PHYSICS-CONSTRAINED NEURAL NETWORKS FOR IMPROVED SHORT-TERM WEATHER FORECASTING: A CASE STUDY OVER THE SOUTH PACIFIC

Egor Bugaev, Fedor Buzaev, Dmitry Efremenko, Denis Derkach & Fedor Ratnikov
Faculty of Computer Science
Higher School of Economics
Moscow, Russia
{ebugaev,fa.buzaev}@hse.ru

ABSTRACT

This study introduces enhancements to physics-constrained neural networks (PC-NNs) that improve the accuracy and stability of hybrid short-term weather forecasting models. Building on the WeatherGFT architecture, three innovations are proposed. First, an upgraded numerical solver—combining a fifth-order weighted essentially non-oscillatory scheme (WENO-5), a β-plane approximation, and subgrid-scale viscosity—permits a fourfold increase in the integration time step (to 1200 s) while reducing the daily mean squared error by up to 26%. Second, a unified autoregressive hybrid block replaces the original chain of 24 specialised modules, eliminating overfitting to specific lead times; and third, the physical core is integrated with two state-of-the-art neural backbones, resulting in PI-PredFormer and PI-IAM4VP. Evaluation on the WeatherBench South Pacific subset (2000–2004) shows that these hybrids reduce root mean squared error at 1–12 h lead times by 8–22% compared to purely neural counterparts, while better preserving physical consistency. These results demonstrate that incremental refinement of hybrid components offers a practical route toward more accurate and efficient short-range weather forecasting.

1 INTRODUCTION

Accurate short-term (1–24 h) weather forecasts are critical for safeguarding lives and infrastructure in sectors including aviation, maritime transport, and emergency response (Molteni et al., 1996; Vitart & Takaya, 2021; Han et al., 2022). While Numerical Weather Prediction (NWP) models generate physically consistent forecasts by integrating governing dynamical equations, they suffer from high computational costs and rely on empirical parameterizations for sub-grid-scale processes (Zängl, 2023). In contrast, pure machine learning (ML) models learn statistical patterns from vast datasets, offering forecasts orders of magnitude faster but often violating fundamental physical laws (Grazzini et al., 2024; Vahid Yousefnia et al., 2024).

A promising path forward lies in hybrid physics-constrained neural networks (PCNNs), which embed the residuals of governing PDEs directly into the learning objective. A seminal architecture in this domain is WeatherGFT (Xu et al., 2025), which interleaves Transformer blocks with a partial differential equation (PDE) solver. In parallel, highly efficient neural backbones have emerged for spatio-temporal prediction. Notably, PredFormer (Tang et al., 2024) and IAM4VP (Seo et al., 2023) achieve state-of-the-art accuracy with lightweight architectures, making them ideal for real-time applications. However, these models lack built-in physical guarantees, which can lead to unrealistic states.

In this study, we advance the hybrid modeling paradigm by combining the physical rigor of physics-constrained frameworks with the expressiveness of recent spatiotemporal architectures. Rather than proposing a wholly new architecture, we systematically enhance WeatherGFT and transfer its physical core to powerful, lightweight neural backbones. Our specific contributions are:

1. **Enhanced solver.** Central-difference schemes are replaced by a fifth-order weighted essentially non-oscillatory scheme WENO-5 (Stevens & Colonius, 2020) with a β-plane Coriolis term and sub-grid viscosity, increasing the integration step from 300 s to 1200 s while reducing 24-h mean squared error by up to 26 %.
2. **Universal hybrid block.** The WeatherGFT original chain of 24 specialized modules is collapsed into a *single, hourly PCNN block* with a learnable weight γ that adaptively balances data-driven and physics-based updates, eliminating over-fitting to pre-set lead times.
3. **Integration pathways.** Two new hybrids are developed:
   - Physics-informed PredFormer (PI-PredFormer), which applies post-hoc physics correction to the temporal Transformer PredFormer outputs every 15 min;
   - Physics-informed IAM4VP (PI-IAM4VP), which incorporates the PCNN block into IAM4VP with physics corrections after each hourly forecast.

Evaluation on the WeatherBench South Pacific subset (2000–2004) shows that these hybrids yield measurable improvements in short-range forecast skill while maintaining physical consistency. The South Pacific region was chosen to ensure a consistent meteorological setting and to enable computationally efficient experimentation.

2 Neural Network Backbones for Spatiotemporal Forecasting

Recent advances in data-driven weather forecasting have in part relied on neural architectures adapted from computer vision, which treat atmospheric states as videos where spatial dimensions (H, W) and variables (channels C) evolve over time. These models learn spatiotemporal correlations directly from data but often lack inherent physical constraints. In this study, we focus on two state-of-the-art computationally efficient architectures, namely, PredFormer and IAM4VP, that serve as foundations for our hybrid models, alongside the established physics-informed baseline WeatherGFT.

PredFormer adapts the Transformer architecture for video prediction by factorizing the computationally expensive self-attention mechanism. Rather than applying attention jointly over all spatiotemporal tokens, it employs two branches: one attends to the temporal evolution of each spatial location, while the other captures spatial interactions within each time step. This factorization substantially reduces computational cost while effectively representing atmospheric patterns. Its compact architecture and high accuracy make it a suitable candidate for enhancement with physical constraints.

IAM4VP (Seo et al., 2023) addresses the trade-off between autoregressive models (prone to error accumulation) and non-autoregressive models (which may lack temporal coherence). It employs a semi-autoregressive strategy, encoding a full sequence of observed frames but decoding forecasts one step at a time. A key innovation is its dual-queue mechanism, which blends ground-truth inputs with autoregressive predictions, thereby anchoring the model to real data at longer horizons. Built with modern ConvNeXt blocks, IAM4VP achieves strong accuracy with only 12M parameters, making it highly computationally efficient.

A Physics-Informed Baseline WeatherGFT (Xu et al., 2025) represents a pioneering hybrid approach, interleaving Transformer blocks with a PDE solver that integrates governing atmospheric equations over 15-minute intervals. This design enforces physical consistency by construction but originally employed a chain of 24 specialized modules, resulting in uneven gradient updates across lead times and overfitting at specific horizons. In addition, its solver was restricted by a stringent Courant–Friedrichs–Lewy (CFL) condition, reducing computational efficiency.

Building on these foundations, we combine the efficient spatiotemporal learning of PredFormer and IAM4VP with the physical consistency principle of WeatherGFT. The next sections describe enhancements to the WeatherGFT framework and the integration of its improved physical core into these neural backbones.

3 Training Data and Experimental Setup

In this section we describe a procedure of neural network training. All experiments are conducted using the WeatherBench dataset (Rasp et al., 2020), which provides reanalysis-derived atmospheric

fields on a regular grid. To focus on a meteorologically coherent domain, the South Pacific subset is extracted, covering the region from 5.625°N–39.375°S and 174.375°E–95.625°W. The dataset is restricted to the years 2000–2004. Data from 2000–2003 are used for training, and the year 2004 is reserved for independent evaluation.

The variables considered are geopotential height, temperature, the zonal (u) and meridional (v) wind components, and specific humidity at levels 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000 hPa, following the standard WeatherBench benchmarks. All variables are taken with the angular resolution of $1.40525°$ making a 128×256 latitude–longitude grid and sampled at 60 min temporal resolution.

More details on model parameters can be found in Appendix A.3

4 Effects of Physical Constraints on Neural Weather Prediction Architectures

This section presents a systematic analysis of how the architecture and placement of hybrid modules, where physical constraints are integrated with neural outputs, affect the performance of neural weather forecasting models. Our aim is twofold: (i) to assess which configurations of the hybrid unit yield consistent improvements under different design choices, and (ii) to provide guidelines for avoiding overfitting and error accumulation when integrating physical priors into data-driven models.

4.1 Transition to a Single Hybrid Block Architecture

The original WeatherGFT employs 24 sequential hybrid blocks, each responsible for a 15-min forecast interval. Because training is supervised only at fixed horizons (1, 3, and 6 h), blocks operating at longer lead times receive fewer gradient updates than those at shorter ranges. This imbalance produces overfitting near the trained horizons and degraded performance at intermediate steps (e.g. 2–5 h), as can be seen on Figure 1.

To address this, a simplified configuration was developed in which a *single* hybrid block is applied autoregressively to generate forecasts of arbitrary length. Details on the difference from the original architecture are presented in the end of Appendix A.2. This ensures that every forecast step contributes equally to parameter updates and removes discontinuities caused by disjoint training coverage across blocks. The consolidated block contains six PDE kernels and two neural modules, arranged to simulate a 60-min forecast via two successive 30-min steps. Its parameter count is approximately twice that of a single 15-min block, providing sufficient capacity while remaining an order of magnitude smaller than the full 24-block system.

The single-block design eliminates horizon-specific overfitting and substantially reduces parameter count, while still delivering accuracy comparable to the full 24-block chain at longer lead times. At the same time, the original WeatherGFT retains an advantage in some early forecast hours, where its larger capacity yields slightly better short-range skill. The new approach offers greater efficiency and more consistent accuracy across horizons, though at the expense of reduced performance in the short range. These trade-offs motivate the exploration of architectures that can exploit multiple temporal inputs, as discussed in the following subsection.

4.2 Numerical Integration Enhancements

Forecast errors in the original WeatherGFT integration arise primarily from the discretization scheme rather than deficiencies in the underlying physical equations. Although the five-point centred stencil is formally fourth-order accurate, it behaves poorly on non-smooth fields, producing dispersive oscillations near grid boundaries. These appear as artificial ripples and distorted moisture or wind fronts, which accumulate over time and degrade forecast quality.

To address these shortcomings and improve stability, we introduced several targeted kernel upgrades that together allow a fourfold increase in the integration time step without compromising accuracy:

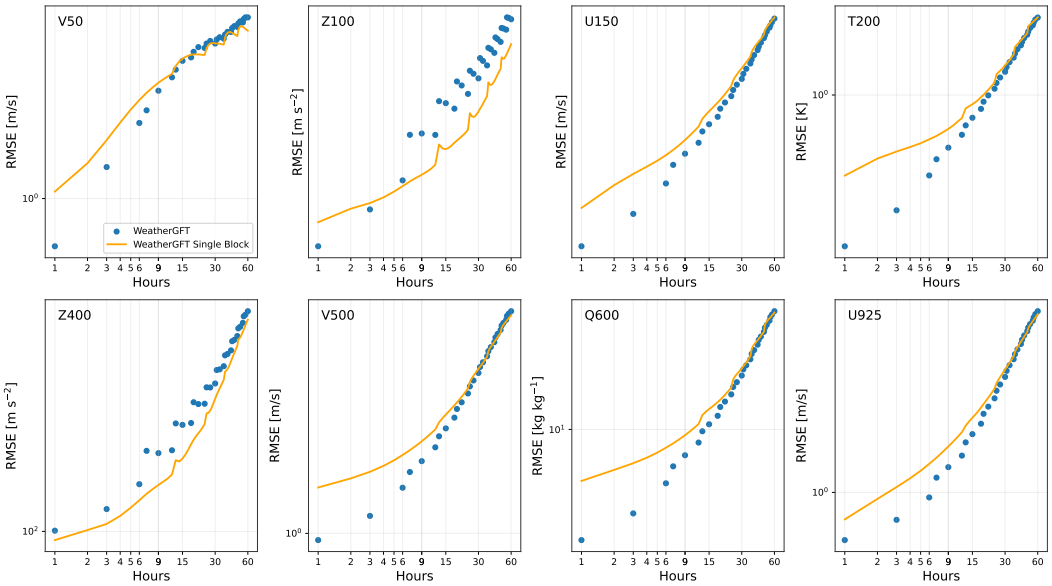

Figure 1: Comparison between the original WEATHERGFT and the improved generalization of the single-block variant. Single-block variation achieves similar or better quality on long-term predictions despite having 10 times less parameters. WeatherGFT results are plotted only at the trained horizons (1, 3, and 6 h of each forecast cycle, i.e. 1, 3, 6, 7, 9, 12 h, etc.), since predictions at intermediate steps (e.g. 2, 4, 5 h) degrade severely due to overfitting and fall outside the plotted range.

WENO-5 for horizontal derivatives. Central differences are replaced with fifth-order *Weighted Essentially Non-Oscillatory* (WENO) derivatives for each prognostic variable $\psi \in \{u, v, T, q, Z\}$. Cell-face values are reconstructed using nonlinear combinations of three parabolas, with weights adapting to local smoothness to suppress oscillations near discontinuities. This reduces spurious wave amplitudes by a factor of three and narrows transition zones from five to three grid points.

Latitude-dependent (β-plane) Coriolis term. The constant Coriolis parameter f is replaced by a latitude-dependent expression

$$f(\varphi) = f_0 + \beta R \varphi, \quad \varphi \in \left[-\tfrac{\pi}{2}, \tfrac{\pi}{2}\right],$$

where R is Earth's radius. This restores meridional variability in Coriolis forcing and reduces artificial ringing near the equator.

Subgrid-scale eddy viscosity. A diffusion term $\nu_e \nabla^2 \mathbf{u}$ is added to damp short-wavelength noise. The Laplacian is computed using a second WENO pass. With $\nu_e \approx 10^4$ m^2 s^{-1}, spectral energy above the Nyquist half-wavenumber is reduced by about 10 dB per model hour, eliminating 2–3 gridpoint artefacts and improving stability at long horizons.

Increased time step. The original model was limited to $\Delta t = 300$ s by CFL constraints. With WENO-5 and eddy damping, stability analysis indicated $\Delta t_{\max} \approx 1400$ s for the 32×64 grid. A step of $\Delta t = 1200$ s was adopted, reducing GPU time by more than 70% due to fewer kernel calls.

Upgrading the discretization scheme with WENO-5 derivatives, a latitude-sensitive Coriolis term, subgrid viscosity, and an increased time step stabilises the integration and improves forecast accuracy across all prognostic variables.

Figure 2 shows RMSE as a function of lead time for geopotential Z, temperature T, specific humidity Q, zonal wind U, and total wind speed V. In all cases, the new integrator (dashed lines) slows error growth compared with the previous scheme (solid lines). For example, Z RMSE remains below 0.2 up to 6 h, compared with 3 h for the baseline, while temperature retains low errors out to nearly 12 h. Humidity benefits most at short leads, with an initial RMSE reduction of almost 30% at 1 h. Wind forecasts also show marked improvements, with oscillatory artefacts suppressed and

mid-range errors reduced. Overall, the new integration scheme yields consistently more accurate and stable forecasts.

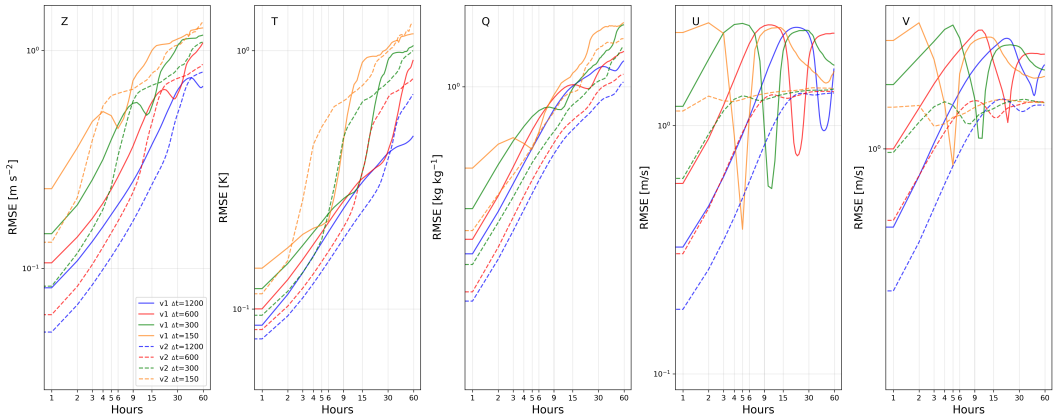

Figure 2: Forecast error growth as a function of lead time for geopotential (Z), temperature (T), specific humidity (Q), zonal wind (U), and total wind speed (V). Solid lines denote the original integrator; dashed lines denote the upgraded scheme. Across all prognostic variables, the enhanced scheme produces slower error growth, suppresses spurious oscillations, and extends the range of low-error forecasts.

4.3 Multi-Input–Multi-Output Architecture: PI-PredFormer

The transition to a single hybrid block also enables the use of more expressive neural architectures, capable of exploiting not only the current atmospheric state but also recent temporal trends. To investigate this potential, we integrate the physics-informed module from WeatherGFT into the PredFormer architecture. PredFormer is designed for spatiotemporal forecasting, using two distinct attention mechanisms: one branch attends along the temporal axis, capturing the evolution of atmospheric states, while the other attends across locations within each time slice, representing spatial dependencies. This dual-branch design allows the network to combine information on both temporal progression and spatial structure, features that are central to weather prediction.

The resulting hybrid, termed PI-PredFormer, employs a two-stage prediction process (Figure 3). In the first stage, PredFormer generates a 12-hour forecast $X' \in \mathbb{R}^{T\times C\times H\times W}$ from a 12-hour input X. In the second stage, independent physics-informed hybrid blocks are applied to X', enforcing dynamical and thermodynamical consistency. Finally, a decoder module merges the neural and physics-informed outputs to produce the forecast Y.

Because the hybrid blocks refine forecasts at each step without relying on their own past outputs, this design avoids the autoregressive error accumulation that typically affects iterative models. At the same time, physical consistency is explicitly reinforced by the PDE solver. To accommodate 1-hour forecast intervals, the hybrid block was redesigned to consist of three 20-minute kernels, which proved more effective than a single long step (see Figure 2).

Despite these advantages, one limitation of PI-PredFormer is that the hybrid blocks are applied independently at each forecast step. This means that their influence diminishes at longer horizons: once the neural predictions begin to drift from physically plausible states, post-hoc corrections cannot fully restore consistency. As a result, the forecasts tend to be sharper but can occasionally appear less realistic.

Figure 5 includes comparison of the performance of PredFormer and PI-PredFormer. The hybrid achieves clear improvements at the 1-hour lead, demonstrating the benefit of applying physical corrections to forecasts that are still anchored to observations. At longer horizons, however, error growth accelerates, highlighting a fundamental challenge in multi-step forecasting: early inaccuracies propagate nonlinearly through both the neural and physical branches, and post-hoc corrections

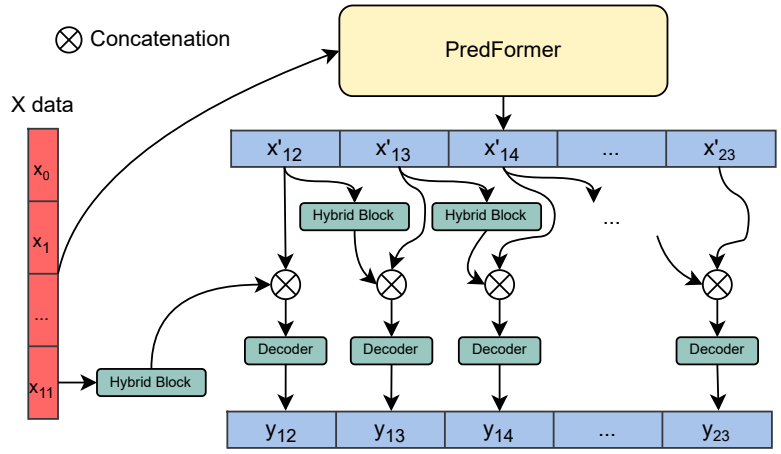

Figure 3: Schematic of the PI-PredFormer architecture. PredFormer first processes the input sequence using temporal and spatial attention branches. Physics-informed hybrid blocks then refine these forecasts, and a decoder adaptively combines neural and physical outputs to produce the final prediction.

cannot completely compensate. This finding underscores the need for tighter coupling between physical and neural components in future hybrid designs.

4.4 Semi-Autoregressive Architecture: PI-IAM4VP

The previous sections demonstrate both the benefits and the limitations of physics-informed modules. Pure autoregressive models suffer from error accumulation, while post-hoc enforcement of physical constraints cannot retroactively correct earlier inaccuracies. This reflects the classical trade-off between Multiple-Input–Multiple-Output (MIMO) strategies, which predict entire sequences at once, and Single-Input–Single-Output (SISO) strategies, which extend forecasts step by step.

To address these challenges, we adapt the IAM4VP architecture, originally developed for video prediction, to the weather forecasting problem. The model separates the input into two streams: historical observations $X \in \mathbb{R}^{T\times C\times H\times W}$ and autoregressive predictions $Y \in \mathbb{R}^{t\times C\times H\times W}$ for the interval between the last observation and the target forecast time. For example, to generate a 5-hour forecast, the model ingests 12 hours of reanalysis data and then uses its own predictions for the subsequent 4 hours before producing the 5-hour output. Intermediate steps are generated in separate forward passes. This design balances the stability of direct prediction from historical states with the flexibility of autoregressive extensions.

The resulting hybrid, PI-IAM4VP, is shown in Figure 4. Its main components are:

- **Spatial Encoder.** A stack of convolutional layers interleaved with axis shifts extracts latent features from X. Four levels of representation are obtained, resulting in $X' \in \mathbb{R}^{4\times T\times C'\times H\times W}$.
- **Hybrid Block.** Three 20-min PDE kernels from WeatherGFT (without neural components) are applied, together simulating a 1-hour forecast step. To accommodate the patch-based nature of the kernels, inputs are divided into 4×4 subgrids.
- **Learned Prior and LP Physics.** These modules refine neural and physics-informed features before fusion. They mirror the encoder structure but use separate parameters and are shared across time steps. Following the convention in WeatherGFT, the physical component is weighted by 0.1 in the final combination.
- **Predictor.** A sequence of ConvNeXt (Woo et al., 2023) blocks generates next-frame predictions from the refined features.

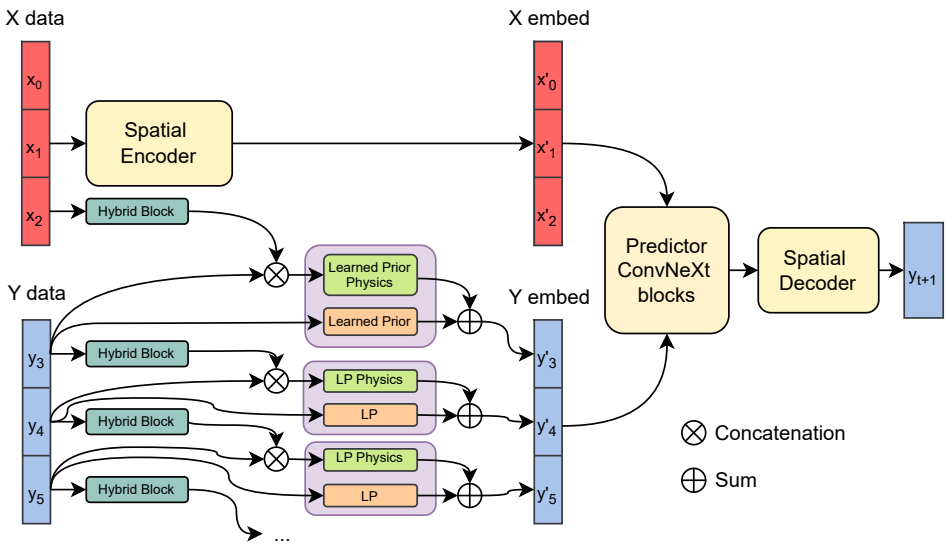

Figure 4: Schematic of the PI-IAM4VP architecture. The input is split into a historical stream (X) and an autoregressive stream (Y). Both streams are processed by hybrid blocks and combined in the decoder to produce the target forecast.

- **Decoder.** This combines outputs from the predictor with latent features from both X and Y. The decoding proceeds in four stages, each aligned with one of the encoder's latent levels.

Evaluation results in Table 2 demonstrate that embedding physics constraints within this semi-autoregressive framework yields consistent improvements across variables. The hybrid block enhances accuracy while maintaining stability at longer horizons. An interesting feature is the sharp reduction in error between the 1-hour and 2-hour forecasts. This may reflect the hybrid block's convolutional operations being better tuned to the smoother synthetic fields generated by the neural module than to the noisier real initial states. Although duplicating the hybrid block for each forecast hour could potentially mitigate this effect, it would substantially increase model complexity. In this work we therefore retain a single shared block, striking a balance between accuracy and computational efficiency.

5 Model Evaluation and Comparison

Having introduced the model components, we now evaluate their forecast performance. Detailed parameters of architectures presented in this study are described in Appendix A.3.

Forecast quality is assessed at prediction horizons of 1, 3, 6, 12, 24, and 60 hours. Results for short-term range are presented in Table 1 and for mid-term range in Table 2.

For geopotential height and temperature, the best-performing models are those with embedded physics modules (PI-PredFormer, WeatherGFT, and PI-IAM4VP), underscoring the value of PDE constraints at very short leads. For wind and humidity, however, the underlying neural architecture proves more decisive: IAM4VP achieves the strongest scores.

At intermediate horizons, PI-IAM4VP delivers the lowest mean errors across most variables, followed by its baseline IAM4VP and by PredFormer. This indicates that physics-informed corrections remain effective up to about half a day, especially when coupled with a semi-autoregressive design.

Table 1: Root Mean Squared Error (RMSE) of forecasts at 1-hour and 3-hour horizons for all evaluated models. Non-bold entries indicate percentage increase in RMSE relative to the best model in the same column.

Model	**z150 (m s^{-2})**		**t150 (K)**		**u500 (m/s)**		**q1000 (kg kg^{-1})**		**Avg. scaled**	
	1 h	**3 h**	**1 h**	**3 h**	**1 h**	**3 h**	**1 h**	**3 h**	**1 h**	**3 h**
WeatherGFT	+123%	+175%	**0.55**	+12%	+24%	+17%	+33%	+13%	+20%	+17%
WeatherGFT Single Block	+125%	+168%	+38%	+52%	+71%	+66%	+70%	+44%	+60%	+50%
PredFormer	+26%	+43%	+20%	+21%	+68%	+43%	+89%	+46%	+60%	+42%
PI-PredFormer	**81.9**	+63%	**0.55**	+25%	+53%	+45%	+85%	+48%	+50%	+42%
IAM4VP	+47%	+66%	+2%	+5%	**1.18**	+3%	**2.30**	**3.02**	**0.10**	+8%
PI-IAM4VP	+40%	**76.1**	+4%	**0.56**	+2%	**1.43**	**2.30**	+1%	**0.10**	**0.12**
PredRNN	+152%	+179%	+60%	+64%	+94%	+78%	+61%	+44%	+70%	+58%
SimVP	+138%	+155%	+155%	+150%	+145%	+103%	+203%	+131%	+140%	+108%

Table 2: The same as in Table 1, but for 6-hour and 12-hour horizons

Model	**z150 (m s^{-2})**		**t150 (K)**		**u500 (m/s)**		**q1000 (kg kg^{-1})**		**Avg. scaled**	
	6 h	**12 h**	**6 h**	**12 h**	**6 h**	**12 h**	**6 h**	**12 h**	**6 h**	**12 h**
WeatherGFT	+141%	+97%	+11%	+7%	+12%	+8%	+2%	**4.69**	+13%	+10%
WeatherGFT Single Block	+135%	+79%	+41%	+26%	+47%	+29%	+26%	+14%	+40%	+20%
PredFormer	+21%	**142.51**	+11%	**0.86**	+21%	+5%	+20%	+3%	+20%	+5%
PI-PredFormer	+38%	+11%	+15%	+2%	+23%	+8%	+21%	+4%	+27%	+10%
IAM4VP	+42%	+23%	+5%	+2%	+3%	+3%	+0%	**4.69**	+7%	+5%
PI-IAM4VP	**97.17**	+3%	**0.66**	**0.86**	**1.80**	**2.41**	**3.79**	**4.69**	**0.15**	**0.20**
PredRNN	+125%	+67%	+52%	+33%	+55%	+35%	+31%	+18%	+47%	+25%
SimVP	+101%	+40%	+114%	+66%	+64%	+30%	+84%	+49%	+67%	+30%

At longer horizons, purely data-driven models (notably PredFormer) take the lead. Here, accumulated forecast errors outweigh the corrective effect of the simplified physics module. Nonetheless, PI-IAM4VP consistently outperforms its baseline, showing that physical constraints still provide benefits relative to the same neural architecture without them. In contrast, PI-PredFormer lags behind plain PredFormer at these leads, reflecting the limitations of post-hoc corrections when neural states have already drifted.

Overall, PI-IAM4VP delivers the lowest mean error across all evaluated variables up to the 12-hour horizon, demonstrating that carefully integrated physics-informed modules yield strong improvements in short- to mid-range forecasts. For longer horizons (24 and 60 hours), plain neural architectures, e.g. PredFormer, outperform their physics-augmented counterparts, likely due to error propagation overwhelming the benefits of embedded physical constraints. These performance trends are summarized in Figure 5.

The evaluation highlights distinct behaviors of the two proposed hybrid architectures. PI-PredFormer performs best at very short lead times (1 h), where physical corrections are applied directly to observation-based states. This indicates that post-hoc enforcement of physical laws can stabilise forecasts when the neural predictions remain close to reality. However, the same mechanism becomes less effective at longer horizons: once neural trajectories diverge from physically

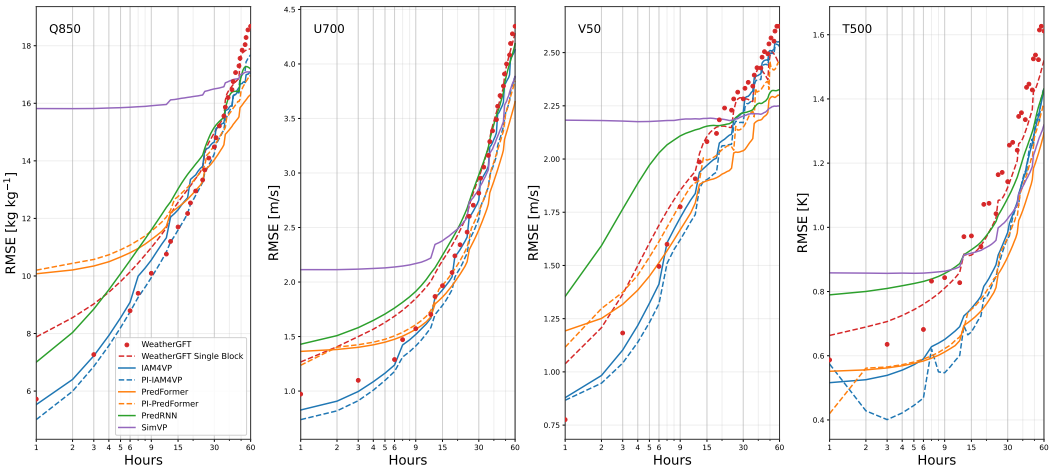

Figure 5: Comparative evaluation of RMSE at the Q850, U700, V50, and T500 levels for various models: WeatherGFT (red dots), WeatherGFT Single Block (red dashes), IAM4VP (blue), PI-IAM4VP (blue dashed), PredFormer (orange), PI-PredFormer (orange dashed), PredRNN (green), and SimVP (purple).

plausible states, subsequent corrections cannot fully restore consistency. This leads to faster error growth beyond 6–12 h, despite initial improvements.

PI-IAM4VP, by contrast, shows more balanced performance across variables and horizons. Its semi-autoregressive design allows the model to incorporate both historical observations and its own predictions, while the embedded hybrid block enforces dynamical constraints at each iteration. This architecture consistently improves upon its baseline up to 12 h and maintains stability at longer ranges. The results suggest that tighter coupling between physical modules and iterative neural forecasting is advantageous for medium-range skill.

6 Conclusion

This study advances the integration of physics-informed components into neural architectures for short-range weather prediction. Building on the hybrid WeatherGFT framework, we proposed and evaluated three enhanced models that address key limitations in computational efficiency and generalisation:

- **WeatherGFT-Single** condenses the original sequence of 24 specialised modules into a single autoregressive hybrid block. This unified design eliminates overfitting to pre-defined lead times and reduces error discontinuities across horizons, while maintaining competitive accuracy with substantially fewer parameters.
- **PI-PredFormer** integrates the factorised spatiotemporal attention of PredFormer with post-hoc physics correction layers derived from WeatherGFT. It achieves strongest performance at very short lead times (1–3 h), where physical constraints act directly on states still anchored in observations.
- **PI-IAM4VP** incorporates physical simulation into the semi-autoregressive framework of IAM4VP. This model delivers consistent improvements across variables and lead times, demonstrating the value of embedding physics within iterative neural forecasting pipelines.

Quantitative evaluation across 1–60 h forecasts shows that physics-informed models enhance accuracy in the short-to-medium range (1–12 h), particularly for geopotential height and temperature.

These findings demonstrate that carefully designed physics-informed neural networks improves accuracy, stability, and physical realism in the operationally critical 1–12 h window. Challenges remain: the PDE cores omit essential sub-grid processes; semi-autoregressive designs accumulate

errors over long horizons; and post-hoc corrections cannot fully compensate for neural drift. Future work should explore tighter coupling of neural and physical components, including joint training of PDE solvers and neural backbones, adaptive weighting schemes responsive to forecast regimes, and incorporation of parameterised sub-grid physics. Extending evaluation to larger domains with additional forcings such as radiation and surface fluxes will be essential to bridge the gap towards robust, operational-scale hybrid forecasting.

ACKNOWLEDGMENTS

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University 139-15-2025-009.

This research is supported in part through computational resources of HPC facilities at HSE University.

REFERENCES

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3170–3180, June 2022.

Federico Grazzini, Joshua Dorrington, Christian M. Grams, George C. Craig, Linus Magnusson, and Frederic Vitart. Improving forecasts of precipitation extremes over northern and central Italy using machine learning. *Quarterly Journal of the Royal Meteorological Society*, 150(762):3167–3181, May 2024. ISSN 1477-870X. doi: 10.1002/qj.4755. URL http://dx.doi.org/10.1002/qj.4755.

Yan Han, Lihua Mi, Lian Shen, C.S. Cai, Yuchen Liu, Kai Li, and Guoji Xu. A short-term wind speed prediction method utilizing novel hybrid deep learning algorithms to correct numerical weather forecasting. *Applied Energy*, 312:118777, April 2022. ISSN 0306-2619. doi: 10.1016/j.apenergy.2022.118777. URL http://dx.doi.org/10.1016/j.apenergy.2022.118777.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

F Molteni, R Buizza, T N Palmer, and T Petroliagis. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, 122(529):73–119, January 1996.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), November 2020. ISSN 1942-2466. doi: 10.1029/2020ms002203. URL http://dx.doi.org/10.1029/2020MS002203.

Minseok Seo, Hakjin Lee, Doyi Kim, and Junghoon Seo. Implicit stacked autoregressive model for video prediction, 2023. URL https://arxiv.org/abs/2303.07849.

Ben Stevens and Tim Colonius. Enhancement of shock-capturing methods via machine learning. *Theoretical and Computational Fluid Dynamics*, 34(4):483–496, May 2020. ISSN 1432-2250. doi: 10.1007/s00162-020-00531-1. URL http://dx.doi.org/10.1007/s00162-020-00531-1.

Yujin Tang, Lu Qi, Fei Xie, Xiangtai Li, Chao Ma, and Ming-Hsuan Yang. Predformer: Transformers are effective spatial-temporal predictive learners, 2024. URL https://arxiv.org/abs/2410.04733.

Kianusch Vahid Yousefnia, Tobias Bölle, Isabella Zöbisch, and Thomas Gerz. A machine-learning approach to thunderstorm forecasting through post-processing of simulation data. *Quarterly Journal of the Royal Meteorological Society*, 150(763):3495–3510, June 2024. ISSN 1477-870X. doi: 10.1002/qj.4777. URL http://dx.doi.org/10.1002/qj.4777.

Frédéric Vitart and Yuhei Takaya. Lagged ensembles in sub-seasonal predictions. *Quarterly Journal of the Royal Meteorological Society*, 147(739):3227–3242, July 2021. ISSN 1477-870X. doi: 10.1002/qj.4125. URL http://dx.doi.org/10.1002/qj.4125.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. Predrnn: recurrent neural networks for predictive learning using spatiotemporal lstms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 879–888, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders, 2023. URL https://arxiv.org/abs/2301.00808.

Wanghan Xu, Fenghua Ling, Wenlong Zhang, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. Generalizing weather forecast to fine-grained temporal scales via physics-ai hybrid modeling, 2025. URL https://arxiv.org/abs/2405.13796.

Günther Zängl. Adaptive tuning of uncertain parameters in a numerical weather prediction model based upon data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 149(756): 2861–2880, August 2023. ISSN 1477-870X. doi: 10.1002/qj.4535. URL http://dx.doi.org/10.1002/qj.4535.

A Appendix

A.1 Comparisons between new architectures and its original variants

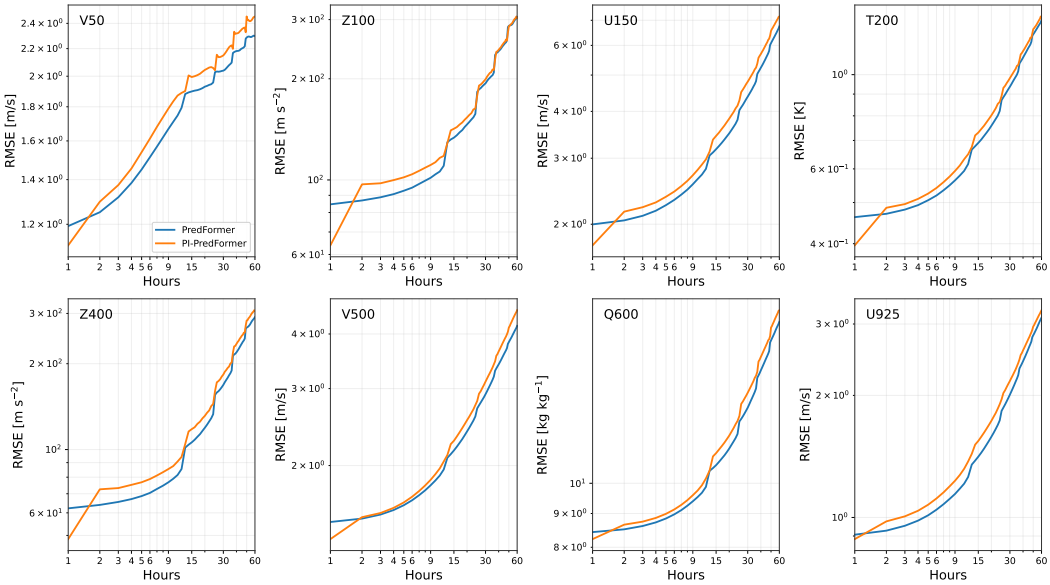

Figure 6: Forecast accuracy of PredFormer and PI-PredFormer in terms of RMSE for geopotential, temperature, wind, and humidity. PI-PredFormer achieves improved accuracy at the 1-hour horizon, when true observations are still available as input. At longer lead times, however, errors grow faster as PDE corrections are applied to increasingly inconsistent neural states.

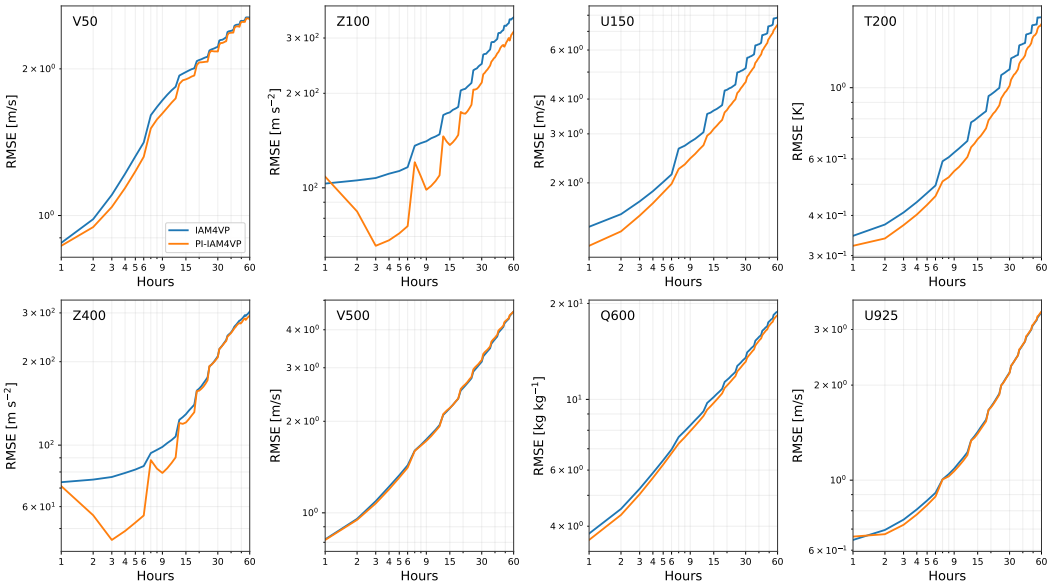

Figure 7: Forecast accuracy of IAM4VP and PI-IAM4VP in terms of RMSE for geopotential, temperature, wind, and humidity. Across all variables and levels, the physics-informed variant provides consistently lower errors without degrading performance at longer horizons.

A.2 Hybrid Block in WeatherGFT

In the WeatherGFT architecture, each hybrid block combines outputs from the neural and physical branches before applying the PDE simulation. This summation originally uses fixed weights:

$$S = 0.1 \cdot P + 0.9 \cdot N, \tag{1}$$

where P and N represent the outputs of the physical and neural components respectively, and S is the resulting tensor passed to the PDE core. The original implementation did not justify the choice of 0.1 and 0.9 weights in Eq. (1). To evaluate the impact of these coefficients, we introduced a learnable tensor $\gamma \in \mathbb{R}^{C \times H \times W}$, yielding a dynamic combination:

$$S = \gamma \cdot P + (\mathbf{1} - \gamma) \cdot N, \tag{2}$$

where the multiplication is element-wise. A separate γ is used for each hybrid block.

Models are trained using the global WeatherBench dataset (2000–2003) and validated with 2004 data. Results showed that learning γ did not improve prediction accuracy. This outcome is attributed to the neural convolution already having sufficient capacity to adapt the scaling internally. However, increasing the fixed weight of the physical branch (i.e., setting γ closer to 1) led to performance degradation, underscoring the critical role of neural corrections (see Figure 8).

Additional analysis of the trained γ values, averaged over spatial dimensions and grouped by variable channel, revealed coefficients clustering around 0.5 - indicating a roughly equal contribution from physics and neural paths across the five PDE-governed variables.

The spatial averaging of the values of γ in the channels highlighted the reduced physical influence near the boundaries of the map (see Figure 9). This effect stems from the limitations of discretized PDE solvers, which lack full neighborhood information at the edges. For example, the northernmost latitude cannot access data beyond the grid, which would be interpolated in a continuous PDE framework. The discrepancy becomes more pronounced in later hybrid blocks, where prediction errors accumulate and neural corrections dominate.

This Hybrid Block is used as basis for novel Single Hybrid Block Architecture, as displayed on Figure 10.

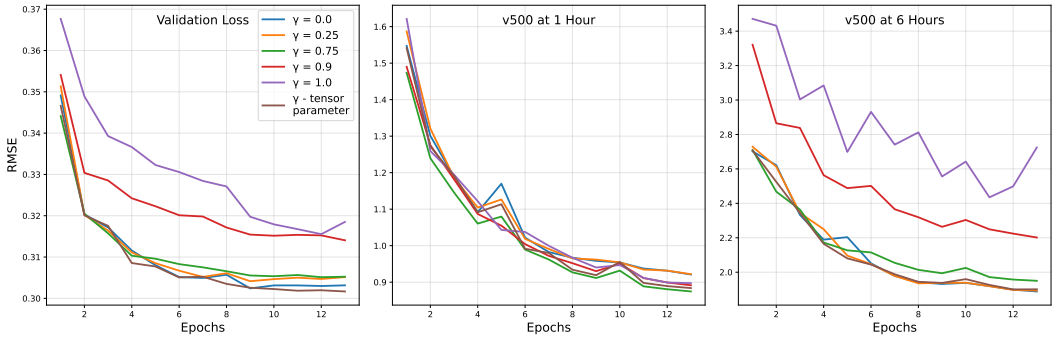

Figure 8: Effect of varying the weight of the physical outputs, γ, on forecast errors: fixed γ values and the dynamic learnable γ tensor across epochs.

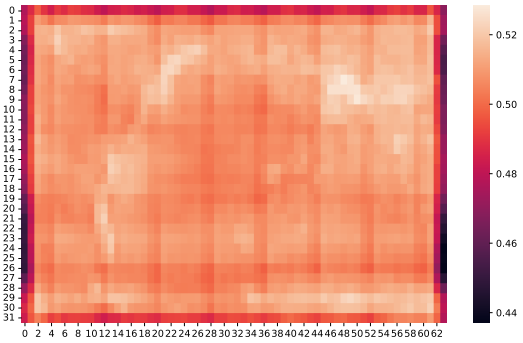

Figure 9: Heatmap of γ averaged over channels C. Larger values indicate greater contribution from the physical branch.

A.3 Architecture details of presented models

All architectures considered in this study, including both the baseline neural networks and our physics-informed extensions are in Table 3.

Two established data-driven models were selected as baselines. SimVP (Gao et al., 2022) is a lightweight convolutional network for next-frame spatiotemporal prediction, while PredRNN (Wang et al., 2017) is a recurrent architecture with spatiotemporal memory, widely used as a benchmark in neural weather forecasting.

All neural components are trained using the AdamW optimizer with cosine annealing learning-rate schedule. The initial learning rate is set to 10^{-4} for PredFormer-based models and $5 \cdot 10^{-4}$ for the other architectures. Batch size is 2 for PredFormer-based models and 16 for IAM4VP-based models, training is performed for 20 epochs with early stopping on plateau. Weight decay is not applied.

The physics-informed modules are integrated directly into the training loop: after the neural network produces forecasts at each step, the PDE solver applies a correction before the loss is computed. The objective function combines mean squared error across all prognostic variables with the implicit enforcement of physical consistency.

All experiments are conducted on NVIDIA GPUs with mixed precision to reduce memory footprint. Training a single model required approximately 150 GPU hours, while inference runs in real time on a single GPU.

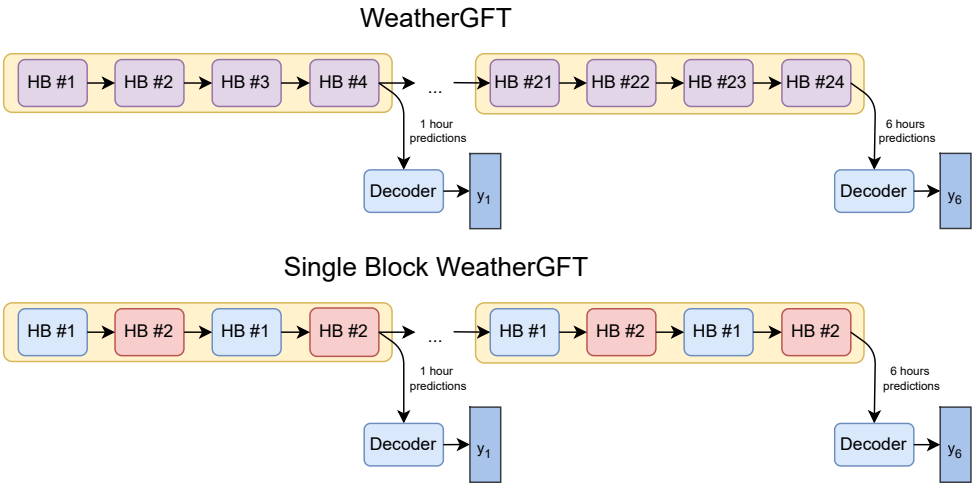

Figure 10: Schematic of the original WEATHERGFT and the single-block variant. New architecture uses 2 Hybrid Blocks (HB) instead of 24, significantly reducing number of model parameters. Single consolidated Hybrid Block is implemented as two original Hybrid Blocks to accommodate the Shifting Windows attention mechanism, as described in (Liu et al., 2021).

A.4 ADDITIONAL COMPARISON BETWEEN MODELS

The anomaly correlation coefficient (ACC) complements RMSE by quantifying spatial pattern fidelity relative to climatology. Figure 11 shows that physics-informed variants maintain higher ACC values up to 12 h, especially for wind and geopotential, indicating improved phase alignment with reanalysis. At longer ranges, ACC values converge across models.

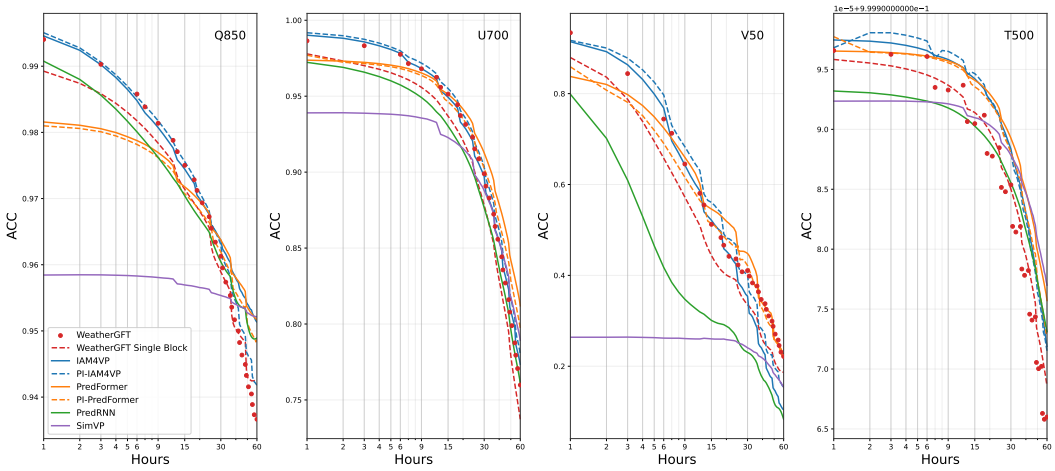

Figure 11: Comparative evaluation of ACC at the Q850, U700, V50, and T500 for various models: WeatherGFT (red dots), WeatherGFT Single Block (red dashes), IAM4VP (blue), PI-IAM4VP (blue dashed), PredFormer (orange), PI-PredFormer (orange dashed), PredRNN (green), and SimVP (purple).

The map-based comparison in Figure 12 illustrates these findings for forecasts of zonal wind at 150 hPa and specific humidity at 1000 hPa. Hybrid models such as WeatherGFT and PI-IAM4VP better reproduce realistic synoptic structures and moisture gradients, whereas purely neural baselines

Table 3: Overview of backbone and hybrid architectures evaluated in this study

Model	Parameters	Type	Forecasting strategy	Key features
PredFormer (Tang et al., 2024)	108M	Transformer	MIMO	Factorized temporal/spatial attention; efficient spatiotemporal learning
PI-PredFormer (this work)	108M	Hybrid (Transformer+PDE)	MIMO + post-hoc correction	PredFormer with physics block applied every 15 min
IAM4VP (Seo et al., 2023)	12M	ConvNeXt semi-autoregressive	Semi-autoregressive	Dual-queue blending of observations and predictions; compact and efficient
PI-IAM4VP (this work)	12M	Hybrid (ConvNeXt+PDE)	Semi-autoregressive + physics	IAM4VP with embedded hourly physics correction
WeatherGFT (Xu et al., 2025)	256M	Hybrid (Transformer+PDE)	SISO (24-step chain)	Original architecture: 24 specialised NN–PDE modules (15 min each)
WeatherGFT-Single (this work)	99M	Hybrid (Transformer+PDE)	SISO (autoregressive)	Single hourly hybrid block, reused iteratively; reduced overfitting
PredRNN (Wang et al., 2017)	41M	Recurrent NN	SISO (autoregressive)	Popular spatiotemporal RNN baseline
SimVP (Gao et al., 2022)	5M	CNN baseline	MIMO	Lightweight fully convolutional model

show grid artifacts or excessive smoothing. These visual diagnostics confirm that hybrid approaches achieve a favorable balance between physical fidelity and adaptability to data.

At longer horizons (24–60 h), purely data-driven models—especially PredFormer—outperform their hybrid counterparts as accumulated errors outweigh the benefits of simplified physical constraints. Spatial analyses further reveal that hybrid models produce more realistic synoptic structures, with sharper gradients and fewer grid artefacts than purely neural baselines.

From a meteorological perspective, these findings underline the potential and the limitations of hybrid approaches. Physics-informed corrections are most beneficial when applied frequently and when forecasts remain anchored to observational information. For extended horizons, however, purely data-driven models currently remain more competitive, reflecting both the incompleteness of the simplified PDE cores and the challenges of mitigating error accumulation. This is further supported by the results presented in Table 4, comparing the models at 24 and 60 hours prediciton range.

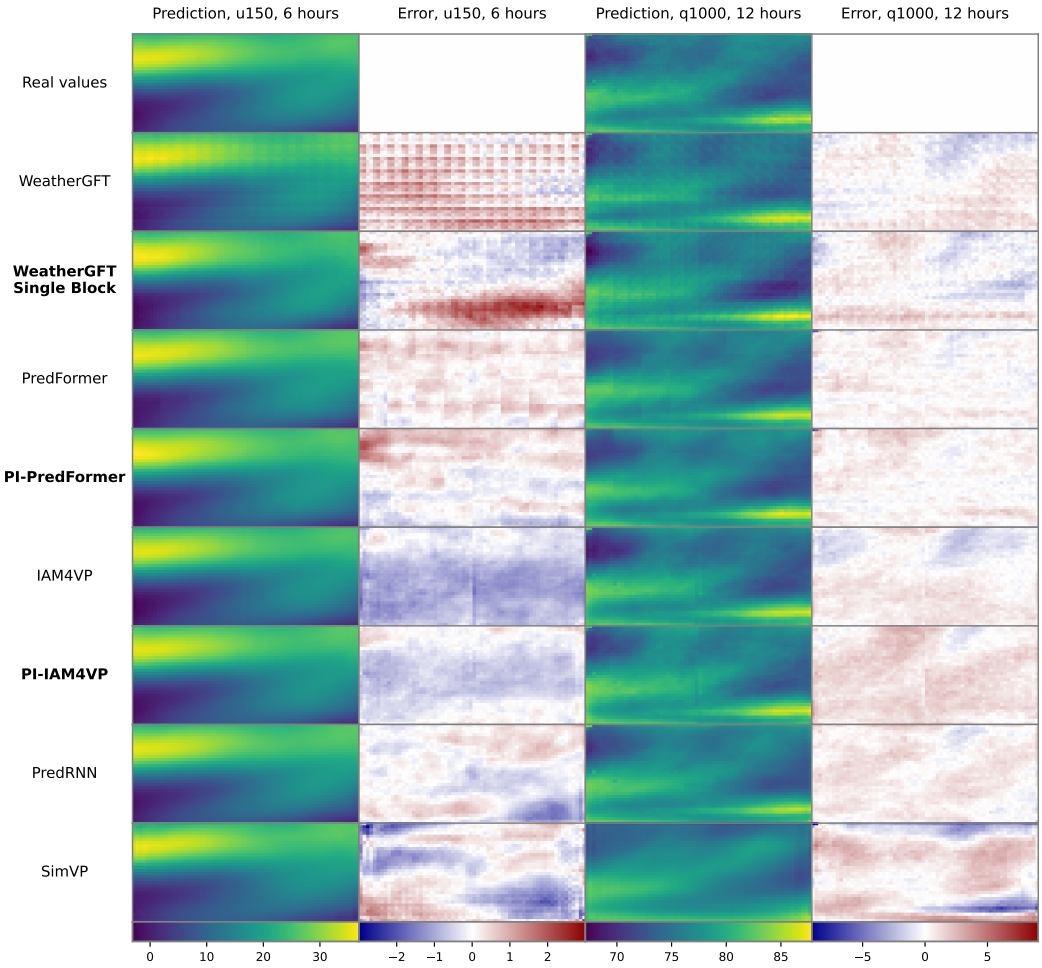

Figure 12: Example forecast maps at 6 h and 12 h for zonal wind at 150 hPa ($u150$) and specific humidity at 1000 hPa ($q1000$). Purely neural baselines exhibit grid artefacts and excessive smoothing, while hybrid models (WeatherGFT-Single, PI-PredFormer, PI-IAM4VP) produce sharper gradients and more realistic synoptic patterns. These maps illustrate the spatial fidelity benefits of embedding physical constraints.

Table 4: The same as in Table 1, but for 24-hour and 60-hour horizons

Model	**z150 (m s^{-2})**		**t150 (K)**		**u500 (m/s)**		**q1000 (kg kg^{-1})**		**Avg. scaled**	
	24 h	**60 h**	**24 h**	**60 h**	**24 h**	**60 h**	**24 h**	**60 h**	**24 h**	**60 h**
WeatherGFT	+65%	+26%	+14%	+19%	+10%	+15%	+5%	+14%	+12%	+18%
WeatherGFT Single Block	+51%	+17%	+25%	+23%	+23%	+19%	+12%	+9%	+19%	+18%
PredFormer	**218.93**	**457.50**	**1.09**	**1.73**	**3.34**	**5.70**	**5.51**	**7.58**	**0.26**	**0.40**
PI-PredFormer	+7%	+3%	+5%	+5%	+6%	+6%	+3%	+5%	+4%	+5%
IAM4VP	+29%	+12%	+10%	+18%	+7%	+10%	+5%	+12%	+8%	+10%
PI-IAM4VP	+19%	+11%	+8%	+10%	+5%	+7%	+6%	+13%	+8%	+10%
PredRNN	+45%	+13%	+26%	+14%	+25%	+14%	+12%	+6%	+19%	+10%
SimVP	+24%	+5%	+40%	+15%	+14%	+4%	+32%	+8%	+15%	+5%