Article

A foundation model for the Earth system

https://doi.org/10.1038/s41586-025-09005-y

Received: 28 May 2024

Accepted: 9 April 2025

Published online: 21 May 2025

Open access

Check for updates

Cristian Bodnar[1,2,11], Wessel P. Bruinsma[1,11], Ana Lucic[1,3,11], Megan Stanley[1,11], Anna Allen[4,11], Johannes Brandstetter[1,5], Patrick Garvan[1], Maik Riechert[1], Jonathan A. Weyn[6], Haiyu Dong[6], Jayesh K. Gupta[2,7], Kit Thambiratnam[6], Alexander T. Archibald[4], Chun-Chieh Wu[8], Elizabeth Heider[1], Max Welling[1,3], Richard E. Turner[1,4,9] & Paris Perdikaris[1,10]✉

Reliable forecasting of the Earth system is essential for mitigating natural disasters and supporting human progress. Traditional numerical models, although powerful, are extremely computationally expensive[1]. Recent advances in artificial intelligence (AI) have shown promise in improving both predictive performance and efficiency[2,3], yet their potential remains underexplored in many Earth system domains. Here we introduce Aurora, a large-scale foundation model trained on more than one million hours of diverse geophysical data. Aurora outperforms operational forecasts in predicting air quality, ocean waves, tropical cyclone tracks and high-resolution weather, all at orders of magnitude lower computational cost. With the ability to be fine-tuned for diverse applications at modest expense, Aurora represents a notable step towards democratizing accurate and efficient Earth system predictions. These results highlight the transformative potential of AI in environmental forecasting and pave the way for broader accessibility to high-quality climate and weather information.

Earth system forecasts are indispensable tools for human societies, as evidenced by recent natural events such as the floods in Valencia, the air quality crisis in New Delhi and hurricanes Helene and Milton in the eastern United States. Such systems not only provide crucial early warnings for extreme events but are also important for diverse applications ranging from agriculture to healthcare to global commerce. Modern Earth system predictions rely on complex models developed using centuries of accumulated physical knowledge, providing global forecasts of diverse variables for weather, air quality, ocean currents, sea ice and hurricanes.

Despite their vital role, Earth system forecasting models face several limitations. They are computationally demanding, often requiring purpose-built supercomputers and dedicated engineering teams for maintenance. Their complexity, built up over years of development by large teams, complicates rapid improvements and necessitates substantial time and expertise for effective management. Finally, forecasting models incorporate various approximations, such as those for sub-grid-scale processes, limiting accuracy. These challenges open the door for alternative approaches that may offer enhanced performance.

Machine learning provides an attractive toolbox for addressing these issues. Breakthroughs in numerous fields have shown that complex prediction systems can be streamlined with machine learning models that deliver superior outcomes[4,5]. This concept was introduced to the Earth sciences as early as the 1990s, with pioneering work on neural networks[6] applied to various Earth forecasting problems[7–15]. However, these early models could not scale to replace full dynamical systems. In 2023, a breakthrough came with Pangu-Weather[2], in which a neural network replaced a numerical solver, outperforming state-of-the-art forecasting systems and sparking a wave of weather prediction models based on AI[3,16–18]. These advancements have mostly centred on global medium-range weather forecasting at 0.25° resolution, leaving substantial gaps in other essential areas, including ocean dynamics, wave modelling and atmospheric chemistry. Furthermore, the potential for machine learning to outperform complex extreme weather prediction systems, which at present rely on human analysis of a wide range of models, remains underexplored.

In this paper, we introduce Aurora, a foundation model for the Earth system, capable of tackling a variety of forecasting tasks. Taking inspiration from recent successes of foundation models in other fields[4,5], we first pretrain Aurora on more than one million hours of diverse Earth system data. We then fine-tune the model on a range of downstream tasks, demonstrating for the first time that an AI model can outperform several existing operational systems while also being orders of magnitude faster. Specifically, Aurora achieves state-of-the-art performance in the following critical forecasting domains:

- 5-day global air pollution forecasts at 0.4° resolution, outperforming resource-intensive numerical atmospheric chemistry simulations on 74% of targets;
- 10-day global ocean wave forecasts at 0.25° resolution, exceeding costly numerical models on 86% of targets;
- 5-day tropical cyclone track forecasts, outperforming seven operational forecasting centres on 100% of targets;
- 10-day global weather forecasts at 0.1° resolution, surpassing state-of-the-art numerical models on 92% of targets while also improving performance on extreme events.

Aurora: an Earth system foundation model

Aurora is a machine learning model that produces forecasts for any collection of Earth system variables at any desired resolution. The model

[1]Microsoft Research, AI for Science, Amsterdam, The Netherlands. [2]Silurian AI, Kirkland, WA, USA. [3]University of Amsterdam, Amsterdam, The Netherlands. [4]University of Cambridge, Cambridge, UK. [5]Johannes Kepler University Linz, Linz, Austria. [6]Microsoft Corporation, Redmond, WA, USA. [7]Microsoft Research, Redmond, WA, USA. [8]National Taiwan University, Taipei, Taiwan. [9]Alan Turing Institute, London, UK. [10]University of Pennsylvania, Philadelphia, PA, USA. [11]These authors contributed equally: Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen. ✉e-mail: pgp@seas.upenn.edu

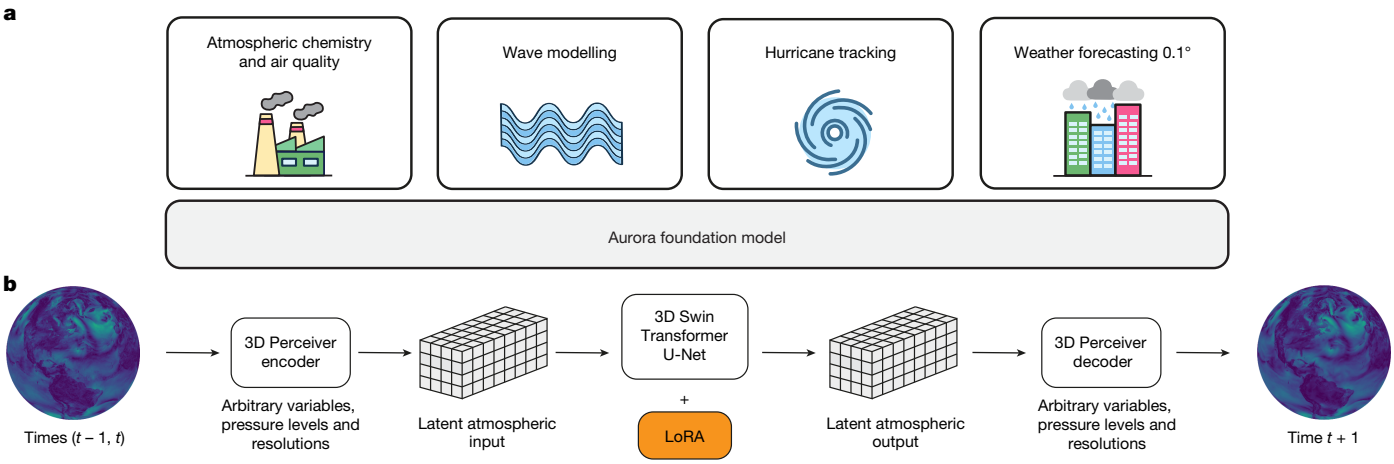


Fig. 1 | Aurora is a 1.3-billion-parameter foundation model for the Earth system. Icons are for illustrative purposes only. **a**, Aurora is pretrained on several heterogeneous datasets with different resolutions, variables and pressure levels. The model is then fine-tuned for several operational forecasting scenarios at different resolutions: atmospheric chemistry and air quality at 0.4°, wave modelling at 0.25°, hurricane tracking at 0.25° and weather forecasting at 0.1°. **b**, Aurora is a flexible 3D Swin Transformer[19] with 3D Perceiver-based[21] atmospheric encoders and decoders. The model is able to ingest inputs with different spatial resolutions, numbers of pressure levels and variables.

consists of three parts: (1) an encoder that converts heterogeneous inputs into a universal latent three-dimensional (3D) representation; (2) a processor that evolves the representation forward in time; and (3) a decoder that translates the standard 3D representation back into physical predictions. The processor is implemented as a 3D Swin Transformer[19,20] and both the encoder and the decoder as Perceiver-based modules[21,22] (Fig. 1). Forecasts for different lead times are generated by recursively feeding predictions back into the model as inputs. For a detailed discussion of the model, see Supplementary Information Section B.

We train Aurora on a large body of Earth system data to learn a general-purpose representation of the dynamics that govern atmospheric and oceanic flow and associated second-order processes. This first training phase is called pretraining and includes a mixture of forecasts, analysis data, reanalysis data and climate simulations (see Supplementary Information Section C.2). After the model has been pretrained, a second training phase can make use of the learned general-purpose representations to efficiently adapt to new tasks, new datasets and new variables. This second training phase is called fine-tuning. Whereas pretraining is expensive and requires large amounts of data, fine-tuning is much cheaper and can typically be performed with little data. We primarily pretrain on atmospheric data, because this is one of the largest sources of information about the dynamical processes underlying the Earth system. Concretely, the pretraining objective is to minimize the next time step (6-h lead time) mean absolute error (MAE) for 150,000 steps on 32 A100 graphics processing units (GPUs), corresponding to approximately 2.5 weeks of training.

Aurora is able to achieve unprecedented performance in fine-tuning tasks by simultaneously scaling the volume of data used during pretraining along with its model size. To evidence this scaling, we demonstrate that pretraining on more diverse data systematically improves validation performance as more datasets are added, especially for extreme values (see Supplementary Information Section G and Extended Data Figs. 1 and 2). Moreover, we demonstrate that validation performance improves by approximately 6% for every ten times increase in model size (see Supplementary Information Section G and Extended Data Figs. 1 and 2). Finally, to measure the benefits of data and model scaling against existing numerical and AI models, we fine-tune Aurora for medium-range weather forecasting at 0.25° resolution, a common task for state-of-the-art AI weather models. Aurora outperforms both the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF)[1], the state-of-the-art numerical weather prediction system and GraphCast[3] on more than 91% of all targets (see Supplementary Information Section H).

Atmospheric chemistry and air pollution

Air quality, a crucial factor in human health, is determined by atmospheric concentrations of specific gases and aerosols[23]. Accurately predicting global atmospheric composition can help mitigate the impact of air pollution events. However, forecasting atmospheric composition is much more complex and computationally costly than weather forecasting. It involves modelling complex chemical reactions through hundreds of stiff equations and accounting for anthropogenic emissions that drive heterogeneous pollution levels globally[24]. The Copernicus Atmosphere Monitoring Service (CAMS) takes this approach and produces global atmospheric composition forecasts and analysis products at 0.4° resolution and reanalysis products at 0.75° resolution[25]. To do this, CAMS extends the IFS with further modules for aerosols, reactive gases and greenhouse gases, which increases computational costs by approximately a factor of ten. So far, no AI method has attempted to produce operational predictions for global atmospheric composition at this scale.

Fine-tuning AI models on CAMS analysis data is extremely challenging for several reasons. First, the CAMS system is relatively new and frequently updated, so training data are limited and change in distribution. Second, air pollution concentrations are highly heterogeneous, sparse and have large dynamic ranges (see 'Discussion'). Finally, pollution is driven by complex anthropogenic factors. These sources underwent complex changes during the global response to the COVID-19 pandemic, further complicating the available training data.

Six air pollutants are the main drivers of poor air quality[23,26]: carbon monoxide (CO), nitrogen oxide (NO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3) and particulate matter at 1 μm (PM_1), 2.5 μm ($PM_{2.5}$) and 10 μm (PM_{10}). Air quality warnings are usually based on threshold values for $PM_{2.5}$ and PM_{10}. Aurora models the five chemical species (CO, NO, NO_2, SO_2 and O_3) across atmospheric levels and as total column (TC) values as well as the particulate matter variables, with CAMS analysis taken to be ground truth. We fine-tune Aurora on CAMS analysis data from October 2017 to May 2022 and test on CAMS analysis data from May 2022 to November 2022 (see Supplementary

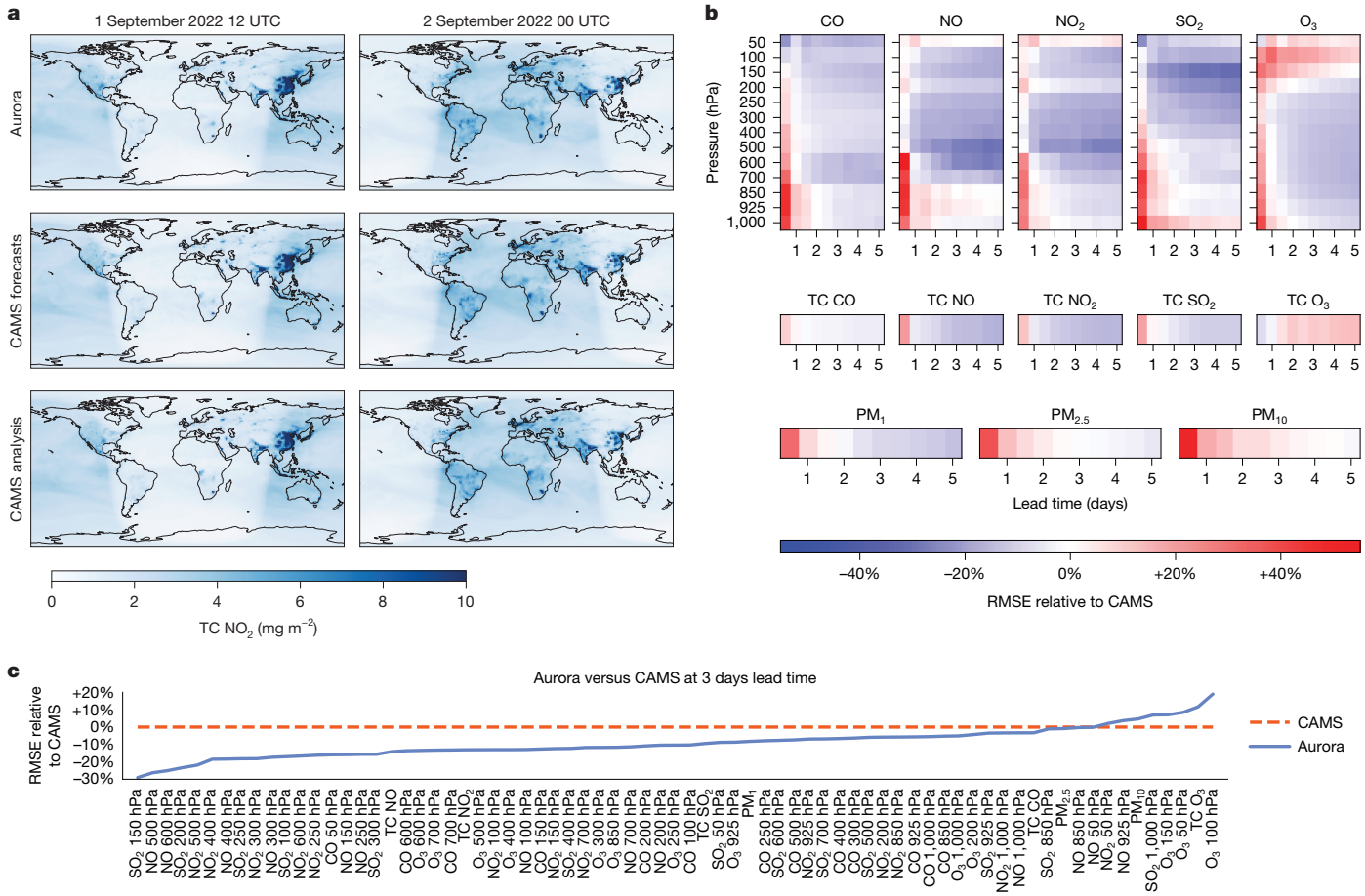


Fig. 2 | In an operational setting, Aurora matches or outperforms CAMS in most comparisons, at orders of magnitude smaller computational expense. a, Predictions for TC NO_2 by Aurora accurately predict CAMS analysis. Predicting atmospheric gases correctly is extremely challenging owing to their spatially heterogeneous nature. In particular, NO_2, like most air pollution variables, is skewed towards high values in areas with large anthropogenic emissions, such as densely populated regions of East Asia. Also, NO_2 exhibits a strong diurnal cycle; for example, sunlight reduces background levels of NO_2 through a process called photolysis. Aurora accurately captures both the extremes and background levels. Aurora and CAMS[25] forecasts are initialized with CAMS analysis on 1 September 2022 at 00 UTC. **b**, Across all lead times, Aurora matches or outperforms CAMS on 74% of all targets. **c**, At a lead time of 3 days, Aurora matches or outperforms CAMS on 89% of all variables. See Supplementary Information Section I.1 for the full results.

Information Section C.4). As the CAMS analysis dataset is very limited in temporal extent, we also incorporate CAMS reanalysis data EAC4 (ref. 27) from January 2003 to December 2021 in the fine-tuning process. We note that CAMS reanalysis data are considered to be lower quality because it uses a lower resolution and a much older version of the underlying model (see Supplementary Information Section C).

Aurora is competitive with CAMS (within 20% root mean square error (RMSE)) on 95% of all targets and matches or outperforms CAMS on 74% of all targets (Fig. 2a). At the 3-day mark, Aurora is competitive with CAMS (within 20% RMSE) on all variables and matches or outperforms CAMS on 89% of all variables (Fig. 2b). CAMS outperforms Aurora on ozone in the very upper atmosphere and the 12-h prediction of all species in the lower part of the atmosphere. Aurora generates these predictions in approximately 0.6 s per hour lead time on a single A100 GPU. This yields roughly a 100,000 times speed-up over CAMS (see Section 2.1.5 in ref. 28 for the cost of the IFS), representing an important advancement in the field of atmospheric composition forecasting. Fine-tuning the pretrained model produces large gains over training a model from scratch, giving improvements for all targets with an average magnitude of 54% (see Fig. I24 in Supplementary Information Section I.1).

We conduct a case study evaluating the predictions of Aurora for PM_{10} on 13 June 2022, when Iraq was hit by a particularly severe sandstorm (see Fig. I21 in Supplementary Information Section I.1), one of a series that led to more than 5,000 hospitalizations in the Middle East[29]. Sandstorms involve complex interactions between particulate matter variables and atmospheric dynamics. Nevertheless, Aurora accurately predicts the sandstorm a day in advance with similar accuracy to CAMS, at a fraction of the computational cost. This case study shows that a foundation model approach for predicting air pollution can generalize to extreme events involving complex interactions between atmospheric dynamics and pollutants.

Ocean wave dynamics

Accurate ocean wave forecasts are critical for shipping, coastal defences, aquaculture, off-shore energy generation and disaster preparedness. The IFS High RESolution WAve Model (HRES-WAM) system[30] produces state-of-the-art wave forecasts up to 10 days lead time. IFS HRES-WAM extends the IFS by adding a coupled ocean wave module. No AI model has yet attempted to produce operational predictions for global wave forecasts at this scale.

Fine-tuning Aurora on the ECMWF's HRES-WAM analysis dataset is challenging. Ocean wave variables include information about the direction, time periods and spectral properties of waves, all of which are complex to model. Wave components can also be absent, meaning

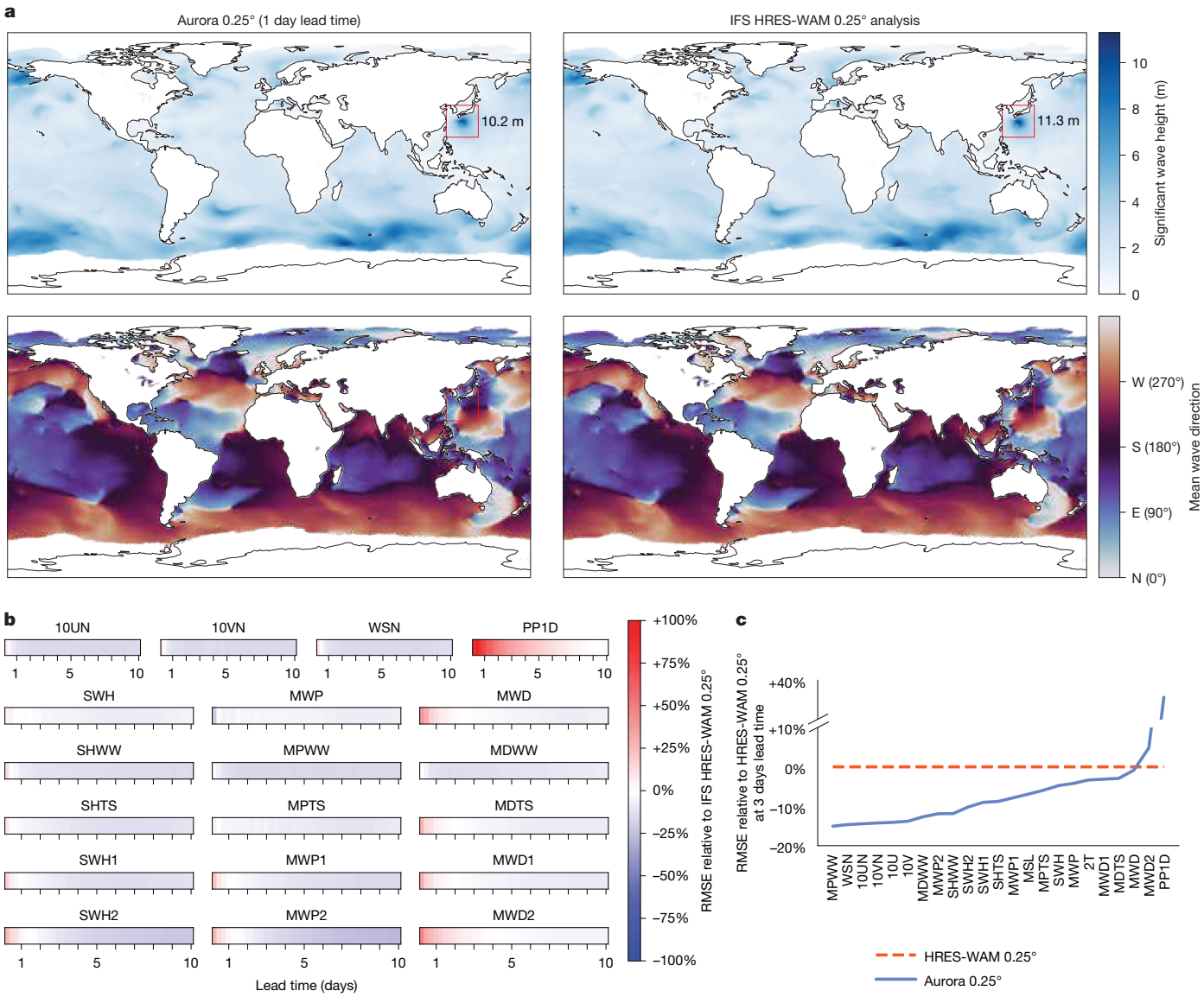


Fig. 3 | In an operational setting, Aurora matches or outperforms HRES-WAM in most comparisons. a, Aurora accurately predicts significant wave height and mean wave direction for Typhoon Nanmadol, the most intense tropical cyclone in 2022. The red box shows the location of the typhoon and the number is the peak significant wave height. Aurora's prediction and HRES-WAM analysis[30] are for 17 September 2022 at 12 UTC, when Typhoon Nanmadol reached peak intensity. Aurora was initialized on 16 September 2022 at 12 UTC. **b**, Across all lead times, Aurora matches or outperforms HRES-WAM on 86% of all wave variables. **c**, At a lead time of 3 days, Aurora matches or outperforms HRES-WAM on 91% of all surface-level variables. See Supplementary Information Section I.2 for the full results.

that the new variables can be undefined at arbitrary and variable spatial locations. Moreover, data for the variables that we consider in this experiment are only available back to 2016, a short record for such a complex task.

The key variables in ocean wave modelling are significant wave height (SWH), mean wave period (MWP) and mean wave direction (MWD). Each of these is predicted for wind waves (WW), total swell (TS), primary swell (1) and secondary swell (2). We also include peak wave period (PP1D) and the components of neutral wind[31] at 10 m, 10UN and 10VN (see Supplementary Information Section C.5). For the full set of variables, see Table C2 in Supplementary Information section C.2. We simultaneously fine-tune Aurora on both wave and meteorological variables by aligning HRES-WAM analysis and HRES T0 in time. HRES T0 refers to the zero-hour forecasts of the high-resolution configuration of the IFS[32], which provides an accurate ground truth for a wide range of meteorological variables. Both HRES-WAM analysis and HRES T0 are regridded to 0.25° spatial resolution. Because the HRES-WAM variables are undefined over land and over oceans whenever sea ice is present, we extend Aurora to support missing data[33] (see 'Discussion'). We use the years 2016–2021 inclusive for fine-tuning and evaluate on 2022 (see Supplementary Information Section C.4).

Aurora is competitive with HRES-WAM (within 20% RMSE) on 96% of all targets and matches or outperforms HRES-WAM on 86% of all wave variables (Fig. 3b). At the 3-day mark, Aurora is competitive with HRES-WAM (within 20% RMSE) on all but one variable, PP1D, and matches or outperforms IFS HRES-WAM on 91% of all variables (Fig. 3c). In particular, Aurora accurately predicts neutral wind speeds, a critical variable for the coupling of atmospheric and wave models[31]. Fine-tuning the pretrained model produces large gains over training a model from scratch, giving improvements for all targets with an average magnitude of 22% (see Fig. I28 in Supplementary Information Section I.2).

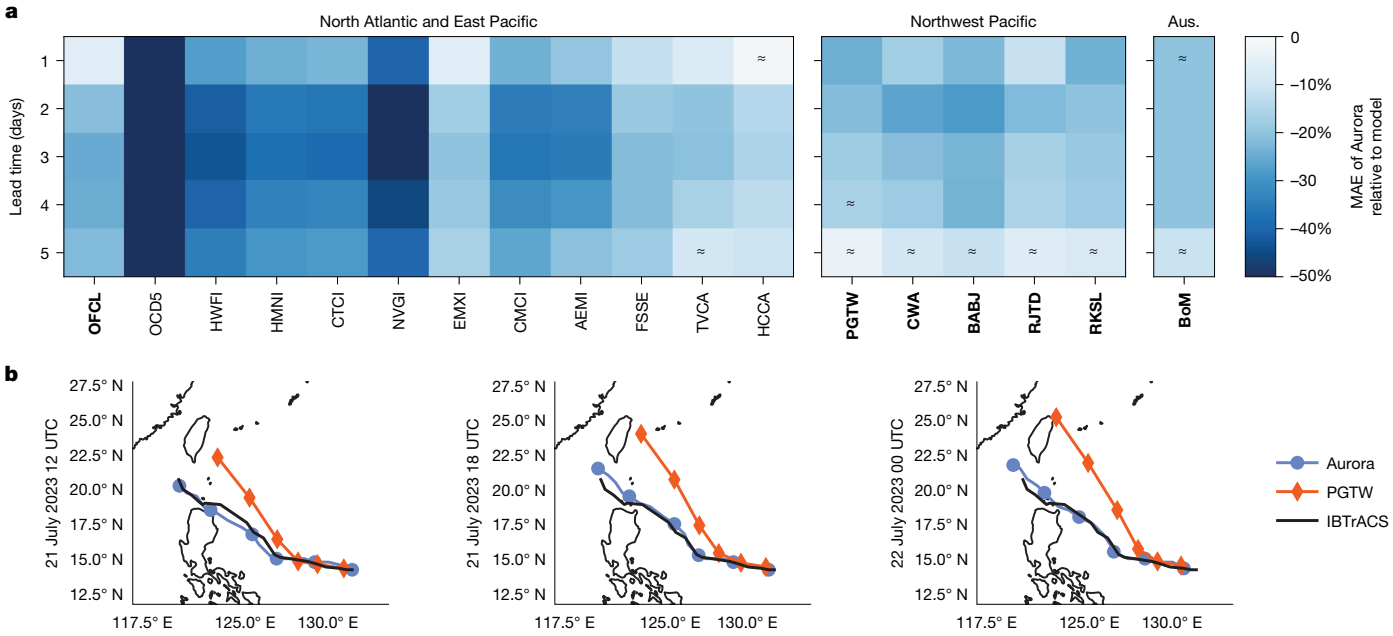


Fig. 4 | In an operational setting, Aurora outperforms state-of-the-art tropical cyclone prediction systems for several agencies and regions worldwide. a, Aurora attains better track prediction MAE than several agencies in various regions. Official forecasts are given by OFCL, PGTW, CWA, BABJ, RJTD, RKSL and BoM (in bold). For the North Atlantic and East Pacific, we also compare with various models used in creating OFCL (not bold). A model does not always make forecasts, which means that different columns are computed over different data. Columns are therefore not indicative of model performance and only indicate the performance compared with Aurora. Here '≈' indicates that the 95% confidence interval for the cell contains zero (see Supplementary Information Section I.3.4 for details). On average, Aurora is 20% better than other agencies in the North Atlantic and East Pacific, 18% in the Northwest Pacific and 24% in the Australian region (Aus.). **b**, On 21 July, a tropical depression intensified into a tropical storm and was named Typhoon Doksuri. Typhoon Doksuri would become the costliest Pacific typhoon so far, inflicting more than US$28 billion in damage. The black lines show its ground-truth paths extracted from IBTrACS[40,41]. Aurora correctly predicts that Typhoon Doksuri will make landfall in the Northern Philippines, whereas PGTW predicts that it will pass over Taiwan.

We conduct a case study of Aurora's prediction of the significant wave height and mean wave direction during Typhoon Nanmadol, which struck the southern coast of Japan on 19 September 2022 (Fig. 3a). Aurora generally produces strong global predictions for significant wave height and mean wave direction that follow the prevailing global wind patterns, with large waves in the typhoon accurately captured.

Tropical cyclone tracking

Tropical cyclones have caused more than US$1.4 trillion in damage since 1950 and pose substantial threats to lives and property[34]. Official forecasts of tropical cyclone tracks are vital for emergency services and the general public. These forecasts are produced by running various dynamical and statistical models, ranging from global ensembles such as the IFS to purpose-built tropical cyclone forecasting systems such as the Hurricane Weather Research and Forecasting model[35]. The output from these systems, together with several consensus products, is analysed by a team of human forecasters who create the final operational product. Here we demonstrate that a single, deterministic run of Aurora fine-tuned to HRES T0 at 0.25° (see Supplementary Information Section H) outperforms the track forecasts from these complex systems for several agencies on a dataset of all tropical cyclones globally in 2022–2023. Aurora fine-tuned to HRES T0 is not specifically fine-tuned for tropical cyclone tracking and therefore illustrates the performance of the model on an unseen downstream task.

Previous comparisons of AI-based tropical cyclone forecasts with official operational forecasts have focused on forecasting track and intensity at short lead times of up to 24 h (refs. 36,37) and showed only marginal improvements at best. The analysis of other large-scale global machine learning models[2,3,38] has been limited to comparisons of tracks, with recent comparisons indicating that performance lags behind that of the official operational forecasts[39].

To generate the track forecasts with Aurora, we run a simple heuristic tracker labelling the centre fix of the vortex as the minimum mean sea-level pressure in consecutive predictions (see Supplementary Information Section I.3.3). We compare the Aurora track predictions with the official forecasts for four basins worldwide, issued by the National Hurricane Center (North Atlantic and East Pacific), China Meteorological Administration, Central Weather Administration Taiwan, Joint Typhoon Warning Centre and Japan Meteorological Agency (Northwest Pacific) and Australian Bureau of Meteorology (Australian region). For all agencies and lead times, Aurora outperforms the official track forecast (Fig. 4a) when compared with the ground-truth paths from the International Best Track Archive for Climate Stewardship (IBTrACS) dataset[40,41]. For example, in the North Atlantic and East Pacific, we observe improvements of 6% at lead time 1 day and 20–25% at lead times 2–5 days. This is the first time that a machine learning model has surpassed full operational tropical cyclone forecasts up to 5 days.

Aurora is able to produce accurate forecasts for several high-impact events. For example, in the case of Typhoon Doksuri in 2023, Aurora accurately predicts landfall in the Philippines at 4 days lead time, in contrast to the official predictions centring the vortex off the coast of Northern Taiwan (Fig. 4b). It is also important to consider the performance of Aurora relative to the wider set of models available to the human forecasters to create the official forecast, as certain models outperform the official prediction at various lead times[42,43]. We therefore compare Aurora with the headline models in the track verification report[42] of the National Hurricane Center (NHC) for the North Atlantic and East Pacific. Aurora outperforms all headline models (Fig. 4a),

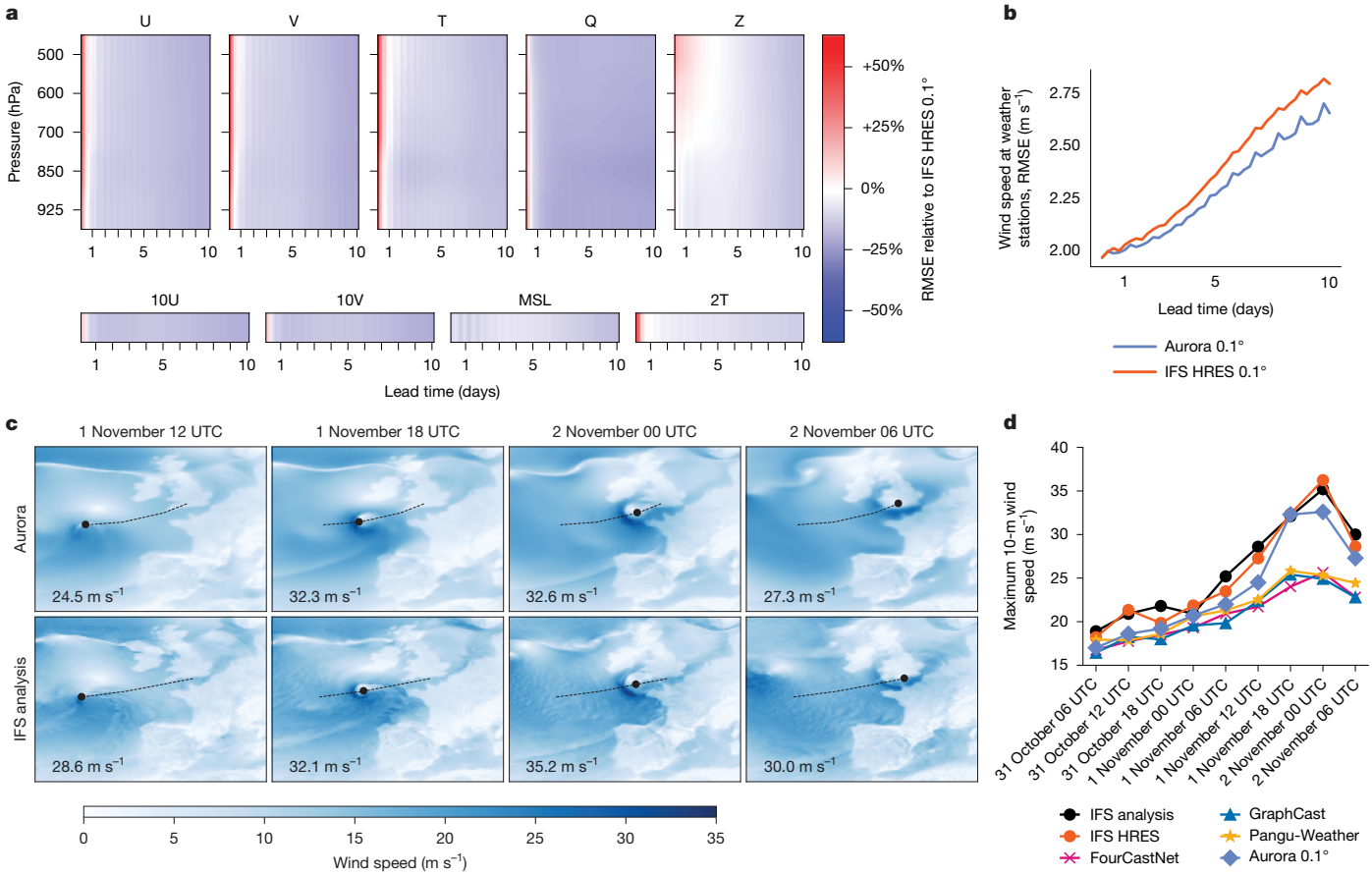


Fig. 5 | In an operational setting, Aurora outperforms IFS HRES in most comparisons and is the only AI model to accurately estimate the maximum wind speeds in Storm Ciarán. a, Aurora outperforms IFS HRES at 0.1° on more than 92% of targets. The scorecard is limited to pressure levels lower in the atmosphere owing to restricted availability of test year data. **b**, Wind speed RMSE computed against measurements at weather stations. Aurora greatly outperforms IFS HRES. **c**, Operational predictions for Storm Ciarán compared with IFS HRES analysis at 0.1°. Black dots show the location of minimum MSL and therefore trace the path of the storm. The maximum 10-m wind speed of the storm is shown in the bottom-left corner of each prediction. To better facilitate the prediction of extreme events, Aurora was run without LoRA. See Supplementary Information Section I.7 for details. **d**, Operational predictions for maximum 10-m wind speed during Storm Ciarán by Aurora, FourCastNet, GraphCast and Pangu-Weather. Aurora is able to predict the sudden increase in 10-m wind speed, unlike the other AI models. The numbers for all AI models except Aurora have been extracted from Fig. 3 in ref. 48.

giving confidence that this is indeed a notable step forward in tropical cyclone track forecast skill.

High-resolution weather forecasting

To accurately resolve high-impact weather events such as severe storms, it is essential that weather prediction systems operate at a high spatial resolution to resolve processes occurring at smaller scales, such as convective and boundary layer effects. HRES[32], the high-resolution configuration of the IFS, operates on a Gaussian grid (TCo1279), which is approximately 0.1° in mid-latitudes. By contrast, current state-of-the-art AI weather prediction models[2,3,16,38,44] can only operate at 0.25° resolution. The reason why state-of-the-art AI approaches are focused on 0.25° is the wealth of high-quality data available at this resolution, whereas 0.1° data are only available from 2016 onwards. Here we demonstrate that a pretraining–fine-tuning protocol can be used to efficiently adapt Aurora to 0.1° and surpass the forecasting skill of IFS HRES under operational evaluation protocols.

We fine-tune Aurora to 0.1° IFS HRES analysis data, which span 2016–2022 (see 'Discussion' and Supplementary Information Section B). For evaluation, we follow the operational protocol in ref. 45, initializing Aurora with IFS HRES analysis and evaluating forecasts against IFS HRES analysis. To ensure that we do not disadvantage IFS HRES, we follow ref. 3 and evaluate IFS HRES against its own so-called zero-hour forecast, referred to as HRES T0, instead of IFS HRES analysis.

Aurora achieves lower RMSE than IFS HRES on 92% of target variables, pressure levels and lead times (Fig. 5a). The performance gains are most pronounced at lead times of more than 12 h into the future, for which we observe a reduction in RMSE of up to 24%. At the shortest lead times, IFS HRES outperforms Aurora for many targets, as is the case for other AI models[3]. We also evaluate the forecasts of Aurora on in situ measurements of 10-m wind speed and 2-m temperature from the WeatherReal-ISD dataset[46], which includes more than 13,000 weather observing stations globally. We find that Aurora outperforms IFS HRES for all lead times up to 10 days (see Fig. 5b and Supplementary Information Section I.5). Owing to the limited availability of 0.1° data, we find that pretraining Aurora is essential in this application. On average, the pretrained model is better than training from scratch by 25% (see Supplementary Information Section I.4).

We conduct a case study of Storm Ciarán, a high-impact mid-latitude storm that took place across Northwest Europe in late 2023, resulting in the lowest recorded pressure in November in England[47]. Following ref. 48, we initialize a selection of AI models at 31 October 00 UTC and compare them with Aurora (see Fig. 5d). We observe that, among the AI models tested[2,3,38], Aurora is the only one capable of accurately predicting the abrupt increase in maximum 10-m wind speed, closely matching IFS analysis, which is taken to be the ground truth.

Discussion

We have introduced Aurora, a large-scale foundation model for the Earth system that outperforms several specialized operational prediction systems at a fraction of the computational cost. We demonstrate state-of-the-art results for air quality, ocean waves, tropical cyclone tracks and high-resolution weather forecasting. From start to finish, each fine-tuning experiment took 4–8 weeks with a small team of engineers, compared with a typical development period of several years for dynamical baseline models. However, it should be noted that such an accelerated timeline is only possible because of the wealth of data that is available as a result of decades of research into traditional numerical approaches.

Improvements are possible along several axes. First, Aurora can easily be extended to generate an ensemble of forecasts, which are crucial in situations in which predictions are uncertain, such as for forecasts at longer lead times or for localized phenomena. Moreover, our scaling results indicate that we have not yet hit a performance ceiling and that improved fine-tuning results can be obtained by scaling pretraining to more diverse data and scaling Aurora to even larger sizes. Although Aurora is fully operational in all experiments, the model does still rely on initial conditions from traditional data assimilation systems. Following recent advances in end-to-end weather forecasting[49], Aurora could be extended to directly operate on observational data. We could also investigate the interpretability of Aurora, aiming to understand whether specific patterns learned by the model can be linked to physical processes.

The potential implications of Aurora for the field of Earth system prediction are profound. Although in this paper we showcase the application of Aurora to four domains, it could be fine-tuned for any desired Earth system prediction task, potentially producing forecasts that outperform the current operational systems at a fraction of the cost. Some examples include predicting ocean circulation, local and regional weather, seasonal weather, vegetation growth and phenology, extreme weather modalities such as floods and wildfires, pollination patterns, agricultural productivity, renewable energy production and sea ice extent. With the ability to fine-tune Aurora to diverse application domains at only modest computational cost, Aurora represents notable progress in making actionable predictions accessible to anyone.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09005-y.

1. European Centre for Medium-Range Weather Forecasts (ECMWF). IFS Documentation CY48R1, Vol. 8. https://doi.org/10.21957/0f360ba4ca (2023).
2. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
3. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
4. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
5. OpenAI et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2024).
6. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
7. Marzban, C. & Stumpf, G. J. A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteorol. Climatol.* **35**, 617–626 (1996).
8. McCann, D. W. A neural network short-term forecast of significant thunderstorms. *Weather Forecast.* **7**, 525–534 (1992).
9. Kuligowski, R. J. & Barros, A. P. Experiments in short-term precipitation forecasting using artificial neural networks. *Mon. Weather Rev.* **126**, 470–482 (1998).
10. Kuligowski, R. J. & Barros, A. P. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather Forecast.* **13**, 1194–1204 (1998).
11. Spellman, G. An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom. *Appl. Geogr.* **19**, 123–136 (1999).
12. Deo, M. & Naidu, C. S. Real time wave forecasting using neural networks. *Ocean Eng.* **26**, 191–203 (1998).
13. Tangang, F., Hsieh, W. & Tang, B. Forecasting the equatorial pacific sea surface temperatures by neural network models. *Clim. Dyn.* **13**, 135–147 (1997).
14. Hsieh, W. W. & Tang, B. Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Am. Meteorol. Soc.* **79**, 1855–1870 (1998).
15. Kolehmainen, M., Martikainen, H., Hiltunen, T. & Ruuskanen, J. Forecasting air quality parameters using hybrid neural network modelling. *Environ. Monit. Assess.* **65**, 277–286 (2000).
16. Chen, L. et al. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, 190 (2023).
17. Han, T. et al. FengWu-GHR: learning the kilometer-scale medium-range global weather forecasting. Preprint at https://arxiv.org/abs/2402.00059 (2024).
18. Chen, K. et al. FengWu: pushing the skillful global medium-range weather forecast beyond 10 days lead. Preprint at https://arxiv.org/abs/2304.02948 (2023).
19. Liu, Z. et al. Swin Transformer: hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 10012–10022 (IEEE, 2021).
20. Dosovitskiy, A. et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, https://openreview.net/forum?id=YicbFdNTTy (Curran Associates, 2021).
21. Jaegle, A. et al. Perceiver: general perception with iterative attention. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 4651–4664, https://proceedings.mlr.press/v139/jaegle21a.html (PMLR, 2021).
22. Jaegle, A. et al. Perceiver IO: a general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, https://openreview.net/forum?id=fILj7WpI-g (Curran Associates, 2022).
23. World Health Organization (WHO). WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. https://www.who.int/publications/i/item/9789240034228 (WHO, 2021).
24. Brasseur, G. P. & Jacob, D. J. *Modeling of Atmospheric Chemistry* (Cambridge Univ. Press, 2017).
25. European Centre for Medium-Range Weather Forecasts (ECMWF). IFS Documentation CY48R1 - Part VIII: Atmospheric Composition. https://doi.org/10.21957/749dc09059 (2023).
26. U.S. Environmental Protection Agency. Technical assistance document for the reporting of daily air quality – the air quality index (AQI). Technical report, https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-quailty.pdf (2024).
27. Inness, A. et al. The CAMS reanalysis of atmospheric composition. *Atmos. Chem. Phys.* **19**, 3515–3556 (2019).
28. Buizza, R. et al. The development and evaluation process followed at ECMWF to upgrade the integrated forecasting system (IFS). Technical Report 829, https://doi.org/10.21957/xzopnhty9 (2018).
29. Francis, D. et al. On the Middle East's severe dust storms in spring 2022: triggers and impacts. *Atmos. Environ.* **296**, 119539 (2023).
30. European Centre for Medium-Range Weather Forecasts (ECMWF). Ocean Wave Model high resolution 15-day forecast (Set II - HRES-WAM). https://www.ecmwf.int/en/forecasts/datasets/set-ii (2024).
31. European Centre for Medium-Range Weather Forecasts (ECMWF). IFS DOCUMENTATION – Cy43r3 operational implementation 11 July 2017. PART VII: ECMWF WAVE MODEL. https://www.ecmwf.int/sites/default/files/elibrary/2017/17739-part-vii-ecmwf-wave-model.pdf (2017).
32. Malardel, S. et al. A new grid for the IFS. *ECMWF Newsl.* **146**, 23–28 (2016).
33. Gordon, J. et al. Convolutional conditional neural processes. In *International Conference on Learning Representations*, https://openreview.net/forum?id=Skey4eBYPS (Curran Associates, 2020).
34. World Meteorological Organization (WMO). Tropical cyclone. https://wmo.int/topics/tropical-cyclone (2024).
35. National Hurricane Center (NHC). NHC track and intensity models. https://www.nhc.noaa.gov/modelsummary.shtml (2019).
36. Boussioux, L., Zeng, C., Guénais, T. & Bertsimas, D. Hurricane forecasting: a novel multimodal machine learning framework. *Weather Forecast.* **37**, 817–831 (2022).
37. Huang, C., Bai, C., Chan, S. & Zhang, J. MMSTN: a multi-modal spatial-temporal network for tropical cyclone short-term prediction. *Geophys. Res. Lett.* **49**, e2021GL096898 (2022).
38. Kurth, T. et al. FourCastNet: accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proc. Platform for Advanced Scientific Computing Conference*, article no. 13 (Association for Computing Machinery, 2023).
39. DeMaria, M. et al. Evaluation of tropical cyclone track and intensity forecasts from Artificial Intelligence Weather Prediction (AIWP) models. Preprint at https://arxiv.org/abs/2409.06735 (2024).
40. Gahtan, J. et al. International Best Track Archive for Climate Stewardship (IBTrACS) project, version 4r01. NOAA National Centers for Environmental Information, https://doi.org/10.25921/82ty-9e16 (2024).
41. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS): unifying tropical cyclone data. *Bull. Am. Meteorol. Soc.* **91**, 363–376 (2010).
42. National Hurricane Center (NHC). National Hurricane Center forecast verification report. 2022 hurricane season. NOAA technical report, https://www.nhc.noaa.gov/verification/pdfs/Verification_2022.pdf (2022).
43. National Hurricane Center (NHC). National Hurricane Center forecast verification report. 2023 hurricane season. NOAA technical report, https://www.nhc.noaa.gov/verification/pdfs/Verification_2023.pdf (2023).

44. Bonev, B. et al. Spherical Fourier neural operators: learning stable dynamics on the sphere. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 2806–2823, https://proceedings.mlr.press/v202/bonev23a.html (PMLR, 2023).
45. Ben Bouallègue, Z. et al. The rise of data-driven weather forecasting: a first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bull. Am. Meteorol. Soc.* **105**, E864–E883 (2024).
46. Jin, W. et al. WeatherReal: a benchmark based on in-situ observations for evaluating weather models. Preprint at https://arxiv.org/abs/2409.09371 (2024).
47. UK Met Office. Storm Ciarán, 1 to 2 November 2023. https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/weather/learn-about/uk-past-events/interesting/2023/2023_09_storm_ciaran_2.pdf (2023).
48. Charlton-Perez, A. J. et al. Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Clim. Atmos. Sci.* **7**, 93 (2024).
49. Allen, A. et al. Aardvark weather: end-to-end data-driven weather forecasting. *Nature* https://doi.org/10.1038/s41586-025-08897-0 (2025).

Methods

Problem statement

We represent the observed state of the atmosphere and surface at a discrete time t as a multidimensional array $X^t \in \mathbb{R}^{V \times H \times W}$, in which V is the total number of variables and H and W are the number latitude and longitude coordinates, respectively. The state can be split into surface (S^t) and atmospheric (A^t) components: $X^t = (S^t, A^t)$, in which $S^t \in \mathbb{R}^{V_S \times H \times W}$ and $A^t \in \mathbb{R}^{V_A C \times H \times W}$ with V_S the number of surface-level variables, V_A the number of atmospheric variables and C the number of pressure levels. The goal is to predict a future state at time $t' > t$. We learn a simulator $\Phi : (\mathbb{R}^{V \times H \times W})^2 \to \mathbb{R}^{V \times H \times W}$, $\Phi(X^{t-1}, X^t) = \hat{X}^{t+1}$, which maps the observed states at the previous time X^{t-1} and current time X^t to a predicted state $\hat{X}^{t+1}$ at the next time step. For predictions at later time steps, we repeatedly apply the simulator, producing an autoregressive roll-out:

$$\begin{aligned}\Phi\left(X^t, \hat{X}^{t+1}\right) &= \hat{X}^{t+2},\\ \Phi\left(\hat{X}^{t+1}, \hat{X}^{t+2}\right) &= \hat{X}^{t+3},\\ &\vdots\\ \Phi\left(\hat{X}^{t+k-2}, \hat{X}^{t+k-1}\right) &= \hat{X}^{t+k}.\end{aligned}$$

For a detailed description of the notation and problem statement, including the specific multidimensional array dimensions and variable definitions, see Supplementary Information Section A.

The Aurora model

3D Perceiver encoder. To accommodate heterogeneous weather datasets with varying variables, pressure levels and resolutions, we design a flexible encoder that maps different datasets into a standardized 3D representation for input into the model backbone (Extended Data Fig. 3a).

The encoder treats all variables as $H \times W$ images. We incorporate static variables (orography, land–sea mask and soil-type mask) by treating them as extra surface-level variables. The images are split into $P \times P$ patches and the patches are mapped to embedding vectors of dimension D using variable-specific linear transformations. For the surface and every pressure level, the embeddings of different variables are summed and tagged with an additive encoding of the pressure level or a learned vector for the surface. A Perceiver module[21] then reduces variable numbers of physical pressure levels C to a fixed number $L = 3$ of latent pressure levels. The result is a $\frac{H}{P} \times \frac{W}{P} \times L$ collection of embeddings. This 3D representation is tagged with additive encodings for the patch position, patch area and absolute time. These encodings use a Fourier expansion scheme with carefully chosen minimum and maximum wavelengths to capture relevant information at appropriate scales. The patch area encoding enables Aurora to operate at different resolutions.

For a detailed description of the encoder architecture, including specifics on input processing, pressure-level aggregation and further encodings, see Supplementary Information Sections B.1 and B.4.

Multiscale 3D Swin Transformer U-Net backbone. The backbone of Aurora is a 3D Swin Transformer U-Net[19,50], which serves as a neural simulator (see Fig. B1 Supplementary Information Section B.1). This architecture allows for efficient simulation of underlying physics at several scales. This architecture falls under the general family of Vision Transformers. However, unlike classical Vision Transformers, here we use local self-attention operations within windows and a symmetric upsampling–downsampling structure.

The backbone is characterized by the following key features: a symmetric upsampling–downsampling structure with three stages each, enabling multiscale processing; 3D Swin Transformer layers performing local self-attention operations within windows, emulating local computations in numerical integration methods; window shifting every other layer to propagate information between neighbouring regions while accounting for Earth's spherical topology; res-post-norm layer normalization[50] for increased training stability; and a flexible design allowing operation at several resolutions without fixed positional biases.

Our backbone contains 48 layers across three stages, compared with the 16 layers and two stages used in ref. 2. This increased depth is made possible by our efficient encoding procedure, which uses a small number of latent levels. For detailed information on the backbone architecture, including window sizes, attention mechanisms and comparisons with previous work, see Supplementary Information Section B.2.

3D Perceiver decoder. The decoder reverses the operations of the encoder, converting the output of the backbone, again a 3D representation, back to the normal latitude–longitude grid (see Fig. 6b). This involves disaggregating the latent atmospheric pressure levels using a Perceiver layer[21] to any desired collection of pressure levels and dynamically decoding into patches by means of variable-specific linear layers. For a detailed description of the decoder architecture, see Supplementary Information Section B.3.

Training methods

The overall training procedure is composed of three stages: (1) pretraining; (2) short-lead-time fine-tuning; and (3) roll-out (long-lead-time) fine-tuning. We provide an overview for each of these stages in the following.

Training objective. Throughout pretraining and fine-tuning, we use the MAE as our training objective $\mathcal{L}(\hat{X}^t, X^t)$. Decomposing the predicted state $\hat{X}^t$ and ground-truth state X^t into surface-level variables and atmospheric variables, $\hat{X}^t = (\hat{S}^t, \hat{A}^t)$ and $X^t = (S^t, A^t)$ (see Supplementary Information Section A), the loss can be written as

$$\begin{aligned}\mathcal{L}(\hat{X}^t, X^t) = \frac{\gamma}{V_S + V_A}\Bigg[\alpha\left(\sum_{k=1}^{V_S} \frac{w_k^S}{H \times W} \sum_{i=1}^{H}\sum_{j=1}^{W} \left|\hat{S}_{k,i,j}^t - S_{k,i,j}^t\right|\right)\\ +\beta\left(\sum_{k=1}^{V_A} \frac{1}{C \times H \times W} \sum_{c=1}^{C} w_{k,c}^A \sum_{i=1}^{H}\sum_{j=1}^{W} \left|\hat{A}_{k,c,i,j}^t - A_{k,c,i,j}^t\right|\right)\Bigg],\end{aligned} \tag{1}$$

in which w_k^S is the weight associated with surface-level variable k, $w_{k,c}^A$ is the weight associated with atmospheric variable k at pressure level c, α is a weight for the surface-level component of the loss, β is a weight for the atmospheric component of the loss and γ is a dataset-specific weight. See Supplementary Information Section D.1 for more details.

Pretraining methods. All models are pretrained for 150,000 steps on 32 A100 GPUs, with a batch size of one per GPU. We use a (half) cosine decay with a linear warm-up from zero for 1,000 steps. The base learning rate is 5×10^{-4}, which the schedule reduces by a factor of ten at the end of training. The optimizer we use is AdamW[51]. We set the weight decay of AdamW to 5×10^{-6}. The only other form of regularization we use is drop path (that is, stochastic depth)[52], with the drop probability set to 0.2. To make the model fit in memory, we use activation checkpointing for the backbone layers and we shard all of the model gradients across the GPUs. The model is trained using bf16 mixed precision. See Supplementary Information Section D.2 for further details.

Short-lead-time fine-tuning. After pretraining Aurora, for each task that we wish to adapt Aurora to, we start by fine-tuning the entire architecture through one or two roll-out steps (depending on the task and its memory constraints). See Supplementary Information Section D.3 for more details.

Roll-out fine-tuning. To train very large Aurora models on long-term dynamics efficiently, even at high resolutions, we develop a new roll-out fine-tuning approach. Our approach uses low-rank adaptation (LoRA)[53] to fine-tune all linear layers in the backbone's self-attention operations, allowing adaptation of very large models in a data-efficient and parameter-efficient manner. To save memory, we use the 'pushforward trick'[54], which propagates gradients only through the last roll-out step. Finally, to enable training at very large numbers of roll-out steps without compromising memory or training speed, we use an in-memory replay buffer, inspired by deep reinforcement learning[55,56] (see Fig. D2 in Supplementary Information Section D.3). The replay buffer samples initial conditions, computes predictions for the next time step, adds predictions back to the replay buffer and periodically refreshes the buffer with new initial conditions from the dataset. For detailed roll-out protocols for each fine-tuning task, see Supplementary Information Section D.4.

Datasets

Aurora was trained and evaluated using a diverse set of weather and climate datasets, encompassing five main categories: analysis, reanalysis, forecast, reforecast and climate simulation datasets. This variety of data sources exposes Aurora to different aspects of atmospheric dynamics, reflecting variability in initial conditions, model parametrizations and chaotic dynamics. Key datasets used in our experiments include ERA5 reanalysis, HRES operational forecasts, IFS ensemble forecasts, GFS operational forecasts, GEFS ensemble reforecasts, CMIP6 climate simulations, MERRA-2 atmospheric reanalysis, as well as CAMS forecasts, analysis and reanalysis data. For a detailed inventory of all datasets used, including specific pressure levels, resolutions and further context for each dataset, see Supplementary Information Section C. These datasets vary in resolution, variables included and temporal coverage, providing a comprehensive basis for training, fine-tuning and evaluating the performance of Aurora across different scenarios.

Task-specific adaptations

Ocean wave forecasting. In the IFS HRES-WAM analysis data, there is a spatially varying absence of data reflecting the distribution of sea ice among other effects. To account for this dynamic nature of the spatial distribution of defined variables, we give each variable an extra channel to represent the presence of a measurement, so we add an extra set of density variables[33] (see Supplementary Information Section B.8).

Data infrastructure

Training Aurora presented substantial technical challenges owing to the large size of individual data points (nearly 2 GB for 0.1° data) and the need to handle heterogeneous datasets with varying resolutions, variables and pressure levels. Owing to the size of data points, training is typically bottlenecked by data loading and not by the model. This means that training smaller models is not always cheaper, because training costs will be dominated by data loading. We developed a sophisticated data storage and loading infrastructure to address these technical challenges.

Data storage and preprocessing. We use Azure Blob Storage with several optimizations to ensure efficient data access. These optimizations include colocating data and compute to minimize latency and costs, storing datasets in appropriate chunks to avoid unnecessary data download and to minimize the number of concurrent connections and compressing these chunks to reduce network bandwidth.

Data loading. We have developed an advanced multisource data loading pipeline to efficiently handle heterogeneous data. We now outline the main design principles of our pipeline. Datasets are instantiated using YAML configuration files specifying loading parameters. Each dataset generates a stream of lightweight BatchGenerator objects. The scope of the BatchGenerator class is to abstract away the details and particularities of datasets by offering a common interface for generating data batches. The streams are combined, shuffled and sharded across GPUs. After sharding, finally the common interface of BatchGenerator is used to do the work needed to download and construct batches for training and inference.

This pipeline enables efficient training on several heterogeneous datasets by batching only samples from the same dataset together and automatically balances workloads across GPUs by using different batch sizes for different datasets. This design offers flexibility needed to experiment with the Aurora model architecture while efficiently handling the challenges of large-scale, heterogeneous weather data processing. For a detailed description of the data loading pipeline, including the BatchGenerator object structure and the unpacking process, see Supplementary Information Section E.

Verification metrics

We evaluate the performance of Aurora using two main metrics: the RMSE and the anomaly correlation coefficient. Both metrics incorporate latitude weighting to account for the non-uniform grid of the Earth. The RMSE measures the magnitude of errors between predictions and ground truth, whereas the anomaly correlation coefficient measures the correlation between the deviation of the prediction and ground truth from the daily climatology.

To assess performance on extreme weather events, we use a thresholded RMSE. The thresholded RMSE uses a threshold to determine which latitude–longitude grid points should be included in the calculation, allowing for evaluation of model performance across different intensity levels of weather phenomena. The thresholds are defined using the mean and standard deviation of the ERA5 reanalysis data over all training years computed separately for each latitude–longitude point. We vary these thresholds linearly for both positive and negative values to obtain RMSE curves for different intensity levels.

For a comprehensive explanation of the verification methods used in this work, including their mathematical formulation and interpretation, see Supplementary Information Section F. Taken together, the metrics used here provide a robust framework for evaluating the performance of Aurora across various weather conditions, from typical to extreme events.

Further details

Further details are available in the Supplementary Information and rely on refs. 26,46,57–75.

Data availability

Most of the data used to train and evaluate Aurora can be obtained from publicly available sources. The ERA5 dataset can be obtained from the Climate Data Store (CDS) (https://cds.climate.copernicus.eu). The HRES Forecasts, HRES T0 and HRES-WAM data can be obtained from the Meteorological Archival and Retrieval System (MARS) (https://confluence.ecmwf.int/display/WEBAPI/Access+MARS). The ECMWF IFS Ensemble data were obtained from the WeatherBench2 repository (https://weatherbench2.readthedocs.io/en/latest/data-guide.html). The GFS Forecasts and GFS T0 datasets can be downloaded from the National Oceanic and Atmospheric Administration (NOAA; https://www.nco.ncep.noaa.gov/pmb/products/gfs/). The GEFS reforecasts dataset is also made available by the NOAA at https://registry.opendata.aws/noaa-gefs-reforecast/. The MERRA-2 dataset is made publicly available by NASA (https://gmao.gsfc.nasa.gov/reanalysis/merra-2/data_access/). The CAMS global reanalysis (EAC4) is available on the ADS (Atmosphere Data Store) (https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4)[76]. The CAMS forecasts and analysis are similarly available at https://ads.atmosphere.copernicus.eu/datasets/cams-global-atmospheric-composition-forecasts. The WeatherReal-ISD weather station dataset can be downloaded from GitHub (https://github.com/microsoft/WeatherReal-Benchmark).

The ground-truth tropical cyclone tracks were obtained from the International Best Track Archive for Climate Stewardship (IBTrACS)[40,41]. Tropical cyclone tracks for baselines in the North Atlantic and East Pacific were downloaded from Automated Tropical Cyclone Forecast (ATCF; https://ftp.nhc.noaa.gov/atcf/)[57] of the National Hurricane Center (NHC) and tracks for baselines in the West Pacific and Australian region were acquired from private communication with the National Taiwan University and the Australian Bureau of Meteorology. All of our plots were made using Matplotlib[77] and the geographical maps were produced using Cartopy[78]. A more detailed description of the data sources is provided in Supplementary Information Section C.

Code availability

Our code and weights are publicly available at https://github.com/microsoft/aurora (refs. 58–75,79–81).

50. Liu, Z. et al. Swin Transformer v2: scaling up capacity and resolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 12009–12019 (IEEE, 2022).
51. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, https://openreview.net/forum?id=Bkg6RiCqY7 (Curran Associates, 2019).
52. Larsson, G., Maire, M. & Shakhnarovich, G. FractalNet: ultra-deep neural networks without residuals. In *International Conference on Learning Representations*, https://openreview.net/forum?id=S1VaB4cex (Curran Associates, 2017).
53. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. In *International Conference on Learning Representations*, https://openreview.net/forum?id=nZeVKeeFYf9 (Curran Associates, 2022).
54. Brandstetter, J., Worrall, D. E. & Welling, M. Message passing neural PDE solvers. In *International Conference on Learning Representations*, https://openreview.net/forum?id=vSix3HPYKSU (Curran Associates, 2022).
55. Lin, L.-J. *Reinforcement Learning for Robots Using Neural Networks*. PhD thesis, Carnegie Mellon Univ. (1993).
56. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
57. Sampson, C. R. & Schrader, A. J. The automated tropical cyclone forecasting system (version 3.2). *Bull. Am. Meteorol. Soc.* **81**, 1231–1240 (2000).
58. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 1263–1272 (PMLR, 2017).
59. Beyer, L. et al. FlexiViT: one model for all patch sizes. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14496–14506 (IEEE, 2023).
60. Rasp, S. et al. WeatherBench 2: a benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.* **16**, e2023MS004019 (2024).
61. Hersbach, H. et al. ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview (2018).
62. Hersbach, H. et al. ERA5 hourly data on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview (2018).
63. European Centre for Medium-Range Weather Forecasts (ECMWF). Section 2.1.2.4 HRES - High Resolution Forecasts. https://www.ecmwf.int/en/forecasts/datasets/set-i (2024).
64. European Centre for Medium-Range Weather Forecasts (ECMWF). Section 5 Forecast Ensemble (ENS) - Rationale and Construction. https://confluence.ecmwf.int/display/FUG/Section+5+Forecast+Ensemble+%28ENS%29+-+Rationale+and+Construction (2024).
65. National Oceanic and Atmospheric Administration (NOAA). NOAA Global Forecast System (GFS). https://registry.opendata.aws/noaa-gfs-bdp-pds (2024).
66. National Oceanic and Atmospheric Administration (NOAA). NOAA Global Ensemble Forecast System (GEFS). https://registry.opendata.aws/noaa-gefs (2024).
67. Scoccimarro, E., Bellucci, A. & Peano, D. CMCC CMCC-CM2-VHR4 model output prepared for CMIP6 HighResMIP hist-1950. Earth System Grid Federation, https://doi.org/10.22033/ESGF/CMIP6.3818 (2018).
68. European Centre for Medium-Range Weather Forecasts (ECMWF). PRIMAVERA: European Centre for Medium-Range Weather Forecasts (ECMWF) ECMWF-IFS-HR model output for the "hist-1950" experiment. NERC EDS Centre for Environmental Data Analysis, https://catalogue.ceda.ac.uk/uuid/470e43e166c44e5990f4f74bc90562d6 (2022).
69. Global Modeling and Assimilation Office (GMAO). MERRA-2: 2d, 1-hourly, time-averaged, single-level, assimilation, single-level diagnostics V5.12.4. Goddard Earth Sciences Data and Information Services Center (GES DISC), https://disc.gsfc.nasa.gov/information/mission-project?title=MERRA-2 (2022).
70. European Centre for Medium-Range Weather Forecasts (ECMWF). CAMS: global atmospheric composition forecast data documentation. https://confluence.ecmwf.int/display/CKB/CAMS%3A+Global+atmospheric+composition+forecast+data+documentation (2024).
71. Lang, S. et al. AIFS – ECMWF's data-driven forecasting system. Preprint at https://arxiv.org/abs/2406.01465 (2024).
72. Dehghani, M. et al. Patch n' Pack: NaViT, a vision transformer for any aspect ratio and resolution. In *Advances in Neural Information Processing Systems 36* (eds Oh, A. et al.) 2252–2274, https://proceedings.neurips.cc/paper_files/paper/2023/file/06ea400b9b7cfce6428ec27a371632eb-Paper-Conference.pdf (2023).
73. Rasp, S. et al. WeatherBench: a benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.* **12**, e2020MS002203 (2020).
74. Hoffmann, J. et al. Training compute-optimal large language models. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 30016–30030 (Curran Associates, 2024).
75. Gunasekar, S. et al. Textbooks are all you need. In *International Conference on Learning Representations*, https://openreview.net/forum?id=Fq8tKtjACC (Curran Associates, 2024).
76. Inness, A. et al. CAMS global reanalysis (EAC4). Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=overview (2024).
77. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
78. UK Met Office. Cartopy: a cartographic Python library with a Matplotlib interface. https://scitools.org.uk/cartopy (2010–2015).
79. Bodnar, C. et al. microsoft/aurora: v1.5.1. Zenodo https://doi.org/10.5281/zenodo.14983584 (2025).
80. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. ClimaX: a foundation model for weather and climate. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 25904–25938, https://proceedings.mlr.press/v202/nguyen23a.html (PMLR, 2023).
81. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020)

Acknowledgements We thank the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Oceanic and Atmospheric Administration (NOAA) for their commitment to open science and their substantial efforts to generate, curate and openly disseminate all of the datasets that enabled our work and we thank M. Chantry for the helpful advice on the ECMWF's data sources. We thank the Copernicus Atmosphere Monitoring Service (CAMS) team at the ECMWF for insightful discussions. We thank W. Shi, Y. Wang, P. Hu and Q. Meng from Microsoft Research, AI for Science and R. T. des Combes and S. Chen from Microsoft Research for helpful inputs in the early stages of this work. We thank D. Kumar, W. Jin, S. Klocek, S. Xiang and H. Sun from MSN Weather for their technical feedback throughout this project. We also thank D. Schwarenthorer for his help with Azure computing and licensing. Finally, we thank A. Foong and F. Noé for constructive feedback during the writing of this manuscript. We are also grateful to N. Shankar for his assistance with the HRESTO dataset. R.E.T. was financed by EPSRC Prosperity Partnership EP/T005386/1 between Microsoft Research and the University of Cambridge during the final stages of the project.

Author contributions C.B., W.P.B., A.L. and M.S. were the four core contributors of this project. They formulated, implemented and evaluated all aspects of Aurora, including the model architecture, the training and fine-tuning pipelines, as well as all experiments and evaluations, except for the tropical cyclone results. A.A. was also a core contributor, provided critical feedback and helped conceptualize, formulate and design the experiments. A.A. originated and carried out the tropical cyclone experiments and evaluations, with the assistance of W.P.B. and C.-C.W. P.G. and M.R. supported the engineering infrastructure of this project. J.B., J.K.G. and M.W. contributed to the initial development and conceptualization of this research. J.A.W., H.D. and K.T. provided regular feedback and carried out all comparisons against weather station data. A.T.A. provided guidance on the CAMS experiments and model evaluation. E.H. provided assistance with programme management and research timelines. R.E.T. and P.P. supervised all aspects of this project. All authors contributed to the writing and editing of this manuscript.

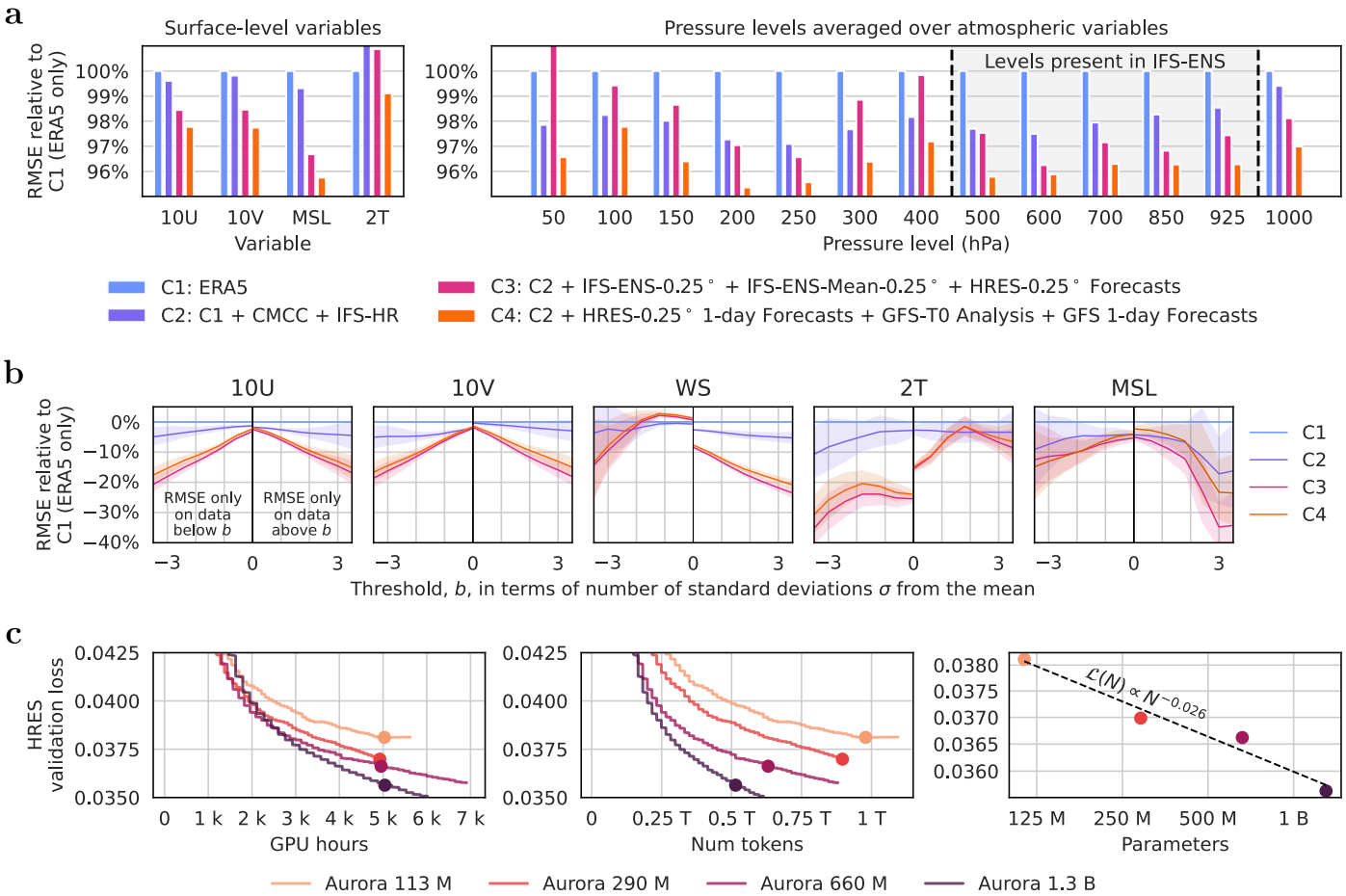
Competing interests C.B., W.P.B., M.S., J.B., P.G., M.R., J.A.W., H.D., J.K.G., K.T., E.H., M.W. and P.P. own Microsoft stock. W.P.B., M.S., P.G., M.R., J.A.W., H.D. and K.T. are Microsoft employees. C.B. and J.K.G. are employees of Silurian AI Inc. and own Silurian AI Inc. stock. The remaining authors declare no competing interests.

Additional information

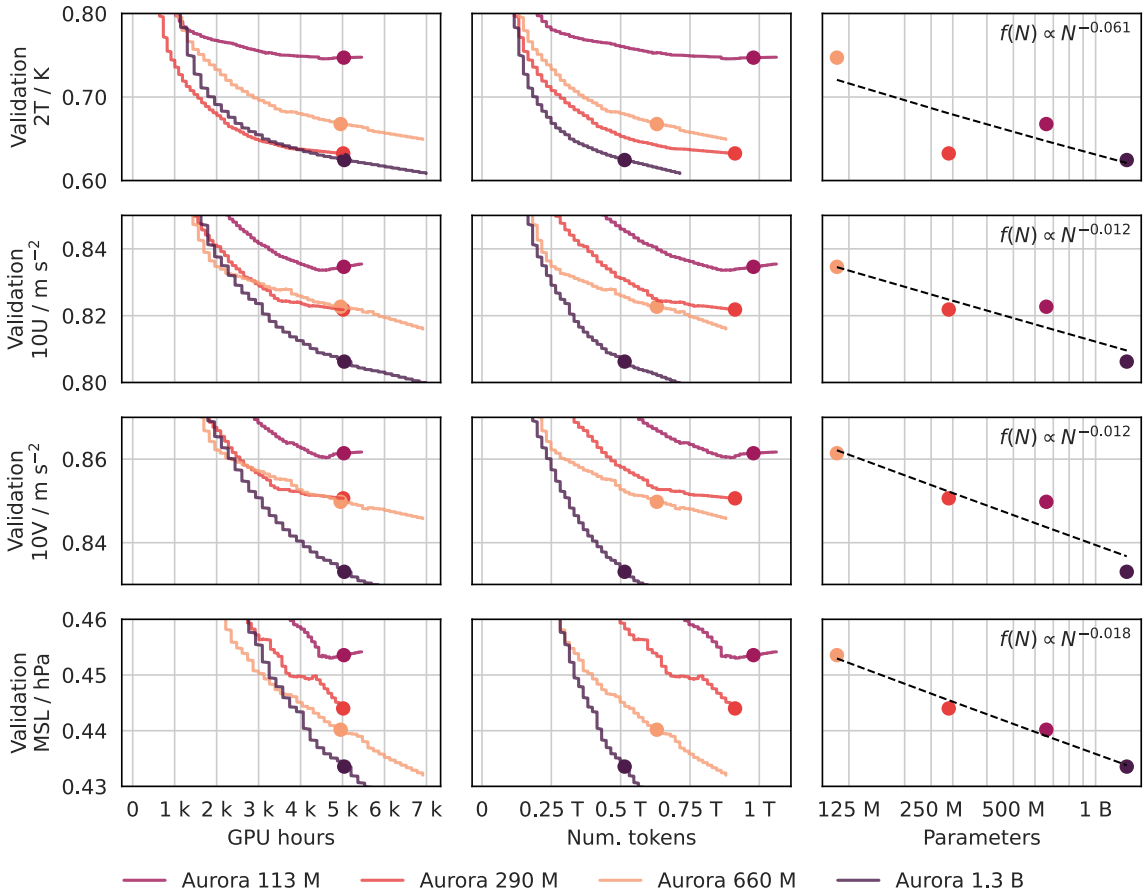
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09005-y.

Correspondence and requests for materials should be addressed to Paris Perdikaris.

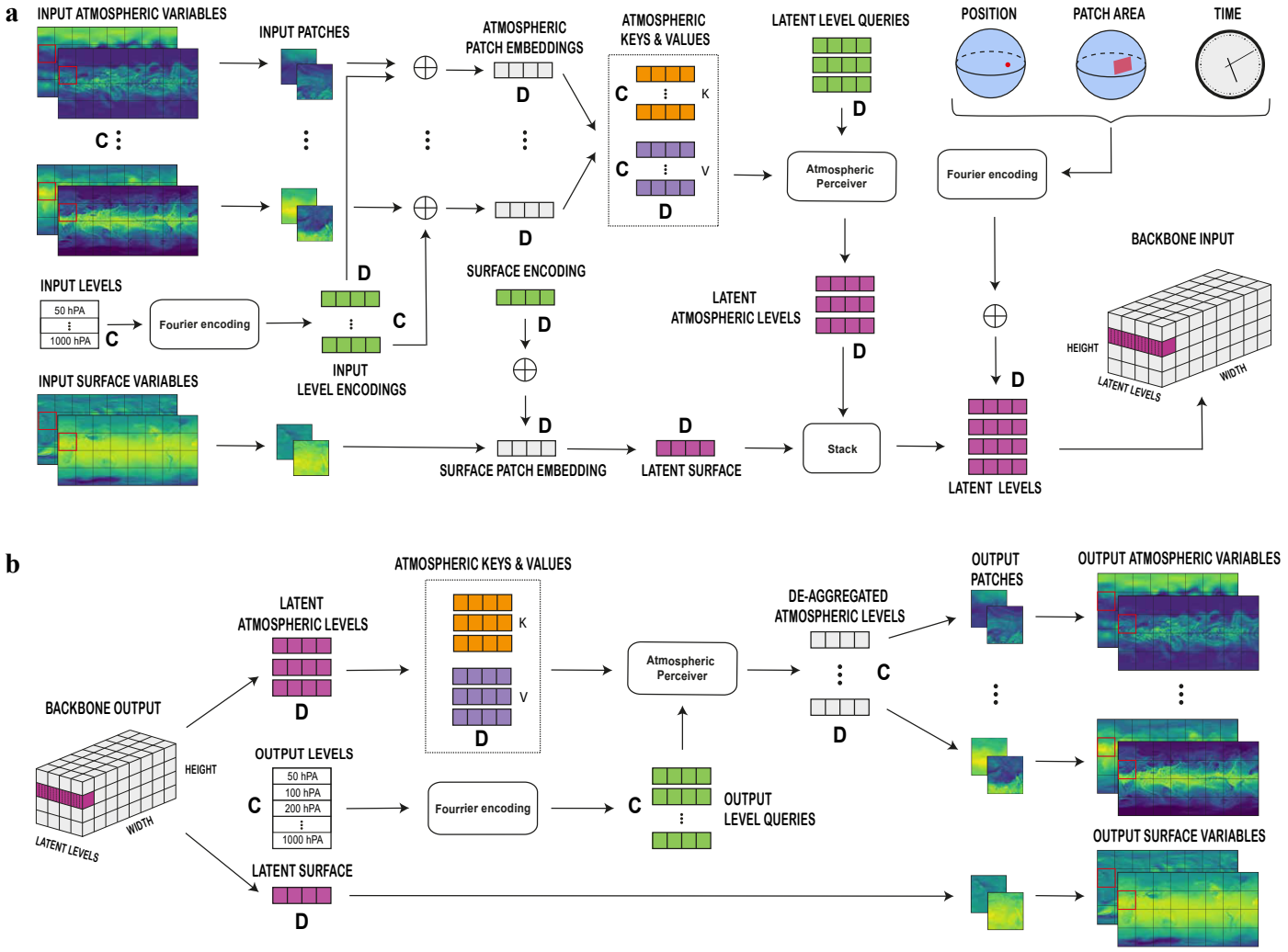
Peer review information *Nature* thanks Qi Tian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Pretraining on diverse data and increasing model size improves performance. a, Performance on ERA5 2021 at 6-h lead time for models pretrained on different dataset configurations, labelled C1–C4, without fine-tuning. Adding low-fidelity simulation data from CMIP6 (that is, CMCC and IFS-HR) improves performance almost uniformly (C2). Adding even more simulation data improves performance further on most surface variables and for the atmospheric levels present in this newly added data (C3). Finally, configuration C4, which includes comprehensive atmospheric coverage and analysis data from GFS, achieves the best overall performance, with improvements across the board. **b**, For the same configurations considered in **a**, performance for extreme values on IFS HRES 2022 at 6-h lead time. Shows RMSEs computed only on data below (left panels) or above (right panels) a threshold b together with a 95% confidence interval obtained through bootstrapping. Pretraining on many diverse data sources also improves the forecasting of extreme values. **c**, Bigger models obtain lower IFS HRES validation loss for the same number of GPU hours. At 5,000 GPU hours, we find that the validation loss behaves like $\mathcal{L}(N) \propto N^{-0.026}$, in which N is the number of parameters, which corresponds to a 6% reduction in validation loss for every ten times increase in model size.



Extended Data Fig. 2 | Validation curves for all surface-level variables during pretraining. For every surface-level variable, at 5,000 GPU hours, we find that the validation loss roughly behaves like $f(N) \propto N^{-\alpha}$, in which N is the number of parameters and $\alpha > 0$ is an estimated parameter.



Extended Data Fig. 3 | Aurora is an encoder–decoder model with a 3D latent representation. The colours are for illustrative purposes only. **a**, Aurora's encoder module. Input weather states are tokenized and compressed into a 3D latent representation using Perceiver-style[21] cross-attention blocks. The resulting latent tokens are augmented with appropriate encodings that provide spatial, temporal and scale information. **b**, Aurora's decoder module. The target output variables are reconstructed in spatial patches by decoding Aurora's 3D latent state using Perceiver-style cross-attention blocks.