Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán

Check for updates

Andrew J. Charlton-Perez [1] ✉, Helen F. Dacre [1], Simon Driscoll[1,2], Suzanne L. Gray [1], Ben Harvey [1,3], Natalie J. Harvey[1], Kieran M. R. Hunt [1,3], Robert W. Lee [1], Ranjini Swaminathan[1,2], Remy Vandaele[1,2] & Ambrogio Volonté [1,3]

There has been huge recent interest in the potential of making operational weather forecasts using machine learning techniques. As they become a part of the weather forecasting toolbox, there is a pressing need to understand how well current machine learning models can simulate high-impact weather events. We compare short to medium-range forecasts of Storm Ciarán, a European windstorm that caused sixteen deaths and extensive damage in Northern Europe, made by machine learning and numerical weather prediction models. The four machine learning models considered (FourCastNet, Pangu-Weather, GraphCast and FourCastNet-v2) produce forecasts that accurately capture the synoptic-scale structure of the cyclone including the position of the cloud head, shape of the warm sector and location of the warm conveyor belt jet, and the large-scale dynamical drivers important for the rapid storm development such as the position of the storm relative to the upper-level jet exit. However, their ability to resolve the more detailed structures important for issuing weather warnings is more mixed. All of the machine learning models underestimate the peak amplitude of winds associated with the storm, only some machine learning models resolve the warm core seclusion and none of the machine learning models capture the sharp bent-back warm frontal gradient. Our study shows there is a great deal about the performance and properties of machine learning weather forecasts that can be derived from case studies of high-impact weather events such as Storm Ciarán.

During the 20th century and the first two decades of the 21st century, numerical weather prediction (NWP) transformed atmospheric science[1]. The combination of physical and mathematical understanding, the availability of high-performance computing and the expansion of the network of Earth system observation led to remarkable and continued progress in the skill and availability of weather forecasts. Numerical weather predictions are a ubiquitous part of modern life, with applications on many different timescales and in sectors as diverse as transport, agriculture, healthcare and recreation.

Over the last two years, machine learning (ML) techniques, a subset of the rapidly developing field of artificial intelligence (AI), have begun to be applied to the weather prediction problem in earnest. Whilst ML has had applications in climate science for many decades[2–4], with these communities aware of its potential[5], and is increasingly used for post-processing weather forecasts[6,7], recent advances in ML and advancements in GPUs (Graphics Processing Units), have enabled the beginning of a 'new dawn' in the application of ML and AI techniques to weather and climate prediction[8].

The publication of the WeatherBench dataset[9] and the 10-year road-map for ML use by the European Centre for Medium-Range Weather Forecasts (ECMWF)[10], amongst other developments, stimulated interest and investment in the development of ML models for weather forecasting. During 2022 and 2023, four ML models were developed by major

[1]Department of Meteorology, Univ. of Reading, Reading, UK. [2]National Centre for Earth Observation, Univ. of Reading, Reading, UK. [3]National Centre for Atmospheric Science, Univ. of Reading, Reading, UK. ✉e-mail: a.j.charlton-perez@reading.ac.uk

technology companies to address the short to medium-range (0–10 day) forecasting problem. These models have all been shown to produce skillful 0–10 day forecasts of the 500 hPa geopotential height field, based on the widely used Anomaly Correlation Coefficient metric[11]. All four models use an encode-process-decode framework but with differing architectures:

- FourCastNet[12], developed by NVIDIA and based on Fourier Neural Operators (FNO) with a vision transformer architecture;
- FourCastNet version 2[13], which builds on FourCastNet by using spherical FNOs;
- Pangu-Weather[14], developed by Huawei and based on a three-dimensional Earth-specific transformer and hierarchical temporal aggregation; and
- GraphCast[15], developed by Google DeepMind and based on graph neural networks.

Similar techniques have been used to develop models for other forecast tasks (e.g., MetNet-3 for 12-h precipitation forecasts in the contiguous United States and 27 European countries[16]). At the present time, ML models primarily produce deterministic forecasts, but rapid progress is being made in producing fully probabilistic forecasts[17–19]. All four ML models are extremely efficient when run on GPU or TPU (tensor processing unit) devices, typically producing 10-day forecasts in a few minutes.

Given the infancy of ML model weather prediction, to the author's knowledge, there are no prior studies that compare how the four ML-models and NWP models capture individual, impactful weather events. Examination of individual weather events available from the papers that describe the ML models is limited to qualitative comparisons of the simulation of tropical cyclones and atmospheric rivers by FourCastNet[12] and quantitative assessment of the tracking error of tropical cyclones by Pangu-Weather[14] and GraphCast[15]. There are no published studies that examine ML model forecasts of extratropical windstorms[20], despite their potential to cause multi-billion dollar damages[21] and increasing severity under climate and population change[22].

In this study, we, therefore, seek to advance knowledge of the comparative performance of ML and NWP models by comparing their forecasts of Storm Ciarán, which affected several European countries during November 2023. This is a valuable out-of-sample test for the ML models because their training datasets all end before the beginning of 2023. We compare the ability of the models to capture the detailed physical structure of the storm and its impacts at two lead times over which operational weather forecasters were actively engaged in issuing weather warnings to the public. An accurate description of the physical structure of this, or any other, storm is a key component of forecasting its compound impact[23] and in constructing plausible storylines for end-users[24].

Results

Storm Ciarán and its associated impacts

Storm Ciarán was first seen as a low-pressure weather system south of Newfoundland at about 00 UTC on 31 October 2023. Based on surface analysis charts issued by the UK Met Office, it then tracked quickly across the North Atlantic, undergoing explosive deepening from 988 hPa at 00 UTC on 1 November to 954 hPa at 00 UTC on 2 November at which time it was located to the northeast of France. This deepening rate, 34 hPa in 24 h means that Ciarán was an extratropical cyclone "bomb"[25]. The lowest pressure recorded, 953 hPa at 06 UTC on 2 November, is a record low pressure for a November storm observed in England[26]. Figure 1 shows surface observations of the 10-m wind speed, cloud cover and mean sea level pressure (MSLP). The cyclonic circulation around the storm centre (with the lowest MSLP observed on the English south coast near the Isle of Wight) has a maximum wind speed of 65 knots on the Normandy coast in France.

Although Storm Ciaran was not a classic Shapiro-Keyser cyclone[27], clear banding in the vicinity of the tip of the cloud as it encircles the storm centre to the poleward side (called the cloud head) could be seen in satellite imagery before it made landfall in northern France. This banding suggests that a sting jet may have been present in Storm Ciarán[28], however its

identification requires methodologies beyond the scope of this study. Gusts of over 100 knots (51 m s^{-1}) were reported in several locations in Brittany[29], with a maximum of 111.7 knots (57.5 m s^{-1}) recorded at Pointe du Raz at approximately 0200 UTC on 2 November[30].

Across Northern Europe, at least 16 people were killed[31]. All flights were cancelled from Amsterdam Schiphol Airport and there were numerous cancellations from Spanish airports. An estimated 1.2 million households in northern France were left without electricity[32] and more than 1 million residents were cut off from the mobile telephone network. Brest and Quimper Airports were also shut and there was disruption to Eurostar operations[33].

Approximately 10,000 homes in Cornwall were left without power, hundreds of schools were closed and many train services were disrupted by fallen trees. Gusts in Channel Islands ranged from 70-90 knots (36-46 m s^{-1})[32] with a maximum gust of 90 knots (46 ms^{-1}) recorded in Alderney at approximately 08 UTC on 2 November[34]. Jersey also experienced a T6 tornado with estimated winds in the region of 161-186 mph (71–83 m s^{-1}). Its 8-km track left a trail of destruction and tens of people needed to leave their homes. It is likely that this is the strongest tornado reported in the British Isles since the Gunnersbury tornado in December 1954[35].

The 10-m wind speed and MSLP structure of Ciarán are shown in Fig. 2a, b at the times when it impacted the land: 00 and 06 UTC 2 November 2023. State-of-the-art model analyses, such as the IFS analysis used in this figure, represent the best three-dimensional estimates of the actual atmospheric state. The low-pressure centre of the storm tracked along the southern UK coast and the strongest winds (turning cyclonically) occurred in an arc in the southwest quadrant of the storm when the strong winds impacted Brittany and later more directly to the south of the low centre when they impacted the Channel Islands. The 10-m winds weakened significantly over land due to surface friction, no longer reaching the threshold for shading in the figure. They also weakened between the two times shown, with the peak winds falling by about 6 ms^{-1}, likely due to a combination of the storm making landfall and having already reached its mature stage. The observed wind speeds, shown by the overplotted colour-filled circles, are consistent with the analysed fields, away from the coastlines but exceed those analysed in some locations, notably some coastal locations and at the narrowest point of the English channel. The winds in these locations will be influenced by local mesoscale processes and so these exceedances are not unexpected given the resolution of the IFS model. The track of the storm, defined as the locations of its minimum MSLP according to IFS analyses, is shown by the black symbols joined by lines in panel (c). Ciarán had its genesis in the western North Atlantic around the time of the first track point shown (06 UTC 31 October) and travelled rapidly eastwards across the North Atlantic. The contours shown illustrate the MSLP and 250-hPa wind speed (i.e., the upper-level jet) at the start, middle and end times of the tracks and show how Ciarán evolved from a weak disturbance (with central MSLP exceeding 995 hPa) to a record-breaking deep storm as it crossed from the equatorward to the poleward side of the jet at about 06 UTC 1 November.

Track, intensification and wind impacts of Storm Ciarán

Ciarán's track was well forecast by both the IFS HRES and ML-models (I. 2(c)) initialised at 00UTC on 31 October, although small differences in the location of the storm centre, and associated wind field, were critical for the accurate predictions of weather warnings along the southern English coast. Two days before Ciarán began to impact land and well before the start of its fast intensification, the spread in the position of the storm in ML models and NWP models is similar.

The evolution of the minimum mean sea pressure (MSLP) at the centre of the developing storm and its associated maximum 10-m wind speed are shown in Fig. 3 for the IFS analysis, IFS HRES forecast and ML model forecasts in (a, c), and for the ERA5 reanalysis, forecasts based on the ERA5 system, and the control (unperturbed) ensemble members of four NWP models in panels (b, d). Considering first the minimum (MSLP) evolution, all the forecasts closely follow both analysis products, capturing both the rapid deepening phase of the storm and its maximum
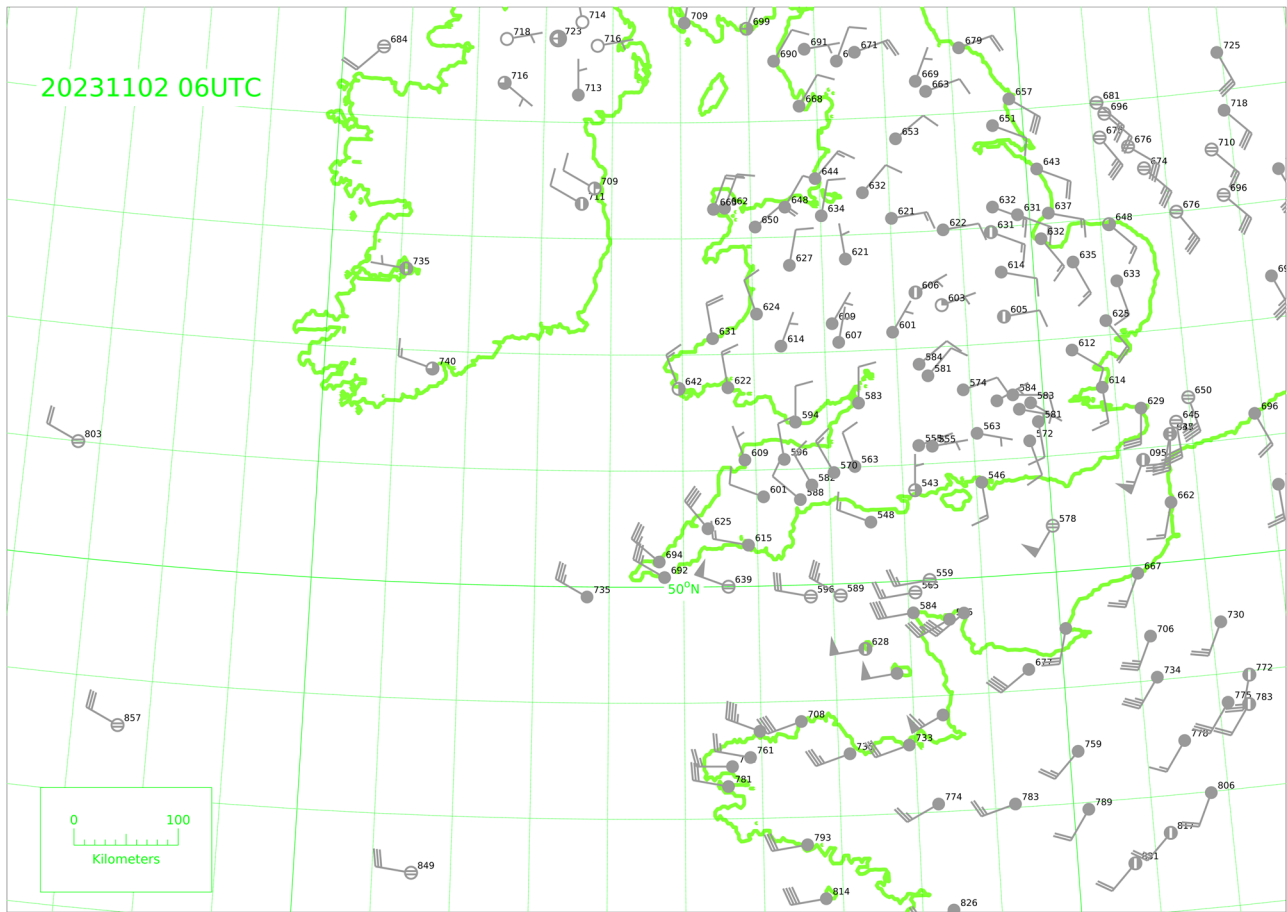
Fig. 1 | Surface land and ship station SYNOP observations of Storm Ciarán at 06 UTC 2 November 2023 extracted from the MetDB database[52,53], which holds data including surface and upper air observations and some satellite data. The observations are shown as simplified station circles using conventional notation[54]. Circle shading indicates cloud cover in octas, wind barbs and feathers indicate wind speed in knots with the wind direction towards the circle, and numbers are the last three digits, including a decimal place, of the MSLP (in hPa) e.g., 543 equates to 954.3 hPa. Some thinning of observations has been performed for clarity and note that two ships both reported at 51.1°N, 1.7°E with different wind speeds and directions.

intensity depth. The minimum MSLP at the end of the forecast (06 UTC 2 November) is 954 hPa in both the IFS analysis and ERA5. This value varies between 951 and 955 hPa for the ML models and between 950 and 953 hPa for the six NWP models (including the IFS HRES). In contrast, the spread in the maximum wind speed evolution is far greater. At the time of peak wind speed in both analyses, 00 UTC 2 November (48-h lead time), the value in the IFS analysis is 34 m s^{-1}. The IFS HRES forecast predicts this well (36 m s^{-1}), while the other, generally slightly coarser resolution, NWP models mostly forecast slightly weaker winds (30-37 m s^{-1}) with the NCEP model being a clear outlier, predicting winds of 40 m s^{-1}. The wind speeds forecast by the ML models are far too weak (25–26 m s^{-1}), even in comparison with the analysis from ERA5 (30 m s^{-1}). The ML models failed to capture the rapid intensification of the winds after about 06 UTC on 1 November (30-h lead time). Forecasts made using the ERA5 analysis system do not suffer from this low wind bias and so the underestimation is unlikely to be the result of training the ML models on the ERA5 data. The economic loss resulting from strong surface winds is often assumed to scale as the cube of normalised wind gust speed over a threshold (such as the 98th percentile value)[36], so even a small underestimation in predicted wind speed can be significant in terms of the subsequent losses.

The differences in maximum 10-m wind speed are explored further in Fig. 4 which shows maps of the 10-m wind speed and MSLP for ERA5, the IFS HRES forecast and the four ML-models valid at 00 UTC 2 November, the time of peak wind speed in both analyses and when the strong winds made landfall in France. All the forecasts were initialised 48 h prior to this time (as for the data shown in Fig. 3). These maps can be compared directly with the IFS analysis fields shown in Fig. 2a. The region of strong winds is located in an arc in the region of the tight MSLP gradient in the southwest quadrant of the Ciarán in all seven maps. However, the ML models fail to predict the strongest winds in a band following the isobars (contours of constant MSLP) in the region of the tightest MSLP gradient, as is seen in the IFS HRES forecast, ERA5 and the IFS analysis. It is notable that, despite all the ML models being trained on ERA5, they fail to capture the structure and magnitude of the winds in ERA5 (including in forecasts made using the ERA5 system, as shown in Supp. Fig. 1a) for this storm, implying that the far weaker winds found for the ML models compared to the NWP forecasts and IFS analysis are not simply a consequence of them being trained on a coarser resolution dataset. Note the NWP models used in Fig. 2 have a similar resolution to ERA5 (equivalent grid spacings of ERA5 ~ 31 km, Met Office ~20 km, JMA ~ 27 km, NCEP ~ 25 km) with the exception of the IFS (~ 9 km).

Dynamical structure of Storm Ciarán

In this section, we evaluate the dynamics of Storm Ciarán during the final stage of its rapid development with a focus on the formation of strong winds at low altitudes. We compare the predictive capability of the ML models by comparing with the IFS HRES forecast and ERA5. The ML models are all trained on ERA5 and have the same resolution as the output provided for ERA5, allowing a fair comparison of model performance. The forecasts are all initialised at 00 UTC on 1 November, during the onset of Ciarán's rapid intensification phase (Fig. 3a). They were evaluated 18 h later (Fig. 5) and 24 h later (Fig. 6), when Storm Ciarán's peak wind speeds were observed. By
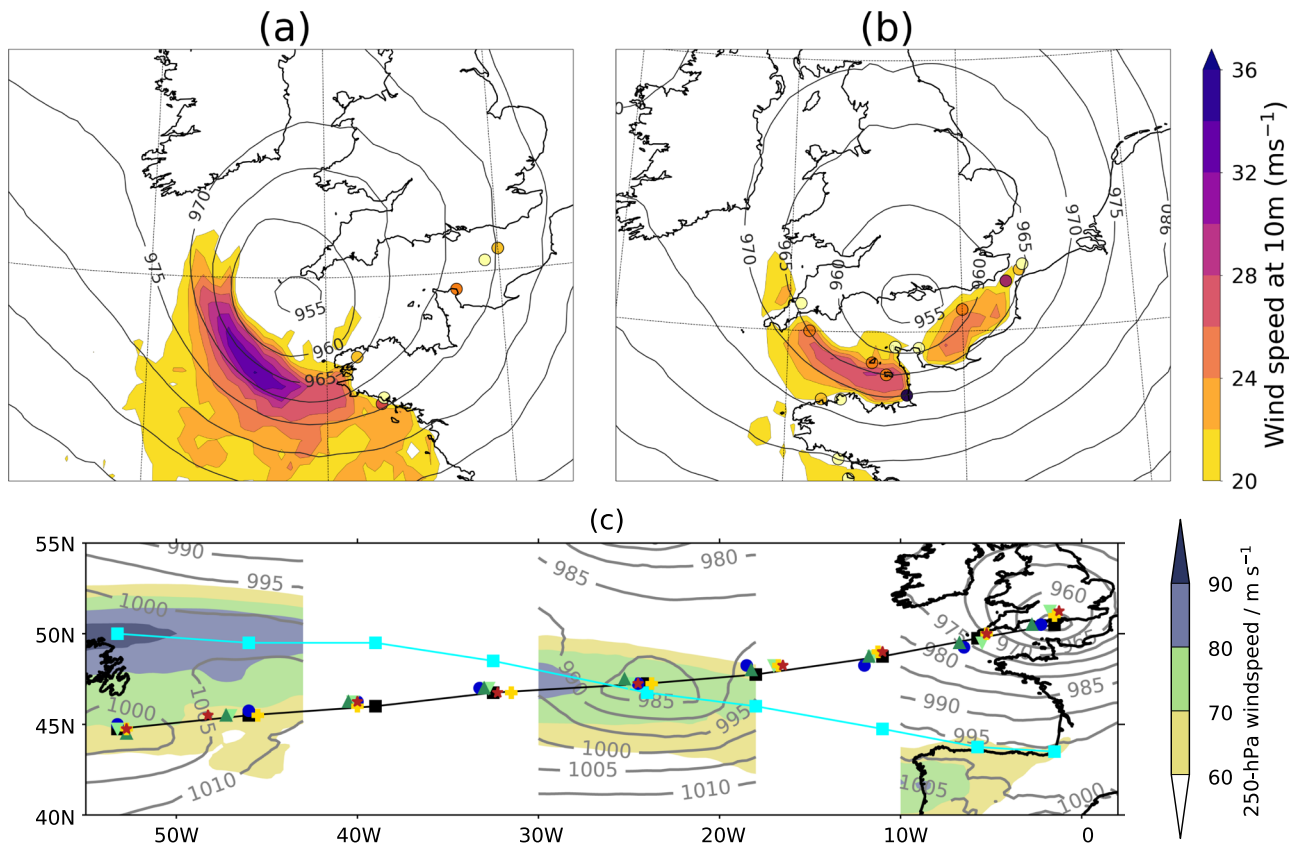
Fig. 2 | Near-surface wind and MSLP structure on landfall and track of Storm Ciarán. a, b Maps of 10-m wind speed (shading) and MSLP (contours) at **a** 00 UTC and **b** 06 UTC 2 November 2023 from the IFS analysis. Synoptic wind observations above 20 m s⁻¹ are shown as coloured dots. **c** Six-hourly track points from the IFS analysis (blacks squares joined by lines) and the IFS HRES forecasts and AI models (coloured symbols as in Fig. 3a, b) from 06 UTC 31 October to 06 UTC 06 UTC 2nd November 2023 (left to right) together with partial MSLP (grey contours in hPa) and 250-hPa wind speed (colour-filled contours) from the IFS analysis at 06 UTC on 31 October, 1 November and 2 November (left to right). The locations of the jet maxima at the longitude points of the MSLP minima at each time are indicated by the cyan squares connected by lines.

shifting the focus to these short lead times from the previous section, the aim is to highlight both the similarities in and differences between, the NWP and ML forecasts on timescales relevant for refining hazard warnings. To aid the reader, some key parts of the storm structure are labelled in Fig. 5a.

On 1 November 2023, Ciarán underwent significant intensification beneath the left exit region of an upper-level jet streak (Fig. 2c). All the ML models captured the position and extent of the upper-level jet streak accurately with the minimum MSLP associated with Ciarán beneath the left exit region at 18 UTC (Fig. 5a–f), a critical aspect of Storm Ciarán's dynamics.

There is also consensus among the ML models concerning the general shape of the cyclone. Figure 5a–f shows the position of the selected moist isentropes, chosen to indicate the frontal locations and, by their separation, the frontal strengths. The position of the warm sector, identified as the region inside the 285 K moist isentrope, is characterised as a hooked feature in ERA5. The shape of the warm sector is well captured by all the ML models. The cloud head, represented by 700-hPa relative humidity above 80% (grey shading), is seen wrapping around the poleward side of the cyclone centre in ERA5. The IFS HRES, Graphcast and PanguWeather forecasts accurately depict the shape of the cloud head; however, the cloud head in FourCastNet v2 appears less curved than the other forecasts. FourCastNet forecast does not output a humidity variable at 700 hPa.

Despite capturing the general shape of Storm Ciarán, there are noticeable differences in the strength of frontal structures, indicated by the gradient in wet-bulb potential temperature (how close together moist isentropes are). This is true for the cold front, denoted by the 285-K and 287.5-K moist isentropes to the southeast of the cyclone centre and also for the "bent-back front", i.e., the gradient between the 282.5-K and 285-K

moist isentropes that wrap around the cyclone centre on its northwestern side. To the southwest of the low centre the moist isentropes indicate the bent-back front diverge, and this is known as the frontal-fracture region. While all the ML forecasts include a frontal-fracture region, they struggle to resolve the sharp across-front temperature gradient to the west and southwest of the low centre.

Cross-frontal wind shear is another indicator of frontal strength, and the values of the vertical component of 850-hPa relative vorticity (green shading in Fig. 5a–f) near the bent-back front provide further evidence of the difficulties that the ML models have in simulating the frontal structures in the region. The hook-shaped narrow strip of high relative vorticity aligned with the bent-back warm front that is present in both IFS HRES and ERA5 (with maximum values up to 9×10^{-4} s⁻¹ and 7×10^{-4} s⁻¹, respectively) becomes broader and weaker in the ML models. The discrepancy between the ERA5 (and also the forecasts based on the ERA5 system, see Supp. Fig. 1b) and ML models in representing the sharpness of the bent-back front indicates that this shortcoming of the ML models is not solely due to model resolution.

This difference in frontal strength is particularly significant since it directly relates to the environment that can be conducive to the descent of a sting jet. Sting jets are coherent air flows that descend over a few hours from inside the tip of the cloud head at mid-tropospheric levels leading to a distinct mesoscale (perhaps 50–100 km across) region of near-surface stronger winds, and particularly gusts[37]. Among the models presented in this study, only the IFS forecasts and analysis have the necessary resolution to resolve a sting jet. It is crucial to recognise that the ML models, trained on coarse resolution data, are not equipped to discern features such as the
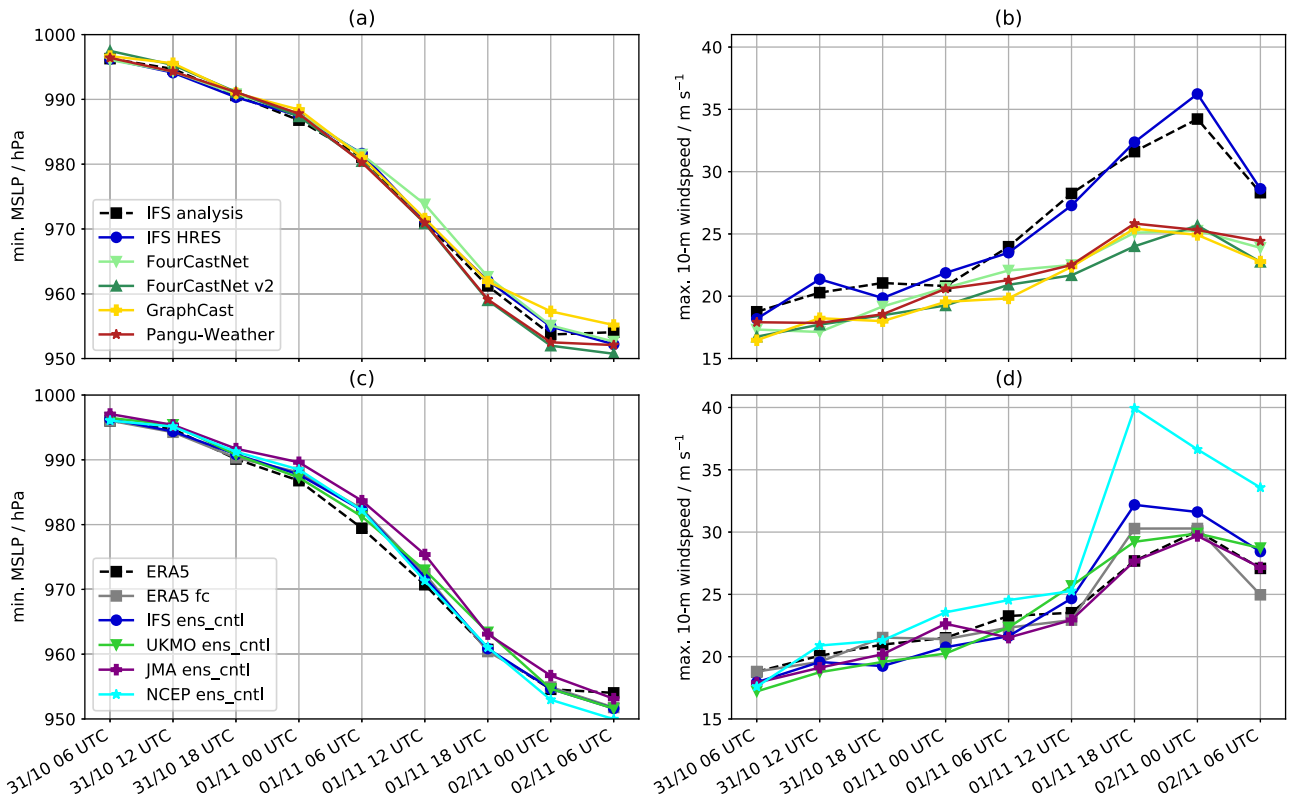
Fig. 3 | Timeseries of key properties of Storm Ciarán in different forecasts and analyses starting from 00 UTC 31 October 2023. a, c minimum MSLP and **b, d** maximum 10-m wind speed. **a, b** IFS analysis, IFS HRES forecast and forecasts from the four ML models. **a, d** ERA5, forecasts made from the ERA5 system and forecasts from control members of the ensemble forecasts from IFS (IFS ens_cntl), the Met Office (UKMO ens_cntl), the Japan Meteorological Agency (JMA ens_cntl) and National Centers for Environmental Prediction (NCEP ens_cntl). Note that the ECMWF HRES and IFS control members use the same model and resolution but are not bit-identical for technical computational reasons.
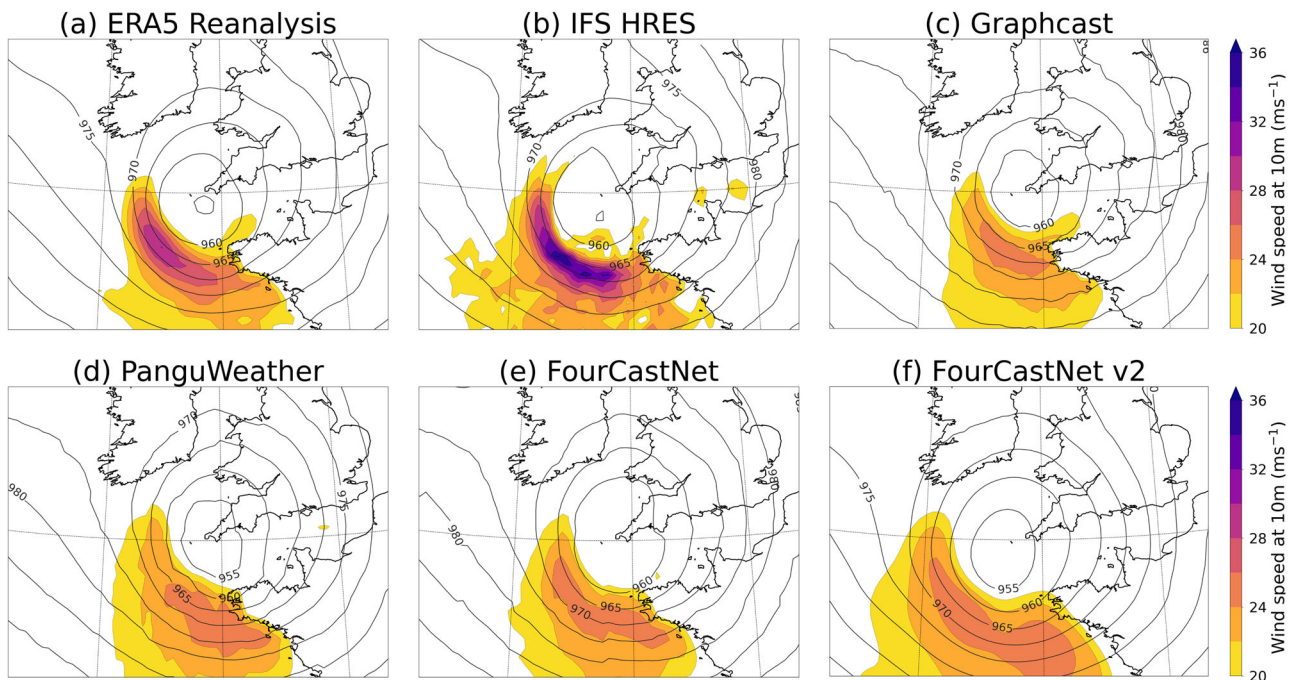

Fig. 4 | Near-surface wind and MSLP structure of Storm Ciarán at 00 UTC on 2 November 2023 from reanalysis and forecasts. Maps of 10-m wind speed (shading) and MSLP (contours) from **a** ERA5 and **b–f** forecasts, initialised at 00 UTC 31 October 2023, from the **b** IFS HRES model and **c–f** ML models, as labelled.
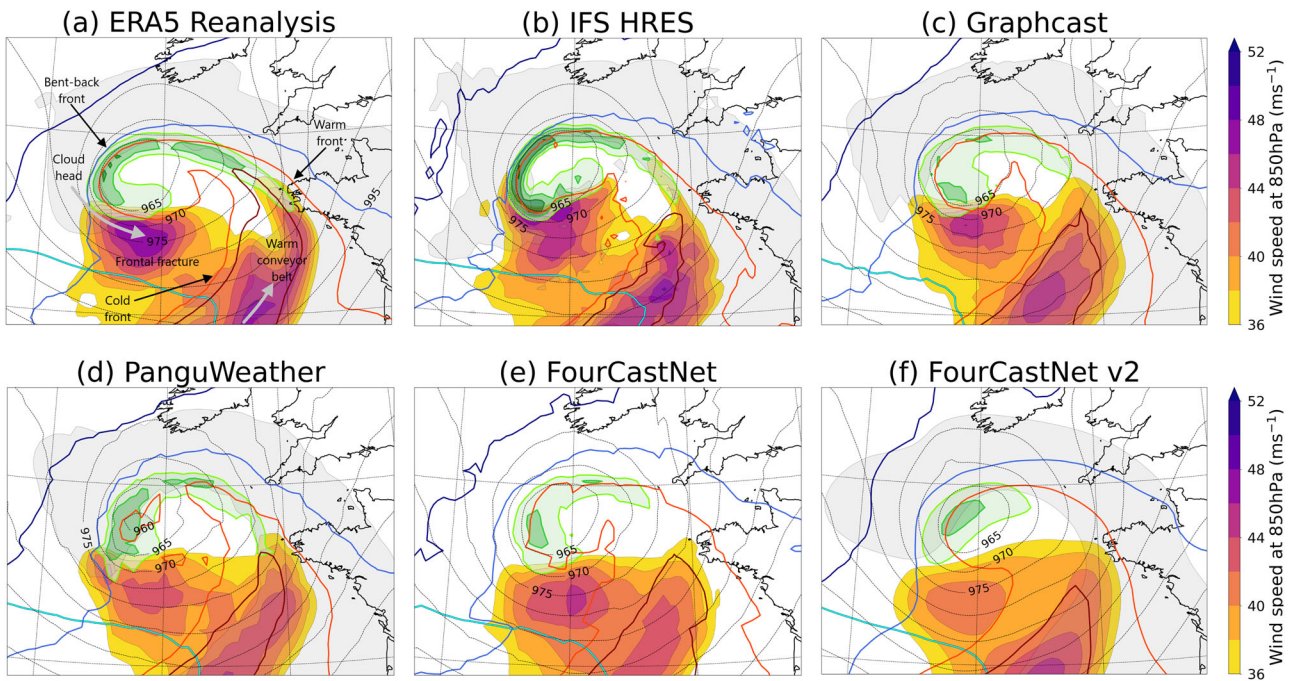
Fig. 5 | Dynamical structure of Storm Ciarán at 18 UTC on 1 November 2023 from reanalysis and forecasts. Maps show wind speed at 850 hPa (shading), wind speed at 250 hPa (65 ms^{-1}, cyan contour with high values in the bottom left of the panels), the wet-bulb potential temperature at 850 hPa (dark blue, light blue, light red and dark red contours indicating values increasing every 2.5 K from 280 K to 287.5 K), MSLP (thin grey contours), relative humidity with respect to water at 700 hPa (grey shading encircling regions above 80%, not shown for FourCastNet (**e**)

as not available), the vertical component of relative vorticity at 850 hPa (light-to-dark green shading, from 3×10^{-4} s^{-1} and then every 2×10^{-4} s^{-1}). **a** shows the structure in ERA5 while **b–f** show the structure from forecasts initialised at 00 UTC 1 November 2023. Note that the range of the wind speed colour bars in Figs. 5 and 6 is different to that in Figs. 2 and 4. The main features of the cyclone described in the text are annotated in (**a**).
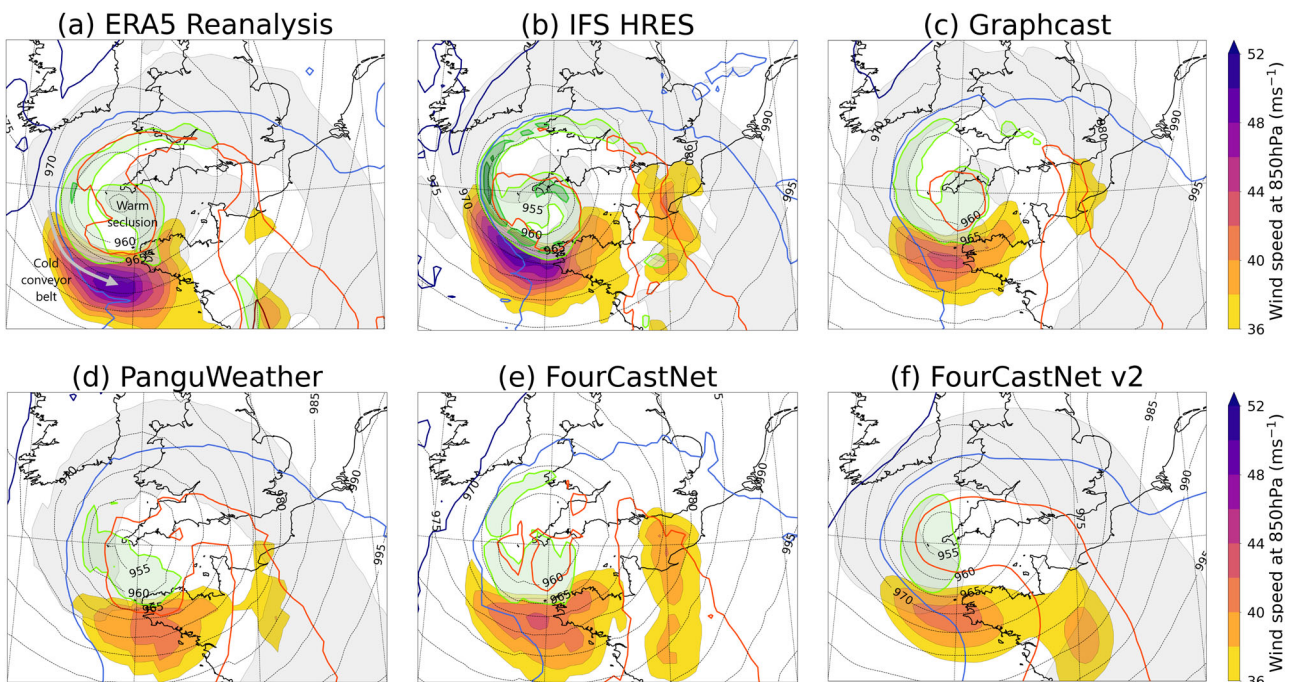
Fig. 6 | As Fig. 5 but for the dynamical structure of Storm Ciarán at 00 UTC on 2 November 2023. Contours of wind speed at 250 hPa and some of the contours of wet-bulb potential temperature at 850 hPa are not present as the associated values are not reached in the maps shown.

presence of mesoscale jets. This limitation highlights the importance of using models with adequate resolution when predicting high-impact weather phenomena occurring at small spatial scales. However, our analysis suggests that the ML models struggle to represent frontal structures conducive to

mesoscale high-impact features even when compared against NWP models with similar resolution, such as that used to generate ERA5.

The lack of sharpness of the bent-back warm front and cold front in the ML models impacts the strength of the wind speed maxima, as can be seen

by turning the focus of evaluation to the region of strong winds. Figure 5a–f shows the 850-hPa wind speed (filled contours) to give an indication of the lower-tropospheric storm structure that is less influenced by the presence of land below than the 10-m winds shown previously; consequently, winds at this pressure level, roughly a km above the ground, are normally stronger than those nearer the surface. All the ML models consistently identify two regions of strong winds: one in the frontal-fracture region and another in the warm sector. The strong winds situated in the frontal-fracture region are associated with the tight pressure gradient near the tip of the bent-back warm front, the associated descent and acceleration where the moist isentropes spread out, and the alignment with the direction of propagation of the storm (with a possible local enhancement due to sting-jet descent in the IFS HRES, see the small-scale areas above 46 ms^{-1}). The strong winds mostly in the core of the warm sector (enclosed by the 287.5-K moist isentropes) are associated with a broad jet, known as the warm conveyor belt jet, which ascends through the depth of the atmosphere from the top of the atmospheric boundary layer. The ability of the ML models to identify both the frontal-fracture and warm conveyor belt wind maxima (albeit with differences in the spatial structure and intensity of the latter) underscores their ability to accurately capture the general structure of extratropical cyclones. However, maximum wind speeds are weaker in the ML models than in ERA5, where they exceed 46 m s^{-1} (and even 48 m s^{-1} in the forecast based on the ERA5 system) in a broad region at the entrance of the frontal fracture. While Graphcast and FourCastNet display a small deficit of around 2 m s^{-1}, PanguWeather and FourCastNetv2 are roughly 4 m s^{-1} and 6 m s^{-1} lower, respectively.

We now turn our attention to the forecasts valid 6 h later, at 00UTC on 2 November 2023 (Fig. 6a–f), the time of peak ERA5 wind speeds. At this stage of Ciarán's evolution, analysis of the frontal-fracture region and the warm sector reveals several interesting features. ERA5 exhibits a region of warm air at the centre of the storm, which is separated from the main warm sector, a process known as warm core seclusion. Warm core seclusion occurs during the mature stage of extratropical cyclone development when cold air wraps around the low centre and cuts it off from the warm subtropical airmass. Relative vorticity starts decreasing along the bent-back front and generally increases near the cyclone centre as the front wraps around it. While the general evolution is captured by all models, the degree of clarity in the presence of a well-defined warm seclusion varies noticeably among the ML-models.

Focusing on the maximum winds in the frontal-fracture region, now compounded by the arrival of the cold conveyor belt (the main low-level jet in the cold sector, behind the cold front, of an extratropical cyclone), reveals that the ERA5, the forecast based on the ERA5 system (Supp. Fig. 1c) and the IFS forecast all have peak 850-hPa wind speeds between 48-50 m s^{-1}. GraphCast and FourCastNet exhibit peak wind speeds 4–6 m s^{-1} lower. PanguWeather and FourCastNet-v2 have a larger weak bias, with wind speeds underestimated by 6-8 m s^{-1}. Wind maxima are consistently underestimated in the ML models when compared to the benchmarks provided by the ERA5 (and its forecast) and the IFS forecast. This discrepancy in predicting wind maxima at the time of peak winds and as they approach land is crucial for assessing the potential impact of Storm Ciaran's surface winds and associated gusts.

By inspection, the structures of the MSLP fields in Figs. 5 and 6 are similar for the different models despite the differences in the wind speed structure and magnitude. This raises a further interesting question, is the discrepancy between the wind maxima in the conventional NWP and ML because the ML models do not reproduce the dynamical balances between the wind and pressure fields inherent in the conventional NWP models? This question is examined in detail in the Supplementary Material (and included Supp. Figs. 2–8) and a short summary is included here.

While the calculation of the geostrophic wind field (resulting from the balance of the pressure gradient and Coriolis forces) is relatively straightforward, calculating the more accurate gradient wind field (with the further inclusion of the centrifugal force associated with the curvature of a parcel trajectory) is more complex. Since both calculations require the evaluation

of horizontal gradients in the geopotential field, an unphysical lack of smoothness on the smallest scales in all of the ML models becomes easily apparent and should be further investigated. Note that while the gradient wind should provide a better approximation to the frictionless large-scale flow (where that flow is curved) than the geostrophic wind, the flow can still differ from gradient wind balance due to unbalanced motions which may be physically realistic, particularly in high-resolution model output (and reality).

Smoothed geostrophic and gradient wind fields have physically plausible structures in both NWP and ML model outputs. While the strongest full winds are found in the NWP model outputs even after smoothing, the strongest gradient winds are not clearly different between the ML and NWP models. The differences between the smoothed full and gradient wind fields for the NWP and ML models have similar characteristic structures and magnitudes in strong wind regions of the storm. Within the limitations of the accuracy of our calculations, we cannot conclude that the weak winds in the ML model forecasts are the result of an inability to resolve the proper dynamical balances, but are likely to instead be related to inadequacies in the geopotential field, i.e., in the gradient and curvature of the geopotential contours.

In summary, the ML models represent the large-scale dynamical drivers key to the development of Storm Ciarán well, including the position of the storm relative to the upper-level jet exit. They also accurately capture the larger synoptic-scale structure of the cyclone such as the position of the cloud head, the shape of the warm sector and the location of the warm conveyor belt jet. The ability of the ML models to resolve the more detailed structure of the storm is more mixed. Only some ML models correctly resolve the warm core seclusion and none of them capture the sharp bent-back warm frontal gradient. ML models underestimate the magnitude of the strongest winds at the surface and in the free atmosphere (above the boundary layer), particularly in the frontal-fracture region near the end of the bent-back front. Note that this underestimation of the strongest wind speeds is not a consequence of the resolution of the output of the ML models or their training data, since it also applies when comparing against the ERA5 (and forecasts based on the ERA5 system) and NWP models with resolution similar to ERA5.

Discussion
The contrasting ability of the four ML models considered to accurately forecast the large-scale dynamical properties of Storm Ciarán and its damaging winds serve to highlight the need for a more comprehensive assessment of this new and potentially transformative forecasting tool. More than 48 h before Storm Ciarán affected communities surrounding the English Channel, forecasts of the rapid MSLP deepening and track of the storms produced by the ML models were essentially indistinguishable from forecasts from an ensemble of conventional NWP models. Our analysis shows that the ML models were able to reproduce the upper-level flow that steered the developing storm into the left exit region of the jet and led to its rapid intensification. Many of the important dynamical features of the storm including the position of the cloud head, the shape of the warm sector and the location of cold and warm conveyor belt jets were also well captured by the ML models. The ML models do not seem to have been limited by the fact that a storm of comparable central pressure has never previously been observed over England during November. However, even in the relatively short ERA5 record, storms developing in a similar way with similar dynamical drivers (such as the upper-level jet) are common throughout winter and the ability of ML models to forecast more dynamically unusual storms, such as small-scale storms that develop rapidly from waves on pre-existing fronts, is an open question.

In contrast, when considering the damaging winds associated with Storm Ciarán in detail, forecasts from the ML models had significant errors and poorer performance than conventional NWP models. All four ML models failed to produce the narrow band of very strong winds at the surface that led to the most severe impacts, The ML models also failed to represent the strength of the cross-front thermal gradient in the bent-back front (a

feature also dynamically linked to strong winds) and had variable success in producing the warm seclusion of air that formed in the centre of the storm in its mature stage.

Much further work, considering other storms, is needed to assess if the biases apparent in the simulation of Storm Ciarán are a systematic feature of this first generation of ML models. Increased scrutiny of the models is likely to lead to the identification of target areas for model improvement, as it has done for NWP models. Since the ML models are available to all through public repositories, this scrutiny is likely to enable rapid model improvement. Detailed documentation of the performance of ML models will also be critical to weather forecasters seeking to make greater use of the ML models as part of the forecasting process. Forecasting centres like ECMWF are already beginning to develop and test alpha versions of ML models that complement their existing capabilities[38].

Based on a single case study, it would be premature to draw conclusions about the relative abilities of the four different approaches to ML weather forecasting exemplified by the different models. In particular, given we only had access to the 'small' version of the FourCastNet-v2 model it might be expected that this model would have a limited ability to produce the detailed properties of Storm Ciarán. Nonetheless, studies like ours are useful for identifying knowledge gaps in ML model development for forecasting, particularly in their ability to capture the structure of extreme weather patterns. This can be direct, such as the inclusion of necessary output variables (a 700-hPa humidity variable to identify the shape of the cloud head missing from the FourCastNet model), or through the formulation of more nuanced hypotheses for investigation. For instance, PanguWeather's ability to capture the vertical component of the 850-hPa relative vorticity could be due to the model integrating height information across levels. Similarly, GraphCast's ability to simulate the warm core seclusion better may be due to its multi-mesh representation rather than the spatial mixing used in the other models. All the ML models failed to capture the intense winds at the surface associated with Ciarán. As shown in the supplementary figures, wind errors in the lower troposphere are smaller. This may indicate that future models could benefit from including variables and the relationships across these variables that better characterise the planetary boundary layer in their training datasets. Most importantly, our analysis makes a strong case for a more robust evaluation of the forecasts from the ML models across all relevant spatio-temporal features of the physical phenomenon considered instead of isolated error metrics on individual variables.

The rapid acceleration of the forecasting capabilities of ML models as exemplified by our study of Storm Ciarán poses many new challenges and opportunities for atmospheric science[39]. Explainable AI (xAI) techniques[40,41] could be powerfully combined with the ML models we have considered to develop a deeper understanding of the reasons that they were able to produce skillful forecasts of Storm Ciarán in line with other attempts to unify ML and causal discovery methods[42]. The development of general-purpose, foundational models[43] could add further to the set of tools available to both forecast and understand high-impact weather events.

Methods

In this study, we compare forecasts produced by four different models based on machine learning methods. All are initialised from the same operational ECMWF analysis allowing a direct comparison with the current operational forecast of the ECMWF high-resolution model (CY48R1). The ML model forecasts are produced using the ai-models toolbox developed by ECMWF (https://github.com/ecmwf-lab/ai-models).

All four models considered are data-driven Deep Learning models and originally trained on a few (~ 4) decades worth of atmospheric and surface variables from the ERA5 dataset[44] at a resolution of 0.25°x 0.25° (~ 30 km resolution at the equator), which translates to 720 x 1440 grid cells. The ML models are all autoregressive, which means model output from a given time step can be used to predict output at the next time step. Model differences arise chiefly from the individual architectures, the selection of variables, parameterisations and training schemes briefly summarised in the technical details below:

- FourCastNet[12], uses the vision transformer (ViT) architecture with an Adaptive Fourier Neural Operator (AFNO)[45]. The AFNO enables dependencies across spatial and channel dimensions to be modelled efficiently at high resolutions where spatial token (feature) mixing occurs as a global convolution in the Fourier domain with FFTs. The model has a pre-training step where the AFNO is trained ahead on the ERA5 data with 20 different surface and atmospheric variables and then used for inference. The pre-training step learns mappings between $\mathbf{X}(t)$ and $\mathbf{X}(t + \Delta t)$ where t is a time step, Δt is a time increment (set to 6 h) and $\mathbf{X}$ is a tensor of features called patches. In the second fine-tuning or inference step, the pre-trained model is used to produce inferences from a defined state $\mathbf{X}(t)$, first for $\mathbf{X}(t + \Delta t)$ and this output from the models is itself then used to generate $\mathbf{X}(t + 2^* \Delta t)$ or the output for the second time step. Thus, while the training of the model is resource intensive, it is a one-time cost and the inference step is very fast.

- FourCastNet v2[13] is a development of the original FourCastNet model that uses Spherical Harmonics Neural Operators for modelling non-linear chaotic and dynamical systems on a sphere as opposed to flat Euclidean spaces. The model is trained with ERA5 data in a two-step process similar to FourCastNet—a single autoregressive step followed by fine-tuning. By learning global convolutions in computationally efficient manners, Fourier Neural Operators (such as those used in FourCastNet) are capable of accurately simulating long-range dependencies in spatio-temporal data. However, the Discrete Fourier Transform that FNOs rely on assumes a flat geometry, resulting in dissipation together with visual and spectral artefacts. The Spectral Fourier Neural Operators (SFNO), forming the basis for the FourCastNet v2 model architecture in its update from the FourCastNet model, in addition to having the desirable properties of FNOs also have translational or rotational equivariance. FourCastNet v2 is trained on a 73-channel subset of the ERA5 reanalysis dataset on single levels and pressure levels.

- Pangu-Weather[14] consists of four deep neural networks with different lead times (time between input and output) of 1 h, 3 h, 6 h and 24 h. 5 upper atmosphere and 4 surface variables at 13 different pressure levels were used to train the model with a combined total of 256 million parameters. The overall deep network architecture is called 3DEST or 3D-Earth-specific Transformer that integrates height information into a new dimension thus capturing relationships between atmospheric variables across pressure levels, unlike similar transformer-based models such as FourCastNet. Data is fed into the neural network and a process called patch embedding is used to downsample the input data from individual grid cells into a 3D cube. This cube is then put through an encoder-decoder based on a ViT called the Swin transformer[25] with 16 blocks. The positional bias in the Swin transformer is replaced with an Earth-specific positional bias to reflect the fact that in a 2D projection of a sphere, distances between neighbouring points are not the same across all latitudes. The decoder is symmetric to the encoder. Although the transformer-based neural network has a large training time similar to FourCastNet, this is partially improved in Pangu-Weather by the use of a hierarchical temporal aggregation scheme that reduces cumulative forecast errors and also the forecast generation time. This is done by employing the neural network with the largest lead time iteratively for a forecast so that neural networks with shorter lead times are used closer to the forecast. The height integration and aggregated forecast schemes are also considered improvements over other transformer-based architectures.

- GraphCast[15] is based on Graph Neural Networks (GNNs)[46] with around 36.7 million parameters. The model is trained with 5 surface and 6 atmospheric variables at 37 pressure levels resulting in 227 variables for every data point or grid cell. In the first step, the Encoder maps information from individual grid cells to nodes in a multi-mesh representation. The multi-mesh is derived as icosahedral meshes of increasing resolution from coarse (12 nodes) to fine (40,962 nodes). The second step has Processors using 16 GNN layers to propagate local

and long-range information across the nodes on the multi-mesh through message passing. Finally, the decoder uses a single GNN layer to map the final processor layer's multi-mesh representation back to the grid cells. GraphCast thus avoids the use of transformers and the associated scaling issues with higher resolutions that could result in large training times.

Numerical weather prediction model forecasts and analysis products

The ML model forecasts are compared to a set of forecasts from conventional numerical weather prediction (NWP) models to assess both systematic differences in the capabilities of the NWP and ML models and how the spread in the forecasts from the two architectures compare. Forecasts from the IFS HRES forecast and forecasts based on the ERA5 system (see below for a description of ERA5) were obtained from ECMWF and control (unperturbed) members of the ensemble forecasts for four models (the IFS[47], the Met Office[48], the Japan Meteorological Agency (JMA)[49], and the National Centres for Environmental Prediction (NCEP)[50]) were downloaded from the TIGGE archive[51] of operational global ensemble weather forecasts out to the medium range. The models chosen differ in their design and the resolution of the numerical model grid. Cycle 48r1 of the IFS was operational at the time of Storm Ciarán. Following the upgrade to this cycle in June 2023, the HRES and ensemble forecasts have the same resolution, equivalent to 9 km grid spacing. The Met Office, JMA and NCEP ensembles have grid spacings of approximately 20 km, 27 km and 25 km, respectively. The data for the four control ensemble members were all obtained regridded to a regular latitude-longitude grid of 0.5 degrees.

Analysis products

Both sets of models are compared to two analysis products (optimal blends of short-range forecasts and observations): the operational IFS analysis and ERA5[44]. The operational IFS analysis is produced using the IFS HRES forecast and has a resolution equivalent to 9 km grid spacing, ERA5 has a resolution equivalent to 31 km grid spacing. The IFS analysis and ERA5 were regridded to a regular latitude-longitude grid of 0.25 degrees.

ERA5 is used as an additional measure of forecasts 'truth' because the ML models all used ERA5 as their training data. Hence comparison with ERA5 indicates the skilfulness of these models relative to the best possible forecast given their training data. It is to be expected that the IFS analysis will include smaller-scale and higher amplitude weather features than ERA5 due to the use of a higher resolution model, despite being regridded to the same grid. It is also expected that the IFS analysis will be closer to the "truth" due to the use of higher resolution and an upgraded modelling system.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Code availability

Apart from the Python packages referenced in the Acknowledgements, the underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

References

1. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).
2. Krasnopolsky, V., Breaker, L. & Gemmill, W. A neural network as a nonlinear transfer function model for retrieving surface wind speeds from the special sensor microwave imager. *J. Geophys. Res.: Oceans* **100**, 11033–11045 (1995).
3. Brajard, J., Jamet, C., Moulin, C. & Thiria, S. Use of a neuro-variational inversion for retrieving oceanic and atmospheric constituents from satellite ocean colour sensor: Application to absorbing aerosols. *Neural Netw.* **19**, 178–185 (2006).
4. Brenowitz, N. D. & Bretherton, C. S. Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.* **45**, 6289–6298 (2018).
5. Jones, N. How machine learning could help to improve climate forecasts. *Nature* **548**, 379–380 (2017).
6. Rasp, S. & Lerch, S. Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.* **146**, 3885–3900 (2018).
7. Grönquist, P. et al. Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. R. Soc. A* **379**, 20200092 (2021).
8. Chantry, M., Christensen, H., Dueben, P. & Palmer, T. Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI. *Philos. Trans. R. Soc. A* **379**, 20200083 (2021).
9. Rasp, S. et al. WeatherBench: a benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.* **12**, e2020MS002203 (2020).
10. Dueben, P. et al. *Machine Learning at ECMWF: a Roadmap for the Next 10 Years*. (European Centre for Medium-Range Weather Forecasts, 2021).
11. Krishnamurti, T. et al. Improved skill for the anomaly correlation of geopotential heights at 500 hPa. *Month. Weather Rev.* **131**, 1082–1102 (2003).
12. Pathak, J. et al. Fourcastnet: a global data-driven high-resolution weather model using adaptive fourier neural operators. https://arxiv.org/abs/2202.11214 (2022).
13. Bonev, B. et al. Spherical fourier neural operators: learning stable dynamics on the sphere. https://arxiv.org/abs/2306.03838 (2023).
14. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
15. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
16. Andrychowicz, M. et al. Deep learning for day forecasts from sparse observations. https://arxiv.org/abs/2306.06079 (2023).
17. Weyn, J. A., Durran, D. R., Caruana, R. & Cresswell-Clay, N. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Modeling Earth Syst.* **13**, e2021MS002502 (2021).
18. Li, L., Carver, R., Lopez-Gomez, I., Sha, F. & Anderson, J. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances.* **10**, eadk4489 (2024).
19. Price, I. et al. GenCast: Diffusion-based ensemble forecasting for medium-range weather. https://arxiv.org/abs/2312.15796 (2023).
20. Magnusson, L. in *ECMWF Newsletter* https://alderney-elec.com/weather-station/ (ECMWF, 2023).
21. Priestley, M. D., Dacre, H. F., Shaffrey, L. C., Hodges, K. I. & Pinto, J. G. The role of serial European windstorm clustering for extreme seasonal losses as determined from multi-centennial simulations of high-resolution global climate model data. *Nat. Hazards Earth Syst. Sci.* **18**, 2991–3006 (2018).
22. Little, A. S., Priestley, M. D. & Catto, J. L. Future increased risk from extratropical windstorms in northern Europe. *Nat. Communications* **14**, 4434 (2023).
23. Zscheischler, J. et al. A typology of compound weather and climate events. *Nat. Rev. Earth Environ.* **1**, 333–347 (2020).
24. Shepherd, T. G. et al. Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. *Clim. change* **151**, 555–571 (2018).
25. Sanders, F. & Gyakum, J. R. Synoptic-dynamic climatology of the "bomb". *Mon. Weather Rev.* **108**, 1589–1606 (1980).
26. Kendon, M. *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.pdf (2023).

27. Shapiro, M. A. & Keyser, D. *Fronts, Jet Streams and the Tropopause*. (Springer, 1990).

28. Gray, S. L. & Volonté, A. *Why Storm Ciarán's Winds Were So Strong*. https://theconversation.com/why-storm-ciarans-winds-were-so-strong-216923 (2023).

29. Barroux, R. *Storm Ciaran: A Night of Destruction and Howling Gales In Brittany*. https://www.lemonde.fr/en/climate-change/article/2023/11/02/storm-ciaran-a-night-of-destruction-and-howling-gales-in-brittany_6222086_152.html (2023).

30. Meteociel.fr. *Tableaux d'observations pour Pointe du Raz*, https://www.meteociel.fr/temps-reel/obs_villes.php?jour2=2&mois2=10&annee2=2023&code2=7103 (2023).

31. France 24. *At Least 16 Dead as Storm Ciaran pummels Western Europe*. https://www.france24.com/en/live-news/20231103-ten-dead-as-storm-ciaran-batters-western-europe (2023).

32. Henley, J. *Storm Ciarán: Seven People Killed and Dozens Injured in Europe*, https://www.theguardian.com/world/2023/nov/02/storm-ciaran-people-killed-injured-storm-ciaran-batters-europe-wind-rain (2023).

33. Henshell, R. *No Trains in Five French Regions and Airport Closes in Brest*, https://www.connexionfrance.com/article/French-news/No-trains-in-five-French-regions-and-airport-closes-in-Brest (2023).

34. Alderney Electricity Limited. Alderney weather stations. https://alderney-elec.com/weather-station/ (2023).

35. Horton, S. *Key Severe Damage Indicators along the Track of the November Jersey Tornado 2023*. (Tornado and Storm Research Organisation, 2023).

36. Klawa, M. & Ulbrich, U. A model for the estimation of storm losses and the identification of severe winter storms in Germany. *Nat. Hazards Earth Syst. Sci.* **3**, 725–732 (2003).

37. Clark, P. A. & Gray, S. L. Sting jets in extratropical cyclones: a review. *Q. J. R. Meteorol. Soc.* **144**, 943–969 (2018).

38. AIFS Team. *ECMWF Unveils Alpha Version of New ML Model*, https://www.ecmwf.int/en/about/media-centre/aifs-blog/2023/ECMWF-unveils-alpha-version-of-new-ML-model (2023).

39. Schultz, M. G. et al. Can deep learning beat numerical weather prediction? *Philos. Trans. R. Soc. A* **379**, 20200097 (2021).

40. Arrieta, A. B. et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).

41. Samek, W. & Müller, K.-R. Towards explainable artificial intelligence. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* 5–22 (Springer, 2019).

42. Camps-Valls, G. et al. Discovering causal relations and equations from data. *Physics Reports.* **1044,** 1–68 (2023).

43. Bommasani, R. et al. On the opportunities and risks of foundation models. https://arxiv.org/abs/2312.15796 (2021).

44. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

45. Guibas, J. et al. Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators. *International Conference on Learning Representations* (2022).

46. Battaglia, P. W. et al. Relational inductive biases, deep learning, and graph networks. https://arxiv.org/abs/1806.01261 (2018).

47. Wedi, N. et al. *The Modelling Infrastructure of the Integrated Forecasting System: Recent advances and Future Challenges*. (European Centre for Medium-Range Weather Forecasts, 2015).

48. Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B. & Beare, S. E. The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.: A J. Atmos. Sci. Appl. Meteorol. Phys. Oceanogr.* **134**, 703–722 (2008).

49. Yamaguchi, H. et al. Introduction to JMA's new Global Ensemble Prediction System. *CAS/JSC WGNE Res. Activities Atmos. Ocean. Model.* **42**, 6.13–16.14 (2018).

50. Zhou, X. et al. The development of the NCEP global ensemble forecast system version 12. *Weather Forecast.* **37**, 1069–1084 (2022).

51. Bougeault, P. et al. The THORPEX interactive grand global ensemble. *Bull. Am. Meteorol. Soc.* **91**, 1059–1072 (2010).

52. MetOffice. *LAND SYNOP Reports from Land Stations Collected by the Met Office MetDB System*. https://catalogue.ceda.ac.uk/uuid/9f80d42106ba708f92ada730ba321831 (2008).

53. MetOffice. *SHIP SYNOP Reports from Ship, Buoy and Fixed Platform Stations Collected by the Met Office MetDB System*. https://catalogue.ceda.ac.uk/uuid/65ca7898647cc3686492bcb8bb483a1c (2008).

54. National Weather Service. *Sample Station Plot*. https://www.wpc.ncep.noaa.gov/html/stationplot.shtml (2022).

Author contributions

A.J.C.P. conceived the analysis and produced ML forecasts and initial characterisation of their performance. S.D., K.M.R.H., R.S. and R.V. contributed to the analysis and description of the ML models. N.J.H. led analysis of the properties of Storm Ciarán and its impacts. H.D., S.G., B.H. and A.V. produced the detailed dynamical analysis of the properties of Storm Ciarán in the forecasts including collecting NWP data and producing all figures. All authors contributed to discussions, writing, proofreading and editing the manuscript. All authors should be considered co-first authors as the work was completed collaboratively. The funders played no role in the study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-024-00638-w.

Correspondence and requests for materials should be addressed to Andrew J. Charlton-Perez.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.