Sonny: An Efficient Approach to Weather Forecasting

Minjong Cheon, Byambatseren Enkhchimeg, Mangare Boby Altein Awumbas, Marie Criz Zaragoza, Hanseon Joo*, Eunjii Lee**
Sejong Univ., Hanyang Univ, Korea Racing Authority**
e-mail:jmj2316@sejong.ac.kr

Sonny: 기상 예측을 위한 효율적인 접근 방식

전민종, Byambatseren Enkhchimeg, Mangare Boby Altein Awumbas, Marie Criz Zaragoza, 주한선*, 이은지**
세종대학교, 한양대학교*, 한국마사회**

Abstract

While deep learning-based weather models show great promise, their high computational cost limits academic accessibility. We introduce Sonny, an efficient hierarchical transformer designed for high-performance forecasting within a modest budget. Sonny features a two-stage StepsNet (narrow slow and full-width fast paths) and employs EMA during training to ensure stable medium-range rollouts without extra fine-tuning. On WeatherBench2, Sonny remains competitive with operational systems and outperforms FastNet in tropical regions. Notably, Sonny can be trained to convergence in just 5.5 days on a single NVIDIA A40 GPU, offering a scalable solution for resource-constrained research.

1. Introduction

For decades, the cornerstone of meteorological science has been Numerical Weather Prediction (NWP) systems. These models rely on the integration of complex partial differential equations to simulate the physical laws governing atmospheric dynamics. While NWP has provided a robust framework for global forecasting, it is increasingly constrained by the immense computational resources required for high-resolution simulations and the inherent difficulties in accurately parameterizing sub-grid scale physical processes. As the demand for more precise and timely climate projections grows, the limitations of these traditional physics-based approaches—specifically their high latency and massive energy consumption—have become more pronounced [1].

To address these growing challenges, the integration of efficient AI-driven architectures has emerged as a critical necessity in modern meteorology. While traditional NWP systems are hindered by the "compute wall," AI models offer a paradigm shift by transforming weather forecasting from a resource-intensive differential equation problem into a highly scalable data-inference task [2].

By leveraging deep learning, these models can bypass the iterative numerical integrations that cause high latency, enabling global forecasts to be generated in seconds on a single GPU rather than hours on massive supercomputing clusters. Furthermore, AI excels at capturing complex, non-linear patterns within sub-grid scale physical processes—such as cloud microphysics and turbulent heat fluxes—that are notoriously difficult to parameterize manually. This efficiency does not merely imply faster results; it fundamentally reduces the energy footprint of climate science, making high-precision forecasting more accessible and sustainable. Ultimately, developing efficient AI is the only viable path to meeting the urgent demand for real-time, hyper-local climate projections in an era of increasing environmental volatility [3].

2. Materials and Methods

2.1 Dataset Description

The model training and evaluation were conducted using the WeatherBench2 (WB2) dataset, configured at a 1.5 degree horizontal resolution 121 x 240 grid points) and a 6-hour temporal frequency. Our input feature set consists of four surface-level variables and five atmospheric variables spanning 13 distinct pressure levels, ranging from 50hPa to 1000hPa. For the experimental setup, we partitioned the dataset chronologically: the period from 1979 to 2018 was utilized for training, while 2019 and 2022 were reserved for validation and testing, respectively [4].

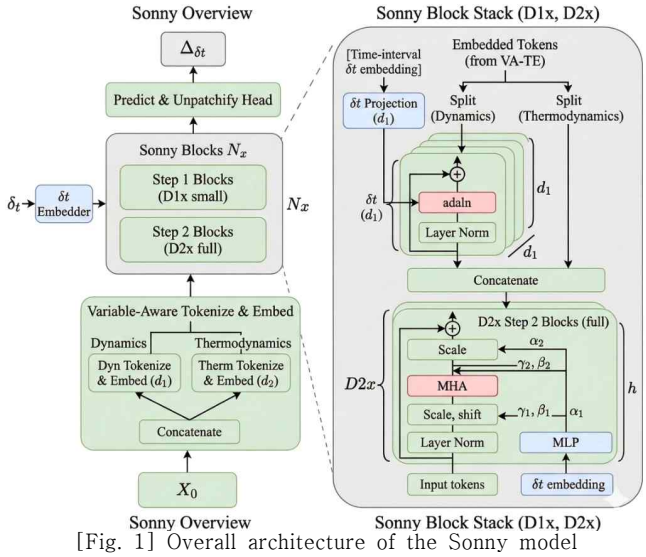
2.2 Sonny

Sonny is a hierarchical weather forecasting model built on the StepsNet architecture. It is specifically designed to leverage the unique physical properties of different atmospheric variables. At the heart of the model is the Variable-Aware Embedding module, which divides input variables into two physically distinct categories: the Dynamics group (comprising wind components U and V, geopotential Z, and surface pressure P) and the Thermodynamics group (comprising temperature T and specific humidity Q). This classification is based on the natural hierarchy of the atmosphere, where large-scale kinematic drivers (dynamics) essentially dictate the environmental evolution that thermodynamic states then follow [4].

Step 1 (The Slow Path): The Dynamics group is projected into a specific d1-dimensional space and passed through a series of N1 Transformer blocks. By limiting the hidden dimensions at this stage, the model can utilize a deeper network to extract the core atmospheric "backbone." This avoids the massive memory costs usually required when processing all variables at once, allowing the model to focus on long-range spatial patterns and the structural changes in pressure fields.

Step 2 (The Fast Path): The refined features from the dynamics stage are combined with the raw thermodynamics embeddings (d2) to create a full representation. These integrated features then move through N2 blocks with a larger combined dimension (d = d1 + d2). This phase is built to model complex, non-linear interactions—such as how moisture moves and releases heat—using the already-processed dynamic skeleton as a structural guide.

To account for different forecast lengths, Sonny uses a Randomized Dynamics Conditioning mechanism. The specific time interval is embedded and integrated into every Transformer block using adaptive Layer Normalization (specifically the adaLN-Zero technique). In the first step, this time information is scaled to match the dynamics dimension, while the second step uses the full-scale embedding.



[Fig. 1] Overall architecture of the Sonny model

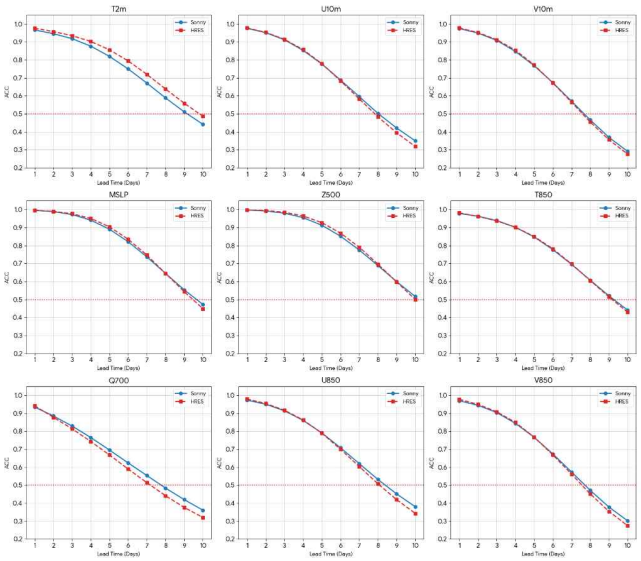
3. Results

Figure 2 illustrates the performance of Sonny compared to the HRES model over a 10-day forecast period, using the Anomaly Correlation Coefficient (ACC) across nine key variables. The results indicate that Sonny maintains a high level of competitiveness across all primary metrics while providing a significantly more efficient pathway for model inference.

The evaluation is further expanded in Figure 4, which compares Sonny against the Met Office GM and FastNet O96 across three major geographical regions: the Northern Hemisphere Extratropics (NHET), the Southern Hemisphere Extratropics (SHET), and the Tropics. In the extratropical regions, Sonny demonstrated remarkably robust performance. For variables such as geopotential at 500 hPa (Z500) and mean sea level pressure (MSLP), the model achieved high ACC scores that were directly comparable to both the Met Office GM and FastNet O96

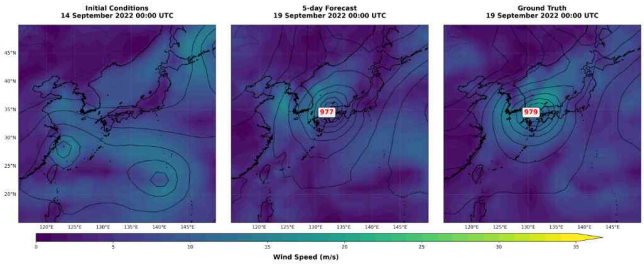
throughout the entire 10-day duration.

A particularly notable finding was observed in the temperature forecasts at 850 hPa (T850). In this category, Sonny showed a much slower rate of performance decline as the lead time increased compared to FastNet O96. In fact, Sonny's accuracy in these forecasts closely aligned with the physics-based Met Office GM, highlighting the model's ability to maintain high fidelity in long-range projections.



[Fig. 2] Comparing Sonny to other weather forecasting models

Figure 3 showcases the regional forecasting results for Typhoon Nanmadol (September 2022), serving as a primary case study for evaluating the model's performance during high-impact weather events over a medium-range horizon. The forecast was initialized at 00:00 UTC on September 14, 2022, and ran for a total of 120 hours (5 days) with data recorded at 6-hour intervals. In terms of intensity, the model predicted a minimum central pressure of 977 hPa, which is remarkably close to the ERA5 reference value of 979 hPa—representing an error of only about 1.4 hPa. These findings are significant because they suggest that the proposed architecture effectively overcomes the "over-smoothing" problem, a common issue in deep-learning weather models where extreme values and sharp atmospheric features tend to be blurred or underestimated.



[Fig. 3] 2022 Typhoon Nanmadol case study

4. Conclusion

This study introduces Sonny, a highly efficient deep learning model tailored for medium-range weather forecasting. We demonstrate that a hierarchical Transformer architecture, utilizing a physically informed variable split, can achieve performance levels competitive with much larger models. The centerpiece of our design is the two-stage StepsNet pipeline. By incorporating randomized dynamics forecasting, we train the model across diverse time intervals, which allows for flexible and adaptive lead-time inference during the forecasting process. Experimental evaluations conducted on the WeatherBench2 dataset show that Sonny possesses strong medium-range forecasting skills, rivaling both traditional operational baselines and current state-of-the-art efficient AI models. From a practical standpoint, Sonny is remarkably resource-efficient, reaching convergence in roughly 5.5 days while running on just a single NVIDIA A40 GPU.

4. Acknowledgement

The authors would like to express their gratitude to Boyoon Choi for her assistance with data preprocessing and English language editing

References

[1] Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... & Anandkumar, A. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214.

[2] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022). Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. arXiv preprint arXiv:2211.02556.

[3] Lam, R., Sanchez-Gonzalez, A., Willson, C., Wirnsberger, P., Fortunato, M., Pritzel, A., ... & Battaglia,

P. (2023). Learning skillful medium-range global weather forecasting. Science, 382(6677), 1416-1421.
[4] Cheon, M. (2026). Sonny: Breaking the Compute Wall in Medium-Range Weather Forecasting. arXiv preprint arXiv:2603.21284.