Submission for AI4X 2026

AI stretch-grid model for kilometre-scale weather prediction over deep tropics

K. W. Chui, Arun Ramanathan, Patel Pratiman, Robert Huva, Chen Chen, Song Chen, Hugh Zhang, Dale Barker

Centre for Climate Research Singapore

Abstract

We present an AI stretch-grid model for kilometre-scale regional weather prediction over the West Maritime Continent. The model is trained on ten years of global ERA5 reanalysis at ~31 km resolution and fine-tuned using three years of kilometre-scale operational weather forecast analyses. A stitched stretch-grid configuration produces 3-hourly forecasts at 6 km resolution over the target region while retaining coarser global coverage elsewhere. Preliminary evaluations against short-range forecasts from the operational physics-based prediction system have been conducted. In particular, 48-hour forecasts are produced within minutes on a single GPU, demonstrating the potential of stretch-grid AI models for efficient, operationally viable regional weather forecasting in the tropics.

1. Introduction

Recent advances in artificial intelligence (AI) have led to the rapid development of data-driven numerical weather prediction (NWP) models trained on large-scale global reanalysis datasets. Over the past few years, several global AI-based NWP systems have demonstrated forecast skill comparable to that of traditional physics-based global NWP models, while offering a significant reduction in computational cost and forecast latency. Prominent examples include FourCastNet3 [1], Pangu-Weather [2], GraphCast [3], Aurora [4], and ECMWF's Artificial Intelligence Forecasting System (AIFS) [5,6]. These models are now being used operationally or experimentally at major meteorological centres, such as ECMWF and NOAA, alongside conventional physics-based NWP systems.

Despite their success, current global AI-NWP models typically operate at horizontal resolutions of approximately 25 km, which remain insufficient to explicitly resolve kilometre-scale, short-lived weather phenomena such as deep convection and organized thunderstorm systems. As a result, their direct application to regional and local forecasting, particularly in the tropics, remains limited. There is therefore a growing need for regional AI-based weather prediction systems that can provide high spatial and temporal resolution forecasts suitable for daily operational use.

Singapore, located in the tropical West Maritime Continent (WMC), is influenced by weather systems spanning a wide range of spatial scales, from large-scale monsoon circulations and regional Sumatra squall lines to highly localized severe thunderstorms. Accurately predicting such multi-scale tropical weather remains challenging, and global AI-NWP models alone are unable to capture these features in sufficient detail. Operational forecasting in Singapore therefore relies heavily on high-resolution physics-based models, such as the SINGV-DA system, to represent kilometre-scale atmospheric processes (Figure 1).

In this study, we develop an AI stretch-grid model that combines global and regional representations within a single AI-based framework. The model employs a dense high-resolution grid over the WMC region and a coarser grid elsewhere, enabling it to retain large-scale global context while resolving regional-scale weather features. The objective is to deliver high-quality, kilometre-scale forecasts with substantially reduced computational latency, making the system suitable for operational deployment

at meteorological agencies in Singapore. The performance of the proposed AI stretch-grid model is evaluated against the operational SINGV-DA system, which currently provides short-range forecasts at 1.5 km resolution for daily weather operations.

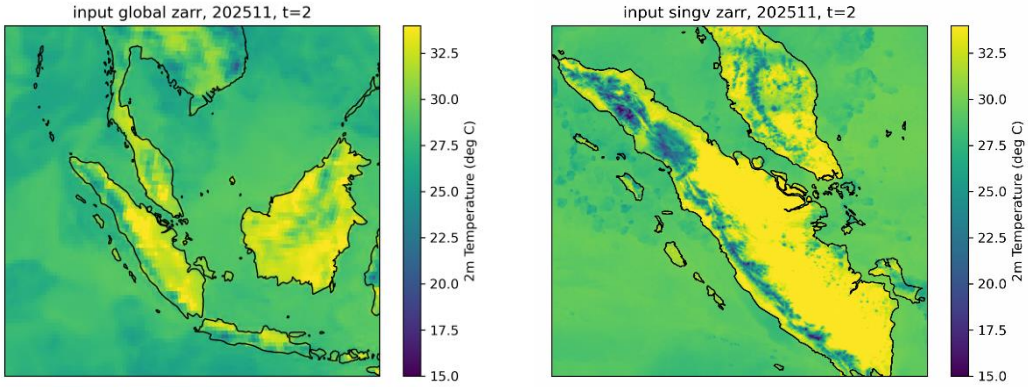

Figure 1. (left) 31km ERA5 2m temperature (right) 6km SINGV-DA (upscaled from the original 1.5km resolution) 2m temperature at 2025-11-01T06Z

2. Data and Methods

2.1 Data for AI model training

Training the AI stretch-grid model (AI-SGM) requires the combination of global coarse-resolution reanalysis data and regional high-resolution numerical weather prediction (NWP) analyses. For the global domain, ERA5 reanalysis data spanning the period 2014–2023 are used. The data are provided on an O-320 Gaussian grid with an approximate horizontal resolution of 31 km and a 3-hourly temporal sampling. These global fields supply large-scale atmospheric context and ensure physical consistency beyond the regional target domain.

For the regional West Maritime Continent (WMC) domain, operational analyses from the SINGV-DA system are employed. The original SINGV-DA analyses at 1.5 km horizontal resolution are spatially aggregated to a 6 km grid to reduce computational cost while retaining kilometre-scale atmospheric structures. The regional dataset covers a three-year period (2023-2025) and is temporally aligned with the global ERA5 data at 3-hourly intervals. The global ERA5 fields and regional SINGV-DA analyses are stitched together to form a unified stretch-grid training dataset, with high-resolution coverage over the WMC and coarser resolution elsewhere.

The set of input variables used for training includes surface-level and pressure-level atmospheric fields, together with static and cyclic variables that encode geographical and temporal information. Surface variables include mean sea level pressure, 2 m temperature, 2 m dew point temperature, 10 m wind components, and accumulated precipitation. Pressure-level variables are extracted at selected levels (1000, 850, 500, and 250 hPa) and include temperature, specific humidity, geopotential height, and horizontal wind components. In addition, static fields such as land-sea mask and cyclic varying fields of latitude, longitude, local time, and Julian day (called positional encodings) are included to improve the model's representation of spatial location and seasonal and diurnal variability. A complete list of variables is summarized in Table 1.

Surface Level Variables	Pressure Level Variables (250, 500, 850 and 1000 hPa)	Cyclic Varying Variables	Static Variables
mean sea level pressure	specific humidity	sine of Julian day	land-Sea Mask
2 m temperature	geopotential height	cosine of Julian day	sine of Longitude
2 m dew point	temperature	sine of local time	cosine of Longitude
10 m u-component of wind	u-component of wind	cosine of local time	sine of Latitude
10 m v-component of wind	v-component of wind		cosine of Latitude
Hourly accumulated precipitation			

Table 1. List of variables used in the training of the AI-SGM.

2.2 Model Training

The AI stretch-grid model (AI-SGM) is trained using the Anemoi framework developed by ECMWF [5,6,7], which implements a graph-based transformer architecture similar to the flagship AIFS model. The model consists of three main components: an encoder, a processor, and a decoder. The encoder maps the input grid data into a latent space representation. The processor evolves this latent representation forward in time to generate forecasts, while the decoder projects the latent representation back onto the physical grid, producing predictions for all required variables.

Forecasting is performed in an auto-regressive manner, where the model predicts the state at the next time step based on previous model outputs. Formally, the prediction at time $t+1$ is given by

$$y(t+1) = f(y(t), y(t-1))$$

where $y(t)$and $y(t-1)$are the predicted atmospheric states at the current and previous time steps, respectively. This formulation allows the model to capture temporal dependencies and dynamically propagate multi-step forecasts, similar to the approach used in AIFS and GraphCast. By learning both spatial and temporal patterns, the AI-SGM can generate stable and consistent forecasts over multiple lead times, from kilometre-scale regional features to larger synoptic structures.

The training strategy follows a multi-stage approach inspired by previous stretch-grid modelling studies [7,8], progressively increasing resolution and regional specificity. The procedure consists of five stages (Figure 2):

- Stage A: Pretraining on global ERA5 data at ~100 km resolution (O-96 grid) to capture large-scale atmospheric patterns.
- Stage B: Fine-tuning the model on global ERA5 data at ~31 km resolution (O-320 grid), refining its representation of medium-scale dynamics.
- Stage C: Stitching the global 31 km ERA5 fields with regional SINGV-DA data aggregated to 6 km resolution. The model is trained on three years of regional data, allowing it to learn high-resolution regional patterns while maintaining global context.
- Stage D: Further fine-tuning on the WMC domain at 3 km resolution to fully resolve kilometre-scale weather features.
- Stage E: Auto-regressive rollout training over 24-hour lead time (8 steps of 3-hourly intervals), enabling the model to learn temporal evolution and multi-step forecasting.

This staged training approach allows the AI-SGM to progressively capture atmospheric dynamics from global synoptic scales down to regional convective scales while maintaining stability and generalization. The combination of global pretraining and regional fine-tuning ensures that the model leverages both large-scale context and high-resolution regional information, producing accurate forecasts across multiple scales with low computational latency.

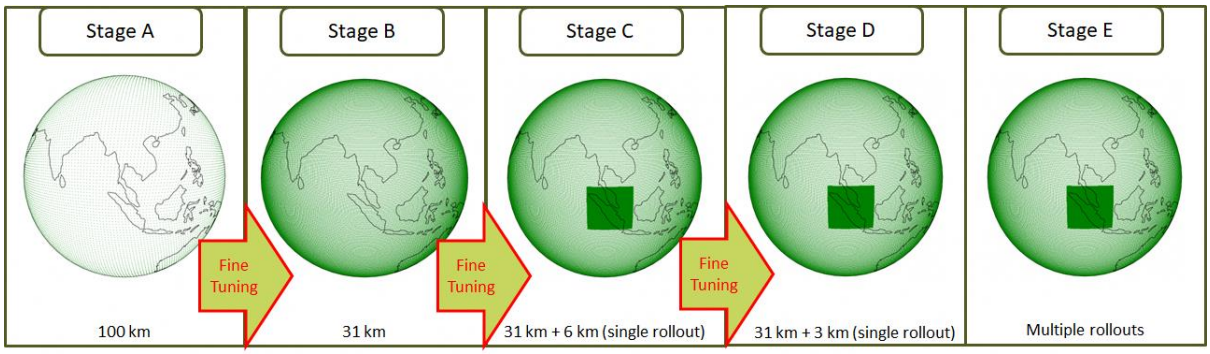

Figure 2. SGM training strategy is designed to have 5 stages starting from the coarse global resolution of 100 km to the final high resolution over WMC region with multiple rollout training over 8-time steps.

3. Evaluation

The performance of the AI stretch-grid model (AI-SGM), trained using high-resolution SINGV-DA analyses, is evaluated against the operational SINGV-DA output as ground truth. Visual comparisons are conducted for key variables, such as 2m temperature, to assess the model's ability to capture regional weather features. Figure 3 shows an example forecast: the left panel displays the SINGV-DA analysis at 2025-11-01T06Z (ground truth), while the right panel shows a representative 3-hour lead prediction from the AI-SGM at the same time. Both plots are presented at 6 km resolution in the regional domain to highlight the model's capability in resolving kilometre-scale structures.

As mentioned in the Methods section, the AI-SGM produces forecasts in an autoregressive manner, where each prediction at a 3-hour step depends on the outputs from the previous two-time steps. For the example in Figure 3, the 3-hour lead forecast is generated using the model states at 2025-11-01T00Z and 2025-11-01T03Z as input. This approach allows the model to capture temporal dependencies and propagate consistent multi-step forecasts.

A major advantage of the AI-SGM is its computational efficiency. A 48-hour forecast at 3-hourly intervals can be generated within minutes on a single GPU, whereas the physics-based SINGV-DA system requires substantially more computational resources (~2000 CPU cores) and several hours of runtime for equivalent forecasts. This low-latency capability demonstrates the potential of data-driven stretch-grid models for operational forecasting in tropical regions, where timely high-resolution predictions are critical for decision-making.

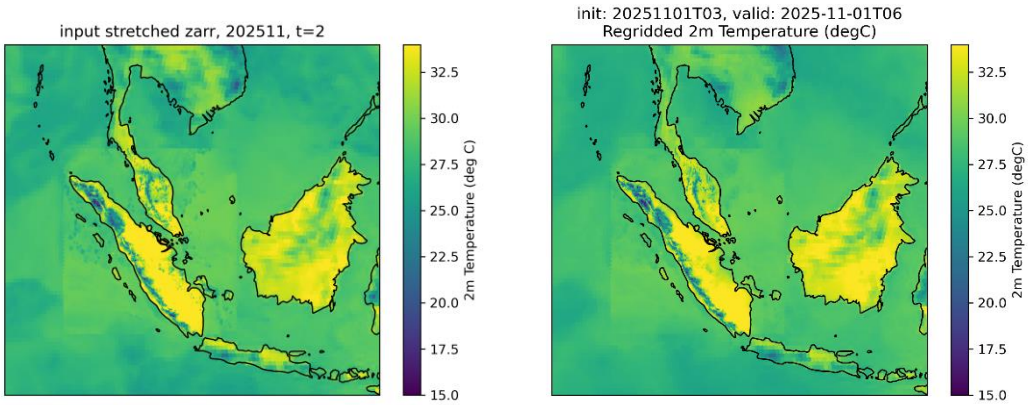

Figure 3. Visual comparison between SINGV-DA analyses at 2025-11-01T06Z (ground truth, left) and SGM prediction at 2025-11-01T06Z (initialized with past two-time steps at 2025-11-01T00Z and 2025-11-01T03Z).

4. Conclusion

In this study, we developed and evaluated an AI stretch-grid model designed to provide high-resolution, low-latency regional weather forecasts over the West Maritime Continent. By combining global ERA5 reanalysis data with regional high-resolution SINGV-DA analyses through a multi-stage training and fine-tuning strategy, the model successfully bridges global-scale context and regional-scale detail within a single unified framework.

Preliminary evaluation results tend to suggest the AI-SGM at 6 km spatial resolution can potentially provide prediction skill comparable to the operational physics-based model for short-range forecasts, trained on a relatively limited regional dataset. Importantly, the AI-SGM offers a substantial computational advantage, producing 48-hour forecasts at 3-hourly intervals within minutes using a single GPU, compared to the significantly higher computational cost and latency of traditional numerical weather prediction systems.

References

[1] Bonev, B., Kurth, T., Mahesh, A., Bisson, M., Kossaifi, J., Kashinath, K., ... & Keller, A. (2025). Fourcastnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *arXiv preprint arXiv:2507.12144*.

[2] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533-538.

[3] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416-1421.

[4] Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., ... & Perdikaris, P. (2025). A foundation model for the Earth system. *Nature*, 1-8.

[5] Chantry, M., Lang, S., Alexe, M., Dramsch, J., Raoult, B., Clare, M., ... & Rabier, F. (2025, January). AIFS-ECMWF's Data-Driven Forecasting System. In *105th Annual AMS Meeting 2025* (Vol. 105, p. 449087).

[6] Moldovan, G., Pinnington, E., Prieto Nemesio, A., Lang, S., Ben Bouallègue, Z., Dramsch, J., ... & Chantry, M. (2025). AIFS 1.1. 0: An update to ECMWF's machine-learned weather forecast model AIFS. *EGUsphere*, *2025*, 1-23.

[7] Nipen, T. N., Haugen, H. H., Ingstad, M. S., Nordhagen, E. M., Salihi, A. F. S., Tedesco, P., ... & Chantry, M. (2025). Regional data-driven weather modeling with a global stretched-grid. *Artificial Intelligence for the Earth Systems*.

[8] Wijnands, J. S., Van Ginderachter, M., François, B., Buurman, S., Termonia, P., & Bleeken, D. V. D. (2025). A comparison of stretched-grid and limited-area modelling for data-driven regional weather forecasting. *arXiv preprint arXiv:2507.18378*.