbation growth or formal invariants, we avoid treating them as a proof of dynamical stability.

4.6 Analysis of the Bounded Transport-Residual Closure (BTRC)

The mathematical motivation for BTRC is given by Eq. (10): the idealized discrepancy is the difference between a fuller coefficient tendency and the retained advection–curvature operator. In the PASSAT-derived horizontal equation, plausible contributors include pressure-gradient, Coriolis, and frictional terms, together with vertical coupling, subgrid processes, proxy-variable error, and discretization mismatch. The network is not supervised with these components separately, so the learned field must be interpreted only as their aggregate effective remainder in normalized model space.

The experiments provide two levels of evidence. The comparison with PASSAT in Table 1 evaluates the entire reformulation, including the changed coefficient equation and the auxiliary encoder refinement. It should not be read as isolating the closure alone. The direct BTRC control is the “Full w/o BTRC” variant in Table 2, which retains the same coefficient pathway but

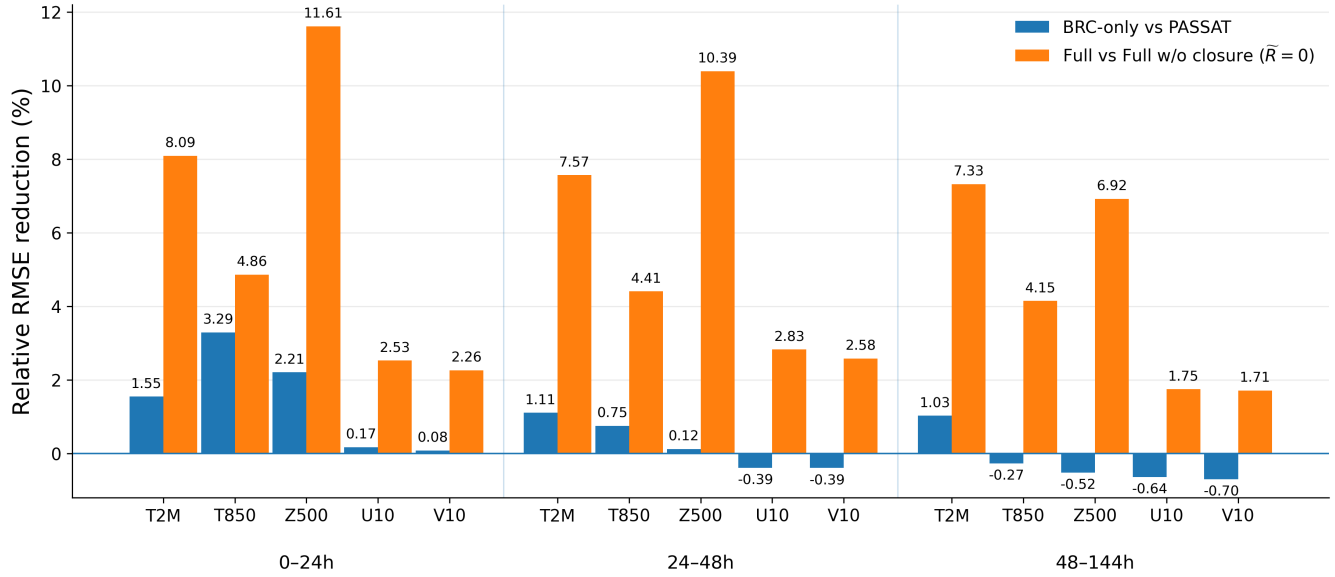


Figure 5. Interval-wise relative RMSE reductions for the BTRC-centered coefficient pathway. Blue bars compare BTRC w/o Node-WT with PASSAT and therefore reflect the broader coefficient-dynamics reformulation without Node-WT. Orange bars compare Full with the otherwise matched Full w/o BTRC control, $\tilde{R} \equiv 0$, providing more targeted evidence for the learned transport residual. Values are first computed as relative RMSE reductions at each forecast lead and then averaged over $(0, 24]$, $(24, 48]$, and $(48, 144]$ h, respectively. Positive values indicate lower RMSE for the first-named model, whereas negative values indicate degradation.

enforces $\tilde{R}_t \equiv 0$. At 144 h, removing the residual increases T2M RMSE from 2.605 to 2.785 and reduces ACC from 0.873 to 0.853; for Z500, RMSE increases from 733 to 771. These differences show that a nonzero learned remainder materially contributes to the performance of the otherwise matched solver pathway.

Fig. 5 further separates the broader coefficient-dynamics reformulation from the direct contribution of the learned closure. For a forecast-lead interval $\mathcal{I}$, the reported relative RMSE reduction is computed as

$$G_{A \leftarrow B}(\mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{\ell \in \mathcal{I}} \frac{\text{RMSE}_B(\ell) - \text{RMSE}_A(\ell)}{\text{RMSE}_B(\ell)} \times 100\%, \quad (32)$$

where A is the first-named model and B is the reference model. The intervals are $(0, 24]$, $(24, 48]$, and $(48, 144]$ h, corresponding respectively to the evaluated 6-h lead times within these ranges.

The blue bars compare BTRC w/o Node-WT with PASSAT. They show that the broader coefficient-dynamics reformulation produces positive short-range gains, particularly for T2M, T850, and Z500, but the gain decreases with lead time and becomes slightly negative for several variables in the medium- and long-range intervals. This comparison should therefore not be interpreted as isolating the learned closure.

The orange bars compare Full with the otherwise matched Full w/o BTRC control. These values remain positive for all five variables and all three intervals, providing more targeted evidence that the nonzero BTRC improves the coefficient pathway.



Table 2. Targeted component analysis in terms of RMSE↓/ACC↑. RMSE units are K for T2M and T850, m^2s^{-2} for Z500, and ms^{-1} for U10 and V10; ACC is dimensionless. “PASSAT + Node-WT” augments PASSAT with the auxiliary node-level wavelet branch. “BTRC w/o Node-WT” uses the BTRC-centered coefficient pathway without Node-WT. “Full w/o BTRC” retains the full pathway but sets $\tilde{R}_t \equiv 0$, providing the direct residual-free control. “Full” combines BTRC and Node-WT. The controls are designed to isolate the BTRC pathway; they are not presented as a complete factorial study of every architectural choice.

	Lead (h)	PASSAT	PASSAT + Node-WT	BTRC w/o Node-WT	Full w/o BTRC	Full
t2m	24	1.252 / 0.970	1.212 / 0.972	1.241 / 0.971	1.254 / 0.970	1.171 / 0.974
	48	1.544 / 0.955	1.511 / 0.957	1.529 / 0.956	1.563 / 0.954	1.451 / 0.960
	72	1.848 / 0.935	1.815 / 0.938	1.828 / 0.936	1.888 / 0.933	1.746 / 0.943
	96	2.158 / 0.911	2.127 / 0.914	2.137 / 0.913	2.220 / 0.907	2.058 / 0.921
	120	2.442 / 0.886	2.417 / 0.889	2.420 / 0.888	2.523 / 0.880	2.350 / 0.897
	144	2.689 / 0.862	2.667 / 0.864	2.668 / 0.864	2.785 / 0.853	2.605 / 0.873
t850	24	1.294 / 0.966	1.270 / 0.967	1.274 / 0.967	1.289 / 0.966	1.238 / 0.969
	48	1.662 / 0.943	1.654 / 0.943	1.655 / 0.943	1.684 / 0.942	1.610 / 0.947
	72	2.118 / 0.907	2.107 / 0.907	2.119 / 0.907	2.163 / 0.903	2.064 / 0.911
	96	2.594 / 0.859	2.574 / 0.859	2.604 / 0.858	2.656 / 0.853	2.542 / 0.864
	120	3.027 / 0.806	2.995 / 0.807	3.041 / 0.805	3.095 / 0.798	2.978 / 0.812
	144	3.387 / 0.756	3.346 / 0.756	3.406 / 0.754	3.467 / 0.745	3.351 / 0.761
z500	24	171 / 0.986	165 / 0.987	170 / 0.986	175 / 0.985	156 / 0.988
	48	293 / 0.958	289 / 0.960	293 / 0.957	304 / 0.955	274 / 0.963
	72	420 / 0.912	419 / 0.914	422 / 0.911	436 / 0.907	399 / 0.920
	96	543 / 0.849	546 / 0.852	546 / 0.848	564 / 0.843	525 / 0.860
	120	651 / 0.780	658 / 0.781	654 / 0.779	676 / 0.772	638 / 0.791
	144	738 / 0.713	751 / 0.712	745 / 0.710	771 / 0.702	733 / 0.721
u10	24	1.539 / 0.929	1.547 / 0.928	1.540 / 0.929	1.553 / 0.928	1.512 / 0.932
	48	2.221 / 0.848	2.229 / 0.847	2.234 / 0.846	2.250 / 0.843	2.187 / 0.853
	72	2.898 / 0.733	2.895 / 0.735	2.916 / 0.731	2.928 / 0.727	2.855 / 0.743
	96	3.463 / 0.611	3.463 / 0.614	3.485 / 0.608	3.490 / 0.603	3.426 / 0.623
	120	3.889 / 0.501	3.899 / 0.504	3.912 / 0.500	3.911 / 0.494	3.864 / 0.514
	144	4.194 / 0.415	4.223 / 0.413	4.227 / 0.411	4.222 / 0.404	4.192 / 0.423
v10	24	1.570 / 0.926	1.580 / 0.926	1.573 / 0.926	1.585 / 0.925	1.547 / 0.929
	48	2.259 / 0.842	2.263 / 0.843	2.271 / 0.841	2.285 / 0.839	2.226 / 0.848
	72	2.954 / 0.722	2.952 / 0.725	2.974 / 0.720	2.987 / 0.717	2.917 / 0.732
	96	3.552 / 0.589	3.552 / 0.593	3.579 / 0.587	3.588 / 0.583	3.523 / 0.602
	120	4.012 / 0.468	4.016 / 0.472	4.038 / 0.468	4.048 / 0.461	4.000 / 0.480
	144	4.341 / 0.372	4.357 / 0.372	4.374 / 0.370	4.387 / 0.361	4.355 / 0.380

The largest reductions occur for T2M and Z500, whereas the gains for U10 and V10 are smaller. Thus, the closure contribution is consistent across the displayed interval averages, although its magnitude remains variable-dependent.

The numerical interpretation is also limited. Equations (17)–(19) guarantee a bound on the direct residual forcing inserted into the coefficient update. The experimental curves indicate reduced error growth for selected variables, but neither result establishes a global stability theorem for the coupled nonlinear rollout. Accordingly, the paper claims a bounded closure and empirically improved autoregressive behavior, not guaranteed atmospheric stability.



4.7 Targeted Component Analysis

Table 2 is organized around the BTRC hypothesis. “PASSAT + Node-WT” measures the auxiliary representation change relative to PASSAT. “BTRC w/o Node-WT” evaluates the BTRC-centered coefficient pathway without Node-WT. “Full w/o BTRC” is the critical internal control because it preserves the full pathway while removing only the learned transport residual. The final column combines BTRC with the auxiliary wavelet refinement.

The results indicate that the wavelet branch alone provides small and nonuniform changes, including degradations for some wind and Z500 settings. It is therefore treated as an auxiliary feature-refinement mechanism rather than a stand-alone core contribution. The strongest controlled difference is between Full and Full w/o BTRC at longer lead times, especially for T2M and Z500. This pattern is consistent with the BTRC hypothesis: when the retained coefficient operator cannot represent the complete coarse-grid tendency, setting its equation discrepancy to zero produces a larger long-lead error.

The component table is not a complete factorial analysis of closure sharing, alternative bounding functions, or alternative multi-scale modules. Those questions remain outside the present experimental evidence and are stated explicitly as limitations.

4.8 Qualitative Analysis

Figs. 6–8 provide qualitative comparisons on three representative variables, namely T2M, Z500, and U10. These variables cover near-surface thermodynamic fields, large-scale circulation fields, and near-surface wind fields, respectively, offering complementary perspectives on forecast quality.

Fig. 6 shows the qualitative rollouts for 2-m temperature. Compared with PASSAT, the proposed rollout visually exhibits a more coherent large-scale temperature distribution and smaller apparent deviations from the ERA5 reference as lead time increases. This observation is consistent with the quantitative improvements in RMSE and ACC reported for T2M.

Fig. 7 shows the qualitative comparison for 500-hPa geopotential (Z500), expressed in m^2s^{-2} . Z500 is particularly informative for evaluating the preservation of large-scale circulation patterns. At longer lead times, the proposed rollout visually exhibits large-scale structures with smaller apparent deviations from the ERA5 reference than PASSAT, which is qualitatively consistent with reduced structural drift.

Fig. 8 shows qualitative rollouts for 10-m zonal wind. Compared with temperature and geopotential, near-surface wind fields contain stronger local variability and finer spatial structures, making them more challenging for autoregressive forecasting. The proposed rollout retains some visually sharper structures than PASSAT, but this qualitative impression should be interpreted together with the nearly neutral U10 RMSE and slightly worse V10 RMSE at 144 h. The images support a possible structural advantage, not a uniform reduction in wind amplitude error.

Overall, the qualitative results are most consistent with the quantitative gains for T2M and Z500. For wind, they mainly illustrate spatial structure and do not override the mixed RMSE evidence. The visualizations therefore complement, rather than replace, the metric-based closure analysis.

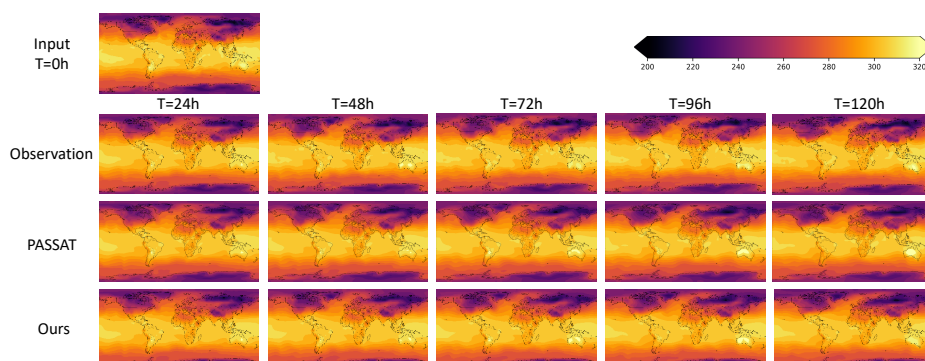


Figure 6. Qualitative rollouts for 2-m temperature (T2M) from 24 to 120 hours. Rows show the ERA5 reference, PASSAT, and the proposed method under the same evaluation setting.

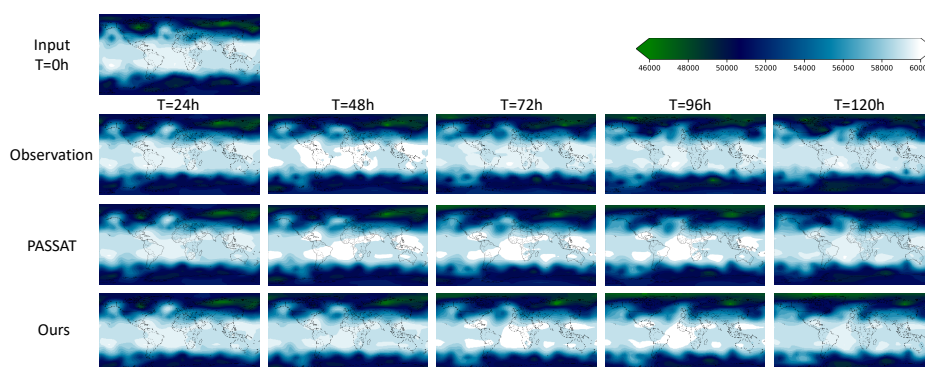


Figure 7. Qualitative rollouts for 500-hPa geopotential (Z500), in $\text{m}^2 \text{s}^{-2}$, from 24 to 120 hours. Rows show the ERA5 reference, PASSAT, and the proposed method under the same evaluation setting.

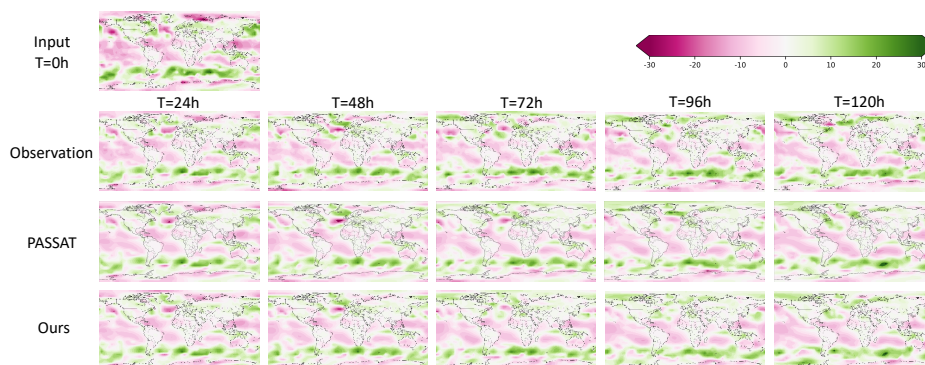


Figure 8. Qualitative rollouts for 10-m zonal wind (U10) from 24 to 120 hours. Rows show the ERA5 reference, PASSAT, and the proposed method under the same evaluation setting.



4.9 Experimental Summary

Overall, the experiments support a focused conclusion: within the evaluated coarse-resolution protocol, the BTRC pathway contributes substantially to T2M and Z500 improvements, while wind benefits are clearer in ACC than RMSE. The wavelet branch is secondary and its gains are not uniform.

4.10 Limitations and Scaling Considerations

This study has several limitations. First, the experiments use a coarse 5.625° grid and five prognostic variables, so the results establish only a low-resolution proof of concept and cannot be extrapolated to operational 0.25° systems. Second, the evaluation ends at 144 h and does not include perturbation-growth tests, conservation diagnostics, or formal invariant analysis; consequently, the bounded transport residual should not be conflated with guaranteed global stability. Third, the common two-

component closure is broadcast across variables. This low-rank design is compact, but it cannot represent all channel-specific pressure, frictional, moist, or boundary-layer effects. Fourth, the use of $\tanh$ is motivated by smooth boundedness and not by an empirical comparison with alternative saturation functions. Fifth, extreme events, seasonal subsets, precipitation, and explicitly unphysical states are not separately evaluated. Finally, the capacity-matched GraphCast-CM, Pangu-CM, and related variants are small reimplementations and are not comparisons with the official large-scale systems.

Scaling the method introduces additional numerical issues. With fixed graph degree, message passing grows approximately linearly with the number of grid nodes, but memory for activations and the cost of repeated Euler sub-steps also increase. Finer grids strengthen spatial gradients and can impose a more restrictive Courant–Friedrichs–Lewy condition, so the present $\delta t = 0.2$ h may not remain appropriate. The normalized closure scale τ , coefficient clipping range, and polar metric regularization are also resolution dependent. Adding vertical levels would further require channel- and level-aware closure structure rather than the present common directional residual. These considerations define the scope of future validation without presuming that the current hyperparameters transfer unchanged.

5 Conclusions

This paper reframed a PASSAT-style solver-coupled forecaster as a reduced coefficient equation augmented by a learned bounded transport-residual closure (BTRC). The retained operator contains spherical advection and curvature, while the residual represents the aggregate discrepancy associated with omitted or approximated pressure-gradient, Coriolis, frictional, vertical, subgrid, and discretization effects. The learned BTRC is inserted inside each coefficient sub-step and influences subsequent atmospheric state updates through the evolved transport trajectory, while $\tau \tanh(\cdot)$ supplies an explicit bound on the directly applied correction.

The existing experiments support this BTRC-centered interpretation. The residual-free control degrades long-lead T2M and Z500 forecasts, providing targeted evidence that the learned BTRC contributes beyond the retained coefficient operator. Wind results are mixed in RMSE despite improved ACC, and the node-level wavelet branch yields smaller, nonuniform gains; both



are reported as qualifications rather than hidden by an aggregate claim. The resulting contribution is therefore a BTRC-centered in-solver closure study for a coarse reduced atmospheric model, not a claim of operational-scale superiority or formally guaranteed dynamical stability.

515 *Code and data availability.* The exact version of the BTRC source code used to produce the results presented in this study is permanently archived on Zenodo at <https://doi.org/10.5281/zenodo.21870203> (Deng et al., 2026). The meteorological data used in this study were obtained from the WeatherBench dataset (Rasp et al., 2020). The WeatherBench dataset is publicly archived by the Technical University of Munich through the mediaTUM research data repository and is available at <https://dataserv.ub.tum.de/index.php/s/m1524895>. The experiments use five prognostic variables (T2M, T850, Z500, U10, and V10) at 5.625° resolution. The scripts and configurations used for data preprocessing, 520 variable selection, and construction of the training, validation, and test splits are included in the archived BTRC software release (Deng et al., 2026).

Author contributions. CD: conceptualization, methodology, software, validation, formal analysis, visualization, and writing – original draft. CL: supervision, conceptualization, project administration, and writing – review and editing. WL: software, investigation, validation, and visualization. ZW: data curation, software, and formal analysis. LZ: supervision, resources, and writing – review and editing. All authors 525 discussed the results and approved the final manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We would like to thank the computational and institutional support provided by Yibin University and Chongqing University of Technology.

Financial support. This work was supported by the Yibin City Science and Technology Project (grant no. 2025JC008), the Science and 530 Technology Department of Sichuan Province (grant no. 2024ZYD0089), and the Yibin University Science and Technology Projects (grant nos. 2024XJYY03 and 2025XJZD01).



References

- Allen, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W. P., Andersson, T. R., Herzog, M., Lane, N. D., Chantry, M., Hosking, J. S., et al.: End-to-end data-driven weather prediction, *Nature*, 641, 1172–1179, 2025.
- 535 Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., and Ott, E.: A machine learning-based global atmospheric forecast model, *Geophysical Research Letters*, 47, e2020GL087776, 2020.
- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.: Relational inductive biases, deep learning, and graph networks, *arXiv preprint arXiv:1806.01261*, 2, 5, 2018.
- Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, *Nature*, 525, 47–55, 2015.
- 540 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, 2023.
- Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., et al.: A foundation model for the Earth system, *Nature*, 641, 1180–1187, 2025.
- Bolton, T. and Zanna, L.: Applications of deep learning to ocean data inference and subgrid parameterization, *Journal of Advances in*
- 545 *Modeling Earth Systems*, 11, 376–399, 2019.
- Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., and Anandkumar, A.: Spherical fourier neural operators: Learning stable dynamics on the sphere, in: *International conference on machine learning*, pp. 2806–2823, PMLR, 2023.
- Brandstetter, J., Worrall, D., and Welling, M.: Message passing neural PDE solvers, 2022.
- Brenowitz, N. D. and Bretherton, C. S.: Prognostic validation of a neural network unified physics parameterization, *Geophysical Research*
- 550 *Letters*, 45, 6289–6298, 2018.
- Brenowitz, N. D. and Bretherton, C. S.: Spatially extended tests of a neural network parametrization trained by coarse-graining, *Journal of Advances in Modeling Earth Systems*, 11, 2728–2744, 2019.
- Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., et al.: Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead, *arXiv preprint arXiv:2304.02948*, 2023a.
- 555 Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H.: FuXi: A cascade machine learning forecasting system for 15-day global weather forecast, *npj climate and atmospheric science*, 6, 190, 2023b.
- Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K.: Neural ordinary differential equations, vol. 31, 2018.
- Cohen, T. S., Geiger, M., Köhler, J., and Welling, M.: Spherical cnns, 2018.
- Deng, C., Li, C., Luo, W., Wang, Z., and Zhu, L.: BTRC, <https://doi.org/10.5281/zenodo.21870203>, 2026.
- 560 Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, *Geoscientific Model Development*, 11, 3999–4009, 2018.
- Esteves, C., Allen-Blanchette, C., Makadia, A., and Daniilidis, K.: Learning so (3) equivariant representations with spherical cnns, in: *Proceedings of the european conference on computer vision (ECCV)*, pp. 52–68, 2018.
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E.: Neural message passing for quantum chemistry, in: *International*
- 565 *conference on machine learning*, pp. 1263–1272, Pmlr, 2017.
- Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., and Catanzaro, B.: Adaptive fourier neural operators: Efficient token mixers for transformers, 2021.



- Gupta, G., Xiao, X., and Bogdan, P.: Multiwavelet-based operator learning for differential equations, *Advances in neural information processing systems*, 34, 24 048–24 062, 2021.
- 570 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, *Quarterly journal of the royal meteorological society*, 146, 1999–2049, 2020.
- Hu, P., Wang, R., Zheng, X., Zhang, T., Feng, H., Feng, R., Wei, L., Wang, Y., Ma, Z.-M., and Wu, T.: Wavelet diffusion neural operator, 2025, 12 291–12 333, 2025.
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.: Physics-informed machine learning, *Nature Reviews*
- 575 *Physics*, 3, 422–440, 2021.
- Keisler, R.: Forecasting global weather with graph neural networks, *arXiv preprint arXiv:2202.07575*, 2022.
- Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al.: Neural general circulation models for weather and climate, *Nature*, 632, 1060–1066, 2024.
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.:
- 580 *Learning skillful medium-range global weather forecasting*, *Science*, 382, 1416–1421, 2023.
- Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C., Lessig, C., Maier-Gerber, M., Magnusson, L., et al.: AIFS–ECMWF’s data-driven forecasting system, *arXiv preprint arXiv:2406.01465*, 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*, 2020.
- 585 Mallat, S. G.: A theory for multiresolution signal decomposition: the wavelet representation, *IEEE transactions on pattern analysis and machine intelligence*, 11, 674–693, 1989.
- Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M. A., Ge, T., Subramaniam, A., et al.: Residual corrective diffusion modeling for km-scale atmospheric downscaling, *Communications Earth & Environment*, 6, 124, 2025.
- Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: Climax: A foundation model for weather and climate, 2023.
- 590 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, *arXiv preprint arXiv:2202.11214*, 2022.
- Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., and Battaglia, P. W.: Learning mesh-based simulation with graph networks, 2020.
- Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al.:
- 595 *Probabilistic weather forecasting with machine learning*, *Nature*, 637, 84–90, 2025.
- Raissi, M., Perdikaris, P., and Karniadakis, G. E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *Journal of Computational physics*, 378, 686–707, 2019.
- Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, *Proceedings of the national academy of sciences*, 115, 9684–9689, 2018.
- 600 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, 2020.
- Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., et al.: WeatherBench 2: A benchmark for the next generation of data-driven global weather models, *Journal of Advances in Modeling Earth Systems*, 16, e2023MS004 019, 2024.



- 605 Scher, S.: Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning, *Geophysical Research Letters*, 45, 12–616, 2018.
- Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 379, 20200097, 2021.
- 610 Tripura, T. and Chakraborty, S.: Wavelet neural operator: a neural operator for parametric partial differential equations, *arXiv preprint arXiv:2205.02191*, 2022.
- Verma, Y., Heinonen, M., and Garg, V.: Climode: Climate and weather forecasting with physics-informed neural odes, 2024, 8408–8430, 2024.
- Weyn, J. A., Durran, D. R., and Caruana, R.: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere, *Journal of Advances in Modeling Earth Systems*, 12, e2020MS002109, 2020.
- 615 Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V.: Integrating scientific knowledge with machine learning for engineering and environmental systems, *ACM Computing Surveys*, 55, 1–37, 2022.
- Zheng, J., Ling, Q., and Feng, Y.: Physics-assisted and topology-informed deep learning for weather prediction, *arXiv preprint arXiv:2505.04918*, 2025.