The 'Butterfly Effect' wall that even Google DeepMind could not overcome... GIST identifies physical limitations of ai weather forecasting models

- *A joint research team led by Professor Jin-Ho Yoon of the Department of Environment and Energy Engineering confirmed discrepancies with actual atmospheric flow in GenCast forecasts compared to conventional numerical weather forecasts due to a lack of interaction between error diffusion and scale*

- *"Verification of physical validity beyond accuracy is needed"… Suggesting the need to redefine evaluation criteria for AI forecast performance*

- *Published in the international academic journal* ***npj Climate and Atmospheric Science***

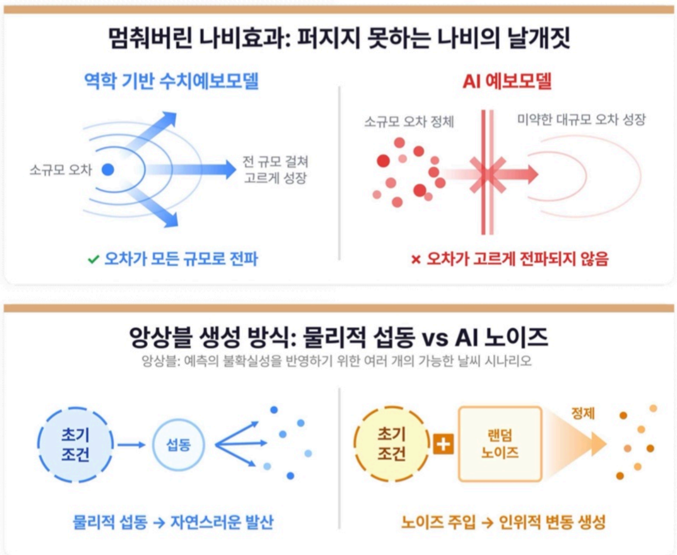
▲ ***Professor Jin-Ho Yoon (corresponding author), Department of Environment and Energy Engineering; master's student Hisu Kim (first author)***

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that a joint research team led by Professor Jin-Ho Yoon of the Department of Environment and Energy Engineering has identified a fundamental limitation in

GenCast, *an AI weather forecasting model developed by Google DeepMind, which fails to adequately reproduce the "Butterfly Effect"*—a core principle of weather forecasting—as realistically as the actual atmosphere.

This study confirmed that similar characteristics appear not only in GenCast but also in AI weather forecasting models using similar methods, demonstrating that it is difficult to determine whether AI models properly reflect actual atmospheric physics based solely on currently widely used performance metrics.

> ** GenCast: An AI weather forecasting model released by Google DeepMind in 2024. After learning from large-scale historical weather data, it utilizes a "Diffusion Model" to predict weather over a probabilistic range of several days to a maximum of about two weeks (15 days).*
>
> ** butterfly effect: A concept first proposed by Edward Lorenz, a professor of meteorology at MIT, which emphasizes the need for probabilistic forecasting to overcome the fact that extremely small differences in initial conditions gradually amplify, ultimately limiting the accuracy of weather forecasts. As a phenomenon occurring in nonlinear systems such as weather forecasting, it developed into chaos theory.*



▲ ***A schematic diagram illustrating the differences between dynamic-based numerical weather prediction models and AI forecasting models.***

Weather forecasts are affected by the "diffuse effect," where very small differences in initial conditions magnify significantly over time. Due to this characteristic, organizations such as the Korea Meteorological Administration probabilistically calculate forecast uncertainty through "ensemble forecasting," which involves performing multiple predictions with slightly different initial conditions.

The key to ensemble forecasting lies in how various initial conditions and prediction results are generated.

The existing model, the European Centre for Medium-Range Weather Forecasts' numerical weather prediction model (ECMWF IFS), utilizes nonlinear characteristics based on weather forecasting equations. It is designed to allow small differences in initial conditions to gradually amplify over time and naturally unfold into future states along various paths. By synthesizing the multiple predictions generated through this process, weather uncertainty and the probability of extreme weather events can be probabilistically evaluated.

In contrast, AI-based weather forecasting models like GenCast perform predictions by learning from data rather than physical equations. Starting from the same initial state, they generate different prediction results (ensemble members) by injecting and removing "random noise" during the forecasting process.

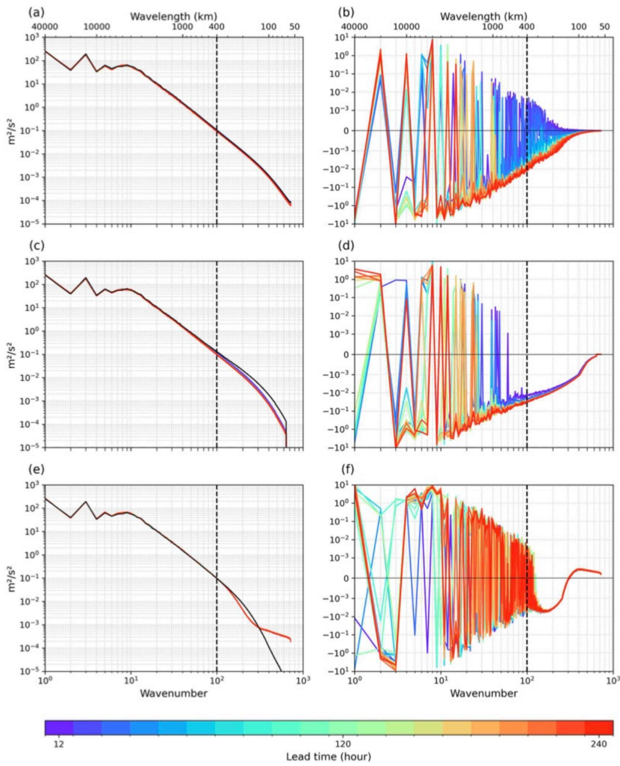
> ** ensemble forecasting: This is a method that performs multiple forecasts by slightly changing initial conditions or model settings, and probabilistically evaluates the uncertainty of the forecast and the probability of extreme weather events based on these results. Derived from the French word meaning "together" or "whole," the term "ensemble" refers to the combined use of multiple predictions or models and is utilized in probabilistic forecasting and machine learning.*

> ** numerical weather prediction (NWP): A weather forecasting model composed of nonlinear equations based on the conservation of energy and momentum, as well as the ideal gas state equation, to explain and predict atmospheric motion. It explains that small differences in initial conditions are sequentially amplified through the interaction of energy and motion within the atmosphere, eventually creating significant differences in weather elements such as clouds, wind, and temperature, which can lead to vastly different prediction results.*

> ** diffusion model: A deep learning technique that generates new data by learning the process of gradually adding noise to data and then removing it. It is widely used in image generation (e.g., DALL-E, Stable Diffusion) and is also being applied to weather forecasting recently.*

The research team confirmed the need to re-examine the reliability of AI ensemble forecasts and conducted a comparative analysis of the European Centre for Medium-Range Weather Forecasts' numerical weather prediction model and GenCAST's forecasts.

NWP For weekly forecasts reset over 52 weeks in 2021, we analyzed how kinetic energy is distributed and changes across spatial scales in the upper atmosphere (approximately 9–10 km), where the jet stream—a narrow band of strong winds—is located.



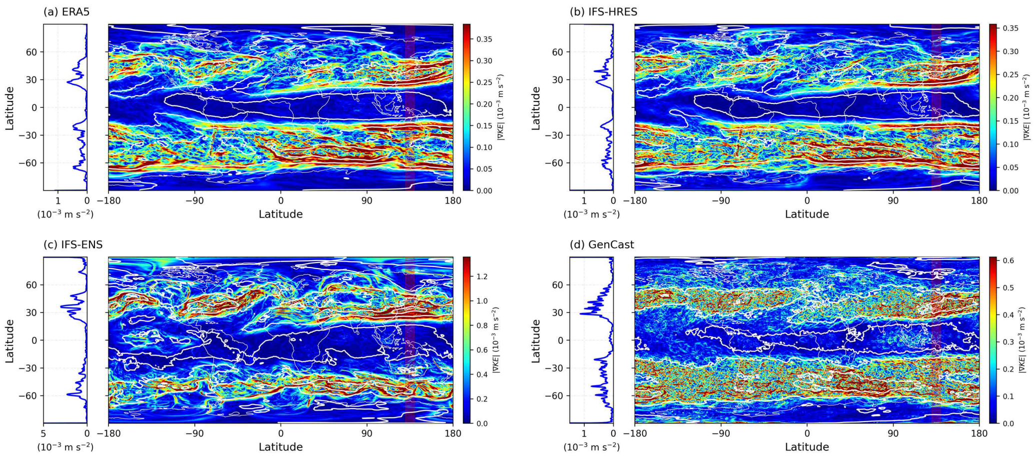
▲ Comparison of 300 hPa kinetic energy spectrum changes across the three models. The left side shows initial energy, and the right side shows the amount of change over time. While the numerical weather prediction models show a decrease in small-scale energy, GenCAST maintains an abnormal distribution caused by noise without change.

As a result, while the existing numerical weather prediction models exhibited a "butterfly effect" where small initial errors gradually amplified along atmospheric flows—such as changes in wind, pressure, and temperature—and naturally spread to various scales, GenCAST revealed a structural limitation: noise injected during the forecasting process did not spread naturally as it does in the actual atmosphere but remained at a specific scale, leaving behind an artificial trace.

Furthermore, while the actual atmosphere involves the interaction of flows of different scales to transfer energy and shape weather, GenCast was found to fail to adequately reproduce realistic atmospheric flow because the interaction between these scales is relatively weak.

In particular, while GenCast reproduces large-scale flows relatively well, it was confirmed that at intermediate scales—which are closely linked to cloud formation and storm development—energy flow remains abnormal, and patterns of "noise forms" different from the actual atmosphere appear.

This suggests that the various prediction results (ensemble) generated by current AI models may be based on statistical variability rather than uncertainty arising from physical laws. This indicates the need to verify physical validity alongside accuracy when evaluating the performance of AI weather forecasting models.



▲ ***Comparison of kinetic energy magnitudes at the upper atmosphere (300 hPa) altitude where the jet stream is located on the 10th day of the forecast (240 hours later). While numerical weather prediction models show distinct, sharp structures along the jet stream, the AI model GenCast displays a blurred jet stream shape where energy is evenly distributed without direction.***

Professor Jin-Ho Yoon stated, “While AI weather forecasting has reached a level similar to existing numerical weather prediction models in terms of conventional indicators and accuracy, how faithfully the results reflect actual atmospheric physics is a separate issue.” He added, “The diversity of current AI ensembles is a limitation based on statistical characteristics rather than uncertainty based on physical laws, so the development of new AI models to overcome this is necessary.”

This research, supervised by Professor Jin-Ho Yoon of the Department of Environment and Energy Engineering at GIST and conducted by master's student Hisu Kim (first author), included co-authors Jihun Ryu, a postdoctoral researcher at Utah State University; Professor Seok-Woo Son of the School of Earth and Environmental Sciences at Seoul National University; Professor Jee-Hoon Jeong of the Department of Environment and Energy at Sejong University; and Professor Hyungjun Kim of the Moon Soul Graduate School of Future Strategy at KAIST.

The results of this research — A spectral test of the butterfly effect and physical consistency in the diffusion-based GenCast's ensembles — which was supported by the Ministry of Science and ICT, the National Research Foundation of Korea's AI-based Future Climate Technology Development Source Research Project and High-Performance Computing Support Project, and the National Science and Technology Council (NST)'s Convergence Research Group Project, were published online in the international academic journal *npj Climate and Atmospheric Science* on March 18, 2026.