

RESEARCH ARTICLE

Error in ERA5 2-m temperature identified using GraphCast

Hannah M. Christensen¹  | Jack Barker¹  | Bobby Antonio¹  |
Massimo Bonavita²  | Mohamed Dahoui²  | Patricia de Rosnay² 

¹Department of Physics, University of Oxford, Oxford, UK

²European Centre for Medium-Range Weather Forecasts, Reading, UK

Correspondence

Hannah M. Christensen, Department of Physics, University of Oxford, Oxford, UK.
Email:

hannah.christensen@physics.ox.ac.uk

Funding information

UK Research and Innovation,
Grant/Award Number: 10049639;
Leverhulme Trust; HORIZON EUROPE
Framework Programme, Grant/Award
Number: 101081383

Abstract

Reanalyses such as the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis Version 5 (ERA5) have long been foundational for weather and climate science. They have also found a new use case, as training and verification data for machine-learned weather prediction (MLWP) models. Here we compare short-lead-time (6 h) forecasts from the MLWP model GraphCast against ERA5. In doing so, we identify a recurrent, spatially coherent error in 2-m temperature centred on the Ethiopian Highlands, which occurs predominantly at 0600 UTC. We find that the same error feature is present in other MLWP models trained on ERA5. We show that these error events arise from errors in ERA5 that are also present in the ECMWF operational analysis. They arise from the two-dimensional (2D) optimal interpolation procedure, when surface reports that are temporally displaced compared with the background forecast are assimilated. This produces spuriously warm analysis increments over Ethiopia on approximately 7% of dates at 0600 UTC across the reanalysis record. The spread from the ensemble of data assimilation partially flags these cases but is underdispersive. We assess the impact on a MLWP system trained on ERA5. While the MLWP model can largely ignore these unphysical error events, a small systematic degradation in forecast skill over the region is observed. We discuss the implications for using reanalysis as truth in machine-learning training and verification, and recommend simple changes to reduce such artefacts in future analyses.

KEYWORDS

2-m temperature, data assimilation, ERA5, GraphCast, machine-learned weather prediction (MLWP), reanalysis error

1 | INTRODUCTION

The European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis Version 5 (ERA5) is a globally consistent, high-resolution reconstruction of past atmospheric, land, and ocean-surface conditions from 1950 to the present (Hersbach *et al.*, 2020). It is one of

the most widely used climate datasets: the ERA5 reference article (Hersbach *et al.*, 2020) has received around 4000 citations per year since publication. ERA5 is used for a diverse range of applications: it is used to assess the severity of climate change by revealing long-term shifts in temperature, precipitation, and extreme events (Masson-Delmotte *et al.*, 2021); it advances physical

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 European Centre for Medium Range Weather Forecasts and The Author(s). *Quarterly Journal of the Royal Meteorological Society* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

climate science through analysis of large-scale circulation patterns, variability, and teleconnections (e.g., Kang *et al.*, 2024; O'Reilly *et al.*, 2024); it is used to provide initial conditions for model simulations (Roberts *et al.*, 2025); and, further, it serves as a benchmark, allowing comparison of model predictions with an observationally constrained dataset (e.g., Strømmen *et al.*, 2025).

Recent years have seen a new use case for ERA5 data: its use as training data for machine-learned weather prediction (MLWP) models (Rasp *et al.*, 2020, 2024). These machine-learned emulators are now positioned as powerful substitutes for physics-based numerical weather prediction (NWP) models (Bi *et al.*, 2023; Lam *et al.*, 2023). MLWP models can produce forecasts up to 1000 times faster compared with NWP (Pathak *et al.*, 2022), and they also show improved forecast skill compared with their dynamical counterparts (Alet *et al.*, 2025; Kochkov *et al.*, 2024; Lam *et al.*, 2023; Price *et al.*, 2024; Yuval *et al.*, 2024). Arguably, it is the availability of ERA5—including its spatio-temporal coverage, comprehensive set of variables, and open-source licence—that has made this machine-learning revolution for weather prediction possible (Rasp *et al.*, 2024).

While initial evaluation of MLWP models focused on their skill compared with conventional numerical weather prediction (NWP) models, subsequent studies have considered their physical behaviour. This includes whether they emulate well-known physical relationships such as geostrophic balance (Bonavita, 2024) and the Matsuno–Gill response to tropical heating perturbations (Hakim & Masanam, 2024). Other studies have considered Kelvin and Rossby waves (Jalan & Sukhatme, 2025), teleconnections (Diao & Barnes, 2025), extreme events (Baño-Medina *et al.*, 2025; Meng *et al.*, 2025; Olivetti & Messori, 2024), and error growth (Selz & Craig, 2023).

One approach to understand the behaviour of a model better is to study the errors it makes. For example, the Transpose–Atmospheric Model Intercomparison Project (AMIP) experiment (Williams *et al.*, 2013) used a reanalysis product to initialise climate models. Comparing their forecast errors over one- to two-day lead times highlighted fast-growing model errors that could be attributed to the parametrised physics. Similarly, Rodwell and Palmer (2007) used the error in short-term forecasts from within the data-assimilation cycle to rule out certain parameter perturbations as unphysical.

The original goal of our study was to analyse errors made by a MLWP model over very short lead times to learn about the physical consistency of the model. However, in studying the characteristics of fast errors, we identified a curious error pattern over the Ethiopian Highlands: on 7% of days, the forecasts validating at 0600 UTC have surface temperature fields that are 5 K colder than ERA5. We find that this error is not due to the MLWP method,

in this case GraphCast (Lam *et al.*, 2023), but rather is a feature of the verification dataset, ERA5. It arises because a cluster of surface observations in this region are provided at non-synoptic times. This leads to a large systematic offset between the observations and the background forecast. The observations involved are usually rejected by the quality-control threshold; when they are included, they lead to a jump in the analysis of the observed magnitude. The presence of this error in ERA5 provides a novel test for the physical behaviour of MLWP models trained using ERA5: can they learn to ignore this clearly unphysical feature in the training data?

This article is structured as follows. In Section 2, we present the ERA5 dataset and the GraphCast model and describe the error database we constructed for our analysis. In Section 3, we present the error and its characteristics. In Section 4, we discuss the cause of the error in ERA5. In Section 5, we discuss the implications for MLWP forecasts trained using reanalysis data such as ERA5 data. Finally, in Section 6, we reflect on the implications of the article and present some conclusions.

2 | METHODS

Several MLWP models have been developed in recent years using the ERA5 reanalysis as their main (or exclusive) training dataset. Among these MLWP models, we focus our attention on GraphCast (Lam *et al.*, 2023), which was developed for medium-range weather forecasting. It is an example of a broad class of MLWP models which use a graph-based neural network architecture, first introduced for meteorological forecast applications by Keisler (2022) and since widely adopted (e.g., Alexe *et al.*, 2024; Lang *et al.*, 2024).

GraphCast evolves the three-dimensional (3D) global state of atmospheric temperature, winds, humidity, and geopotential on 37 pressure levels, as well as the state of five surface fields. It can produce a 10-day forecast in under a minute on a single Google tensor processing unit (TPU). Predictions are made at a spatial resolution of 0.25 deg latitude/longitude (approximately 25 km) and with a six-hour timestep. It receives the two previous states of the atmosphere as input—in other words, the atmospheric state at 0 and –6 h is used to predict the state at +6 h. A forecast is then made by calling GraphCast iteratively to step forward the atmospheric state. GraphCast is trained using a mean squared error loss. Training is autoregressive out to lead times of three days, to improve the model's ability to predict the atmospheric evolution over several timesteps. The loss function does not weigh all variables equally: the single most important variable in the loss function is the surface 2-m temperature (2mT) field. This

variable receives a weighting in the loss function equal to that of an entire 3D variable (on 37 pressure levels) such as temperature (Lam *et al.*, 2023).

Similarly to other MLWP models, GraphCast is trained to emulate the training data, in this case ERA5 (Hersbach *et al.*, 2020) using data from 1979 to 2017. ERA5 is based on the data assimilation system of the ECMWF Integrated Forecasting System (IFS) Cy41r2, which became operational on March 8, 2016. The reanalysis is produced using a four-dimensional variational (4D-Var) data assimilation process, which blends observations taken over a period of 12 hours with short-range (background) forecasts from the previous analysis update. In 1979, these observations numbered approximately 0.75 million per day, increasing to around 24 million a day by 2019 (Hersbach *et al.*, 2020). ERA5 is produced at a horizontal resolution of 31 km, with 12-hour cycling from 9 pm to 9 am and vice versa, with hourly output saved. In addition to the main reanalysis, uncertainty information is generated by producing a lower resolution 10-member ensemble of 4D-Var reanalyses. This ensemble provides background-error covariance estimates for the high-resolution 4D-Var computation.

In this study we choose to focus on 2mT, because of the high weighting given to this field in the GraphCast loss function. This means that GraphCast assigns high importance to accurate forecasts of this field during training. Errors in this field therefore have particular significance when considering the possible limitations of GraphCast's forecasting ability. However, 2mT is not a prognostic variable in the IFS. It is diagnosed by interpolating between the temperature at the lowest model level and the temperature at the surface, using the same assumed profiles as in the surface flux calculations (ECMWF, 2024b). It is analysed (alongside 2-m relative humidity) through an intermittent two-dimensional optimal interpolation (2D-OI) scheme, which combines screen-level observations with the background forecast diagnosed from the 4D-Var short-range forecast as discussed above (ECMWF, 2024a).

As described in ECMWF (2024a), during 2D-OI the background forecast at each analysis time (0000, 0600, 1200, and 1800 UTC) is first interpolated horizontally to the location of the observations. There is usually a mismatch between the altitude of the interpolated forecast and the true observation, because of the smoothed representation of orography in the forecast model: the observation is corrected to account for this discrepancy using a fixed lapse rate of 4.5 K/km. A quality-control (QC) check is then carried out and observations are rejected if they differ from the background forecast by more than 7.5°C. Background departure points, ΔX_i , are calculated at each observation site, i , as the difference between each corrected observation and the interpolated background forecast. The final analysis increment, ΔX_p^a , at each model grid point

is computed as the weighted sum over the background departure points from the $N = 50$ closest sites:

$$\Delta X_p^a = \sum_{i=1}^N W_i \cdot \Delta X_i, \quad (1)$$

where the weighting, W_i , is computed using the assumed background and observation-error covariance matrices in a standard Kalman filter update. In the weighting computation, the horizontal structure function has a decorrelation length-scale set to 300 km, which results in a horizontal e-folding distance of 420 km, while the vertical structure function has a length-scale of 800 m. If a given site has no valid observation at the time of interest, the departure point corresponding to the nearest in time (and most recent) valid observation within a plus/minus three-hour window is used instead. The standard deviations of background and observation errors are taken to be 1.5 K and 2 K, respectively.

We initialise GraphCast from ERA5 every six hours between 0000 UTC on January 1, 2014 and 1200 UTC on December 31, 2016. For each initial condition, we produce forecasts for one time step, corresponding to a lead time of 6 hours. We compute the error in the GraphCast 2mT forecast as

$$\epsilon = 2mT_{\text{GraphCast}} - 2mT_{\text{ERA5}}. \quad (2)$$

The recent period 2014–2016 was chosen for our analysis to ensure the best observational data coverage and therefore constraint of ERA5. It also falls within the training period for GraphCast. This is to assess GraphCast's ability to emulate the behaviour of ERA5 as observed in the training data, as opposed to its ability to generalise to unseen data. Three years was found to be sufficient for robust analysis.

3 | ERRORS IN 2-M TEMPERATURE

To summarise the error database, we first performed an empirical orthogonal function (EOF) analysis of the six-hour GraphCast errors for forecasts validating at 0600 UTC. EOFs 3, 4, and 5 all showed that a strong component of 2mT error variability was localised over the Ethiopian Highlands in central and northern Ethiopia (Figure S1). This merited further investigation. We also performed the EOF analysis for forecasts validating at other times. Although they show some signal of an error component over Ethiopia, the signal is strongest at 0600 UTC.

Figure 1 (bottom row) shows a sequence of 0600 UTC error maps for six consecutive days in 2016. On January

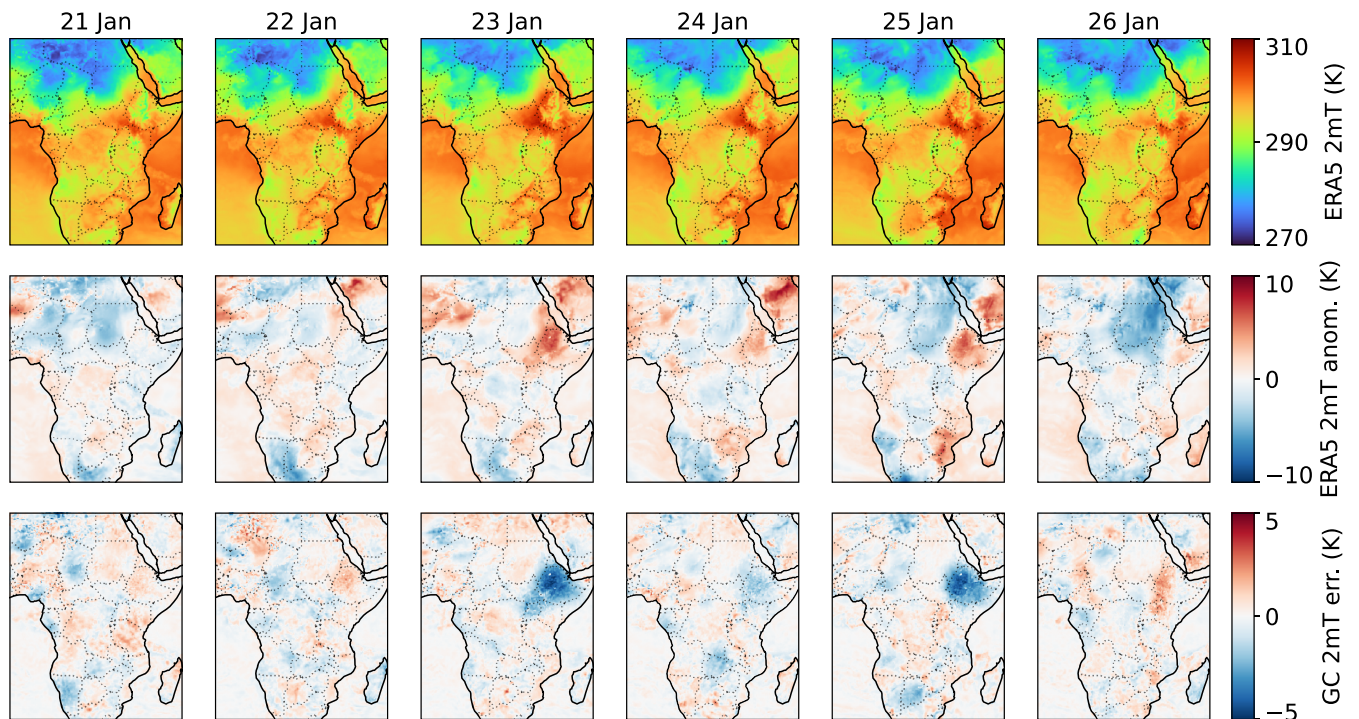


FIGURE 1 ERA5 2-m temperature (top row), ERA5 2-m temperature anomalies (centre row), and errors in 6-h lead-time GraphCast forecasts (bottom row), all shown at 0600 UTC for six consecutive days in 2016. ERA5 anomalies are computed by subtracting the 0600 UTC 2014–2016 January mean from the instantaneous fields. Error maps are defined as the GraphCast forecast validating at 0600 UTC minus the corresponding ERA5 field.

23 and 25, a large cold error of up to -5 K is observed in the GraphCast forecast. We call these “error events”. The error events are not observed on the other days in the period. The error is approximately centred on Addis Ababa, and is of order 1000 km in diameter. The top row of Figure 1 shows the ERA5 2mT fields at 0600 UTC for the same days. No unusual features are obvious on the days corresponding to the error events. The centre row shows ERA5 anomalies for the same days, computed by subtracting the 2014–2016 January 0600 UTC mean from each raw field. Here, we observe that ERA5 is warmer than average over Ethiopia on the days with the error events, though the anomaly is not out of the ordinary compared with anomaly features in other regions. For example, there are strong warm anomalies over both Saudi Arabia and Mozambique on certain days within the selected period, but these other warm anomalies do not correspond to an error feature in GraphCast forecasts.

We define a region of interest over Ethiopia spanning 7.5°N – 13.5°N , 35°E – 41°E and compute the average error in this region. Figure 2a shows a time series of this 0600 UTC error over the first half of 2016. The error events are clearly visible. We plot the ERA5 analysis and the GraphCast 6-h forecasts in Figure 2b, demonstrating that it is ERA5 and not GraphCast that shows abrupt jumps for these error cases. Figure 2c plots the same time

series for the ECMWF operational analysis and the corresponding background forecasts valid at 0600 UTC for each date, and the error in the operational analysis is included in Figure 2a. The ECMWF operational analysis shows similar behaviour to ERA5, while the corresponding operational background forecasts show similar behaviour to GraphCast.

There are two possible causes of the observed error. Firstly, the error could point to a problem with GraphCast. For example, there could be a short-lived weather phenomenon that forecasting models struggle to predict, but that is captured by the reanalysis. We note that, if this is the case, Figure 2c indicates the same problem in forecasts using the IFS. To identify what this phenomenon could be, we consider the distribution of variables in ERA5 on days with and without the error events. This includes skin reservoir content, total evaporation, cloud cover, various radiative fluxes, precipitation, and humidity and temperature at a range of vertical levels. We find no significant difference in distribution between error and non-error cases for any of these variables (not shown). We also assess whether related fields in ERA5 show similar fluctuations, including temperature on pressure levels between 900 and 500 hPa and skin temperature. These variables do not show the same unusual behaviour (Figure S2). We finally compared 6-h forecasts from two additional MLWP models with

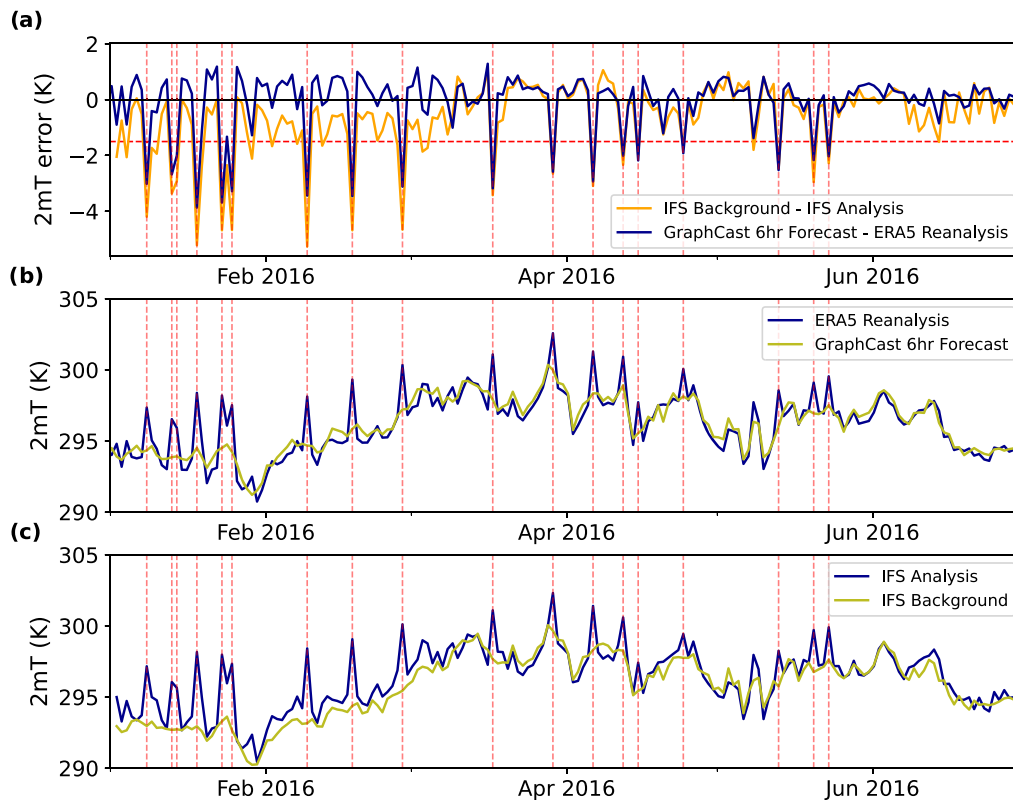


FIGURE 2 (a) Time series of the error in GraphCast compared with ERA5 reanalysis (dark blue), and the difference between the background forecast and the analysis from within the ECMWF operational data assimilation system (yellow). (b) GraphCast 6-h forecasts compared with ERA5 reanalysis. (c) ECMWF background forecasts and analysis from the operational data assimilation system. All panels show results for 2-m temperature, are averaged over 7.5°N – 13.5°N , 35°E – 41°E , and are valid at 0600 UTC on each date. The horizontal dashed line in panel (a) shows the threshold used to define error events in GraphCast compared with ERA5. The vertical dashed lines in each panel highlight the detected error cases.

TABLE 1 Percentage of ERA5 2-m temperature fields with error cases, as a function of time of day, for all dates within different reanalysis years. Here, we define error events as points in time for which the GraphCast error averaged over the rectangular region 7.5°N – 13.5°N , 35°E – 41°E is below -1.5 K .

	0000 UTC	0600 UTC	1200 UTC	1800 UTC
1980	1.1	8.0	0.3	0.0
1990	0.8	3.6	0.0	0.0
2000	0.3	0.5	0.0	0.0
2010	0.8	2.7	0.3	0.0
2014–2016	1.5	7.2	0.2	0.1

ERA5: FuXi (Chen *et al.*, 2023) and Pangu (Bi *et al.*, 2023). Results are shown in the Appendix A (Figure A1): FuXi and Pangu behave similarly to GraphCast and do not show large fluctuations in 2mT. We therefore conclude that the error is likely an issue with ERA5 that is also present in the operational analysis.

To compute statistics of the error, we define error events as points in time for which the GraphCast

error averaged over the rectangular region 7.5°N – 13.5°N , 35°E – 41°E is below -1.5 K . This is shown as the horizontal dashed line in Figure 2a. A visual comparison of the maps of errors on those days indicates that this approach was effective at selecting error events, with very few false positives or missed events. Table 1 shows the number of error events at each validation time for the 2014–2016 period. To assess whether these errors are a recent phenomenon, or whether they occur throughout the ERA5 record, we perform additional GraphCast forecasts to compute the statistics over 1980, 1990, 2000, and 2010, respectively. The results are shown in Table 1. The error cases are strongly confined to 0600 UTC and occur on around 7% of dates. They are also present throughout the historical record. Maps of example error cases in the earlier part of the record are shown in Figures S3 and S4. For earlier years, the error cases over Ethiopia have a similar spatial structure. However, many additional error features are visible, which could be worth investigating.

The ERA5 ensemble of data assimilation (EDA) aims to represent uncertainty in the analysis due to the observing system, boundary conditions, and forecast model. It

accounts primarily for random observation and model uncertainties, as opposed to systematic errors, but it can still provide information about relative uncertainty in the reanalysis. To assess whether the spread of the EDA correctly indicates higher uncertainty in the reanalysis for the error cases, we compute spread–error scatter plots (Leutbecher, 2010) for two point locations.

All dates in 2014–2016 are first sorted according to the EDA ensemble spread and grouped into 10 equally populated bins. Ideally, we would compare the EDA spread on each day with the reanalysis error, that is, the difference between the reanalysis and the real world, but this is not available. We are interested primarily in whether the EDA spread captures the uncertainty associated with very large error cases. We therefore use GraphCast as a proxy for the real world, since for these large error cases the difference between GraphCast and ERA5 will be dominated by the reanalysis error (as opposed to the forecast error). Therefore, for each bin, the EDA root-mean-square (RMS) spread is computed and compared with the RMS error in GraphCast forecasts for those corresponding dates. The results are shown in Figure 3. The two locations were chosen as points close to international airports, to benefit from the automated screen-level observations available there. The spread–error plot for Addis Ababa, within our region, shows a clear relationship between EDA spread and GraphCast error. More error cases are typically found in high-spread bins than in low-spread bins, giving rise to differences in RMSE across those bins: the EDA spread provides useful information about the possibility of large error cases. However, the ensemble is underdispersive. The spread is approximately half the observed RMSE and does not reach an amplitude that reflects the true error. In contrast, there is no clear relationship over Nairobi, which is outside our region of interest. Errors between GraphCast and ERA5 at this location are likely dominated by

random error growth, as opposed to uncertainties in the reanalysis.

Finally, we assess whether there is a systematically larger EDA spread associated with error cases compared with non-error cases. This can be answered by sorting the data by error instead of by spread and binning as above. If we do this, the largest error bin has an RMS spread of 0.51, compared with its RMSE of 3.0. The spread of all other bins is approximately 0.45 and shows no relationship with RMS error. Therefore there is a systematic, albeit modest, increase in spread for error cases compared with non-error cases.

4 | THE CAUSE OF THE ERROR

An analysis product represents the average weather conditions over the grid box, whereas a surface observation is a measurement of the local weather conditions. This philosophy underpins the 2D-OI approach. A mismatch between the background forecast and the observation is to be expected, because of the difference between the grid-box average and local weather conditions. The QC threshold is therefore set high to ensure that valid observations are retained and used to constrain the analysis. To represent the grid-box average better and reduce computational costs, the 2D-OI scheme uses the nearest 50 observations to constrain the analysis at any given point. Regions with complex orography, such as Ethiopia, are particularly challenging. ERA5 has a horizontal resolution of 31 km, which is not sufficient to resolve small-scale surface features or the resultant localised weather conditions. This means there is often a very substantial mismatch between a point observation and the corresponding first-guess forecast. Since microclimates can vary over short distances in such regions, pointwise observations

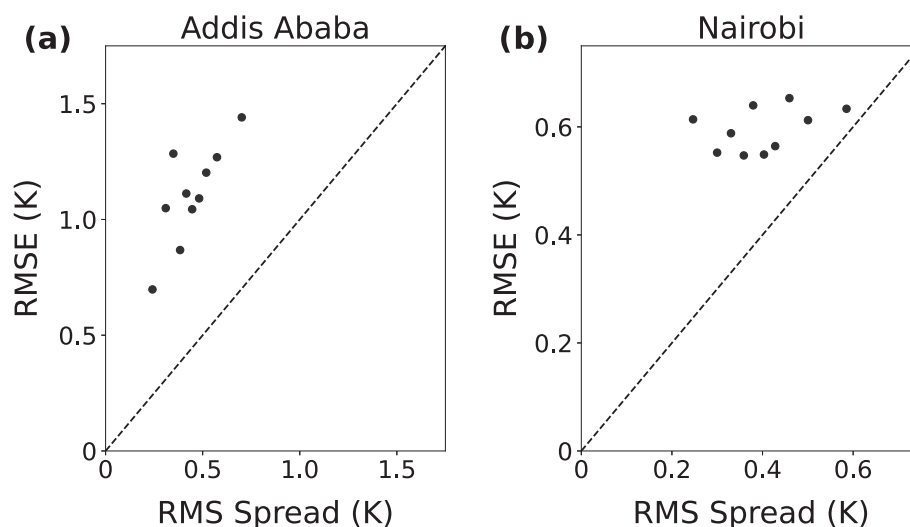


FIGURE 3 Root-mean-square (RMS) ensemble spread compared with RMS error for (a) Addis Ababa, Ethiopia (9°N, 38.75°E) and (b) Nairobi, Kenya (1.25°S, 37°E). The ensemble spread is taken from the ERA5 ensemble of data assimilation, whereas the error is computed between GraphCast and ERA5.

are very different from the grid-box average and averaging over many observations becomes critical.

A further challenge in this region is that surface observations are relatively sparse. At each analysis time, 20–30 observations are typically used to constrain the analysis across the Ethiopian region of interest; an area of approximately $700 \times 700 \text{ km}^2$. For comparison, Figure S5 shows a typical global distribution of surface observations assimilated into the 0600 UTC 2mT field, highlighting substantially more observations in other global regions. Because the nearest 50 observations are used to constrain the analysis at a given point, in regions where surface observations are sparse a single observation can influence the analysis over a very large region, including at locations with very different local weather effects and associated model biases.

Initially, our assumption was that the reanalysis errors were due to sparsity of observations in space; however, on further investigation, we identified that the ultimate source of the error is actually sparsity of observations in time. We found that a cluster of stations in Ethiopia have unusual reporting practices. These stations often miss reporting observations at 0600 UTC (9 am local time), but report instead at 0900 UTC (local noon). However, the 2D-OI scheme only has a background forecast valid at the synoptic time of 0600 UTC. Since a report at 0900 UTC falls within the acceptable plus/minus three-hour window, the background at 0600 UTC is compared with the raw measurement at 0900 UTC. The diurnal cycle shows very large increases in temperature over this window, so the observation at 0900 UTC is always substantially warmer than the background forecast at 0600 UTC and is usually QC rejected. However, occasionally it slips below the threshold, leading to a large jump in analysis temperature at 0600 UTC. Because a cluster of stations in the region has this reporting practice, and because of the overall sparsity of stations, this leads to a Gaussian-shaped uplift in the temperature over a large region centred on Ethiopia. In addition to a measurement at 0900 UTC, a measurement is also often available at 0300 UTC. However, the 2D-OI scheme always selects the nearest and *most recent* (i.e., later) observation, such that the 0900 UTC measurement is preferentially selected, and the errors always have a positive sign.

To demonstrate this mechanism, Figure 4 shows the available observations for the 0600 UTC analysis for all stations within the Ethiopian region for January 21–26, 2016, compared with the area-mean ERA5 reanalysis over the same domain. On January 21, 22, and 26, there are many observations at 0600 UTC with which to constrain the analysis. In contrast, relatively few stations report at 0600 UTC on January 23, 24, and 25. For each sta-

tion, the 2D-OI procedure therefore selects the nearest available observation in time at each station. For January 23 and 25, many stations report at 0900 UTC, data that are then used to constrain the 0600 UTC analysis, leading to the observed warm error. On January 24, only one station provided reports, at each of 0700, 0800, and 0900 UTC. The analysis at 0600 UTC was therefore constrained by the single closest observation, taken at 0700 UTC: the warm error was less severe on January 24, such that this date was not classified as an error event.

5 | IMPLICATIONS FOR MLWP

Reanalyses provide the backbone of training datasets for many modern AI weather prediction systems, including GraphCast. While they are among the most thoroughly evaluated geophysical datasets available, they are not free from errors or deficiencies. The identification, documentation, and correction of such issues are an integral part of the reanalysis production and quality-assurance process, with known issues routinely recorded and communicated through Copernicus Climate Change Service quality-control and user-support frameworks. In this section we explore the implications of reanalysis errors for MLWP trained on reanalysis.

Firstly, we assess whether the presence of the error in the training data has made forecasts in this region measurably worse compared with other regions. We hypothesise that GraphCast (and other MLWP models trained on reanalysis) might show systematically warmer 2-m temperature predictions over Ethiopia at 0600 UTC to hedge against the error. This would be unphysical, but would match the training distribution better.

We exclude all error cases from the database, leaving only forecasts where we are confident in the ERA5 validation field. Figure 5 shows the 2mT GraphCast bias at 0600 UTC. Points within our region of interest have a slight warm bias. When averaged over our $6^\circ \times 6^\circ$ region (rectangle marked on Figure 5), the warm bias is $+0.14 \text{ K}$. However, other regions over Africa have similar magnitude biases. To assess the significance of the bias in the region of interest, we randomly sample one million $6^\circ \times 6^\circ$ boxes, where the lower left-hand corners of these boxes are sampled from -90°N – 84°N and 0 – 354°E . Sample boxes containing any ocean points are discarded, before the average 0600 UTC bias is computed for each of the remaining $\sim 186\,000$ samples. Figure 6 shows the distribution of the resultant regionally averaged biases. GraphCast 2mT forecasts over land are, on average, slightly too cold at 0600 UTC. The Ethiopian region has an average bias of

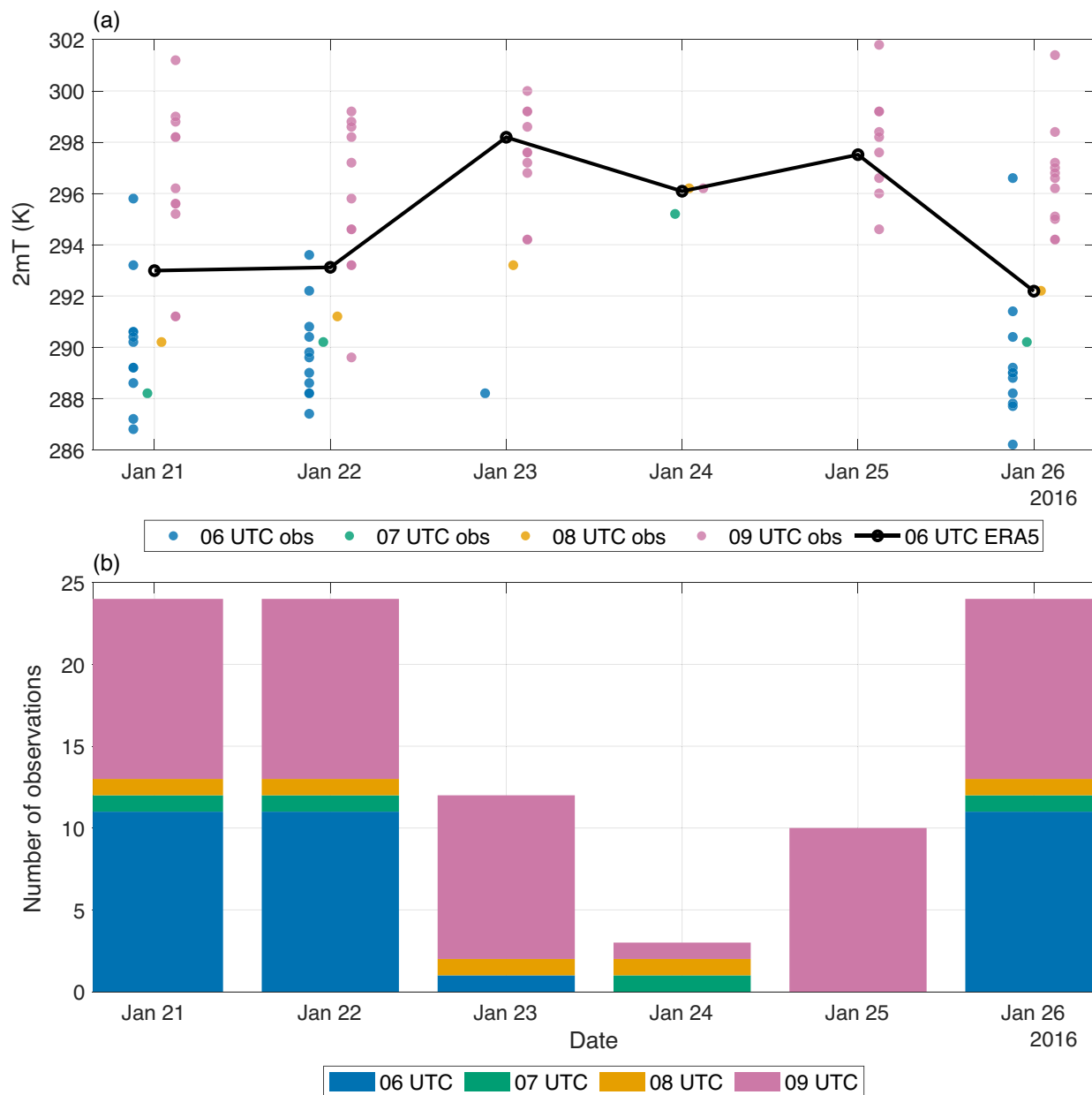


FIGURE 4 (a) Surface observations available to the 0600 UTC 2-m temperature analysis within the Ethiopian region (7.5°N – 13.5°N , 35°E – 41°E) for January 21–26, 2016. Individual observations are coloured according to their observation time (0600, 0700, 0800, or 0900 UTC), while the black line shows the corresponding area-mean ERA5 2-m temperature. (b) Number of observations available to the 0600 UTC analysis, separated by observation time (0600, 0700, 0800, and 0900 UTC).

0.14, which is 2.46σ from the mean. This corresponds to the 98.8th percentile, and is therefore a significant warm bias at the 5% level, though we note that it is within the distribution of biases for regions of this size. Furthermore, an average bias of -0.14 K falls at the 9.3rd percentile, such that biases of this *magnitude* are not significant.

Secondly, we assess whether the presence of an error case in the initial conditions impacts forecast skill systematically at subsequent times. We sort the data into

two groups according to the presence of an error case at 0600 UTC. We then compute the forecast bias at 1200 UTC for each group of forecasts, averaged over all initial conditions within the group, separately for each location. Figure 7a shows the distribution of biases for each point within the region of interest over Ethiopia, for the initial conditions with and without the error cases. If Graph-Cast were able to ignore the warm error in its starting conditions, the bias distributions would be indistinguishable on error and non-error days. Instead, the forecasts

FIGURE 5 Bias in 6-h GraphCast forecasts over Africa validating at 0600 UTC for non-error cases. The bias is computed over 2014–2016, compared with ERA5.

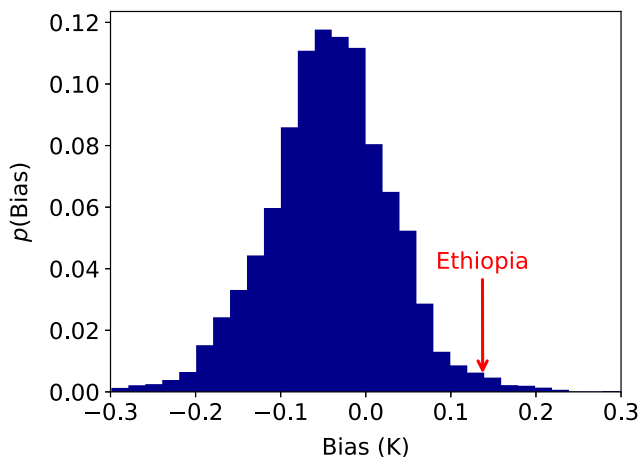
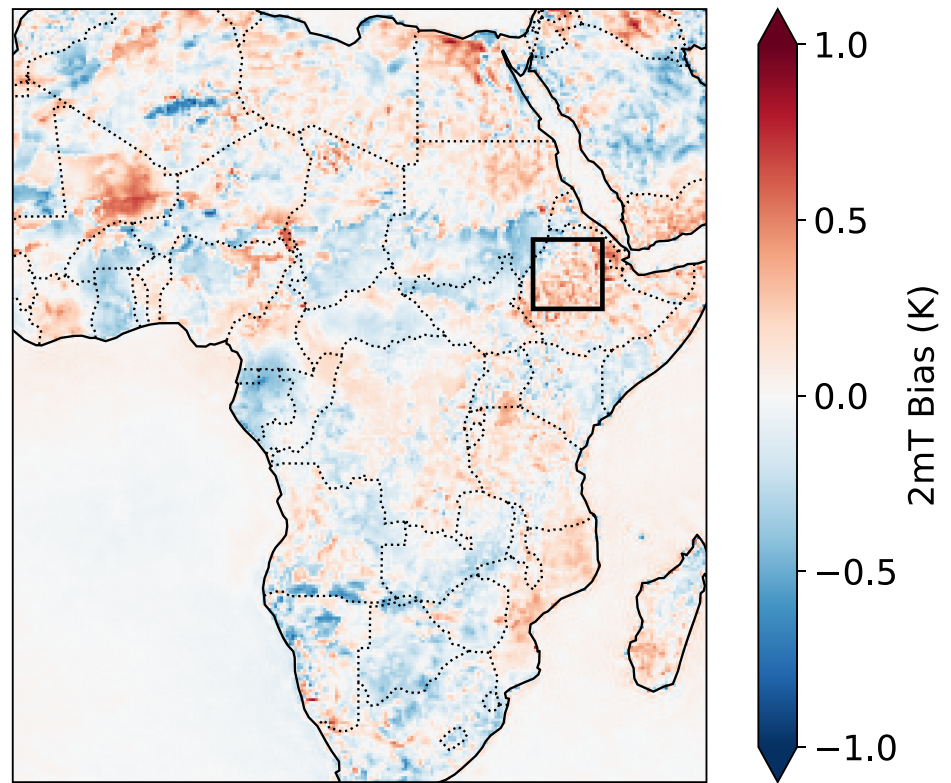


FIGURE 6 Distribution of regionally averaged 2-m temperature biases in 6-h GraphCast forecasts at 0600 UTC. The histogram contains only land regions sampled globally. The bias measured for the Ethiopian region is marked by the red arrow.

initialised with the error cases show a systematic cold bias (average of -0.15 K) at 1200 UTC compared with forecasts initialised without the error cases (average of -0.045 K). GraphCast forecasts are therefore systematically colder for initial conditions with a warm error. However, we note that, while there is a systematic shift, it is small compared with the size of the error in the initial conditions, which is typically approximately -3 K (see Figure 2). Figure 7b

shows the distribution of instantaneous errors, in both space and time, for these two groups of forecasts. This demonstrates further that the impact is small.

6 | DISCUSSION AND CONCLUSIONS

Here, we summarise and expand the principal conclusions of this study and then offer practical recommendations going forward.

6.1 | The error is significant because 2-m temperature is widely used

The 2mT field is one of the most widely downloaded ERA5 fields. It is commonly used for impact studies, climate-change trend analysis, and verification of weather and climate models. The error we identify over the Ethiopian highlands is pervasive throughout the dataset and occurs at 0600 UTC on an appreciable fraction of dates (approximately 7% in our sample). A limited-area EOF analysis over the African continent identified similar error features over Niger, Nigeria, Angola, Namibia, Congo, and Chad (Supplementary Figure S1). These regions also rely on relatively few, manual observations (Supplementary Figure S5), leaving them prone to this analysis error.

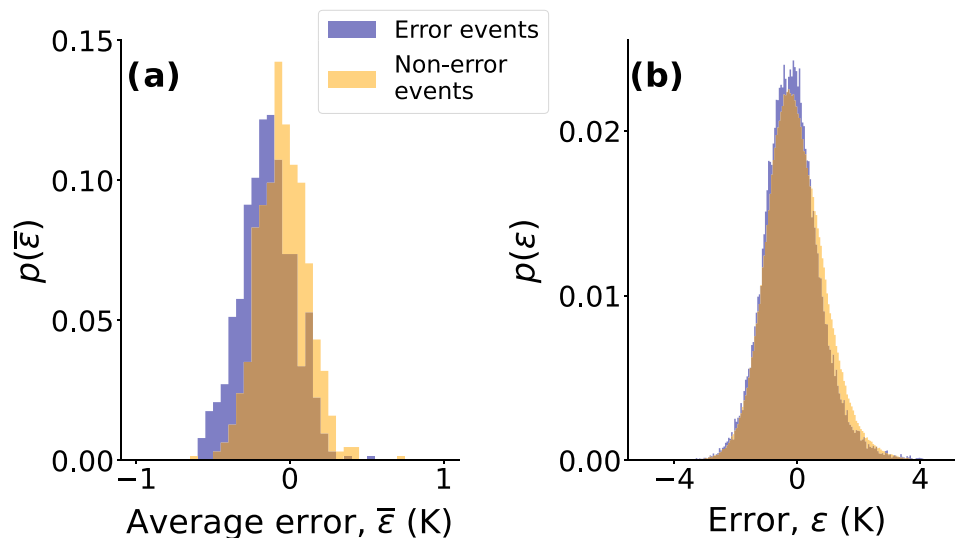


FIGURE 7 Distribution of 1200 UTC error in GraphCast forecasts initialised with (blue) and without (yellow) an error case in the initial conditions. (a) Bias distribution, defined as error averaged in time over all start dates, and (b) instantaneous errors in both space and time. Both show data for all locations within the Ethiopian region of interest.

Consideration of samples from 1980 and 1990 indicates that this may be a more widespread issue in the earlier part of ERA5, with more geographical regions affected. It is important to report such errors to users of ERA5, so they can account for them in any use of ERA5 for impacts or model evaluation.

6.2 | Uncertainty estimates from ERA5 should be utilised better

Reanalysis products such as ERA5 blend observations with background forecasts from a numerical model. They inherit key benefits, but also limitations, from both these data sources. ERA5 includes an ensemble of data assimilation, which produces an uncertainty estimate used within the data assimilation (DA) procedure and which accounts for both model and observation error (Hersbach *et al.*, 2020). This uncertainty estimate is also available to users of the reanalysis product. Despite its availability, this uncertainty estimate is not used routinely. For example, it was not included with WeatherBench (Rasp *et al.*, 2020, 2024) and is not widely used by the MLWP community. Where possible, users and machine-learning practitioners should make explicit use of the EDA (or other uncertainty estimates) to reduce the weighting assigned to uncertain analyses during training and verification, for example, through including analysis uncertainty in the loss function. Furthermore, a better understanding of the DA process could guide how MLWP training objectives are constructed. For example, weighting surface 2mT equally with the 3D prognostic variables benefitted GraphCast's overall skill, but surface fields may contain artefacts due to the 2D-OI algorithm and sparse surface observations, whereas 3D fields are constrained relatively well by 4DVar.

6.3 | MLWP in the Tropics requires caution because of observation sparsity

MLWP methods are often promoted as promising for tropical forecasting, where NWP models can have lower skill. However, our study indicates that caution must be taken. Tropical regions are often more sparsely observed than the extratropics, and complex terrain can amplify the impact of reporting practices or DA choices. However, the sparsity of surface observations means that end-to-end MLWP, which takes observations as input and predicts the forecast at station locations, would leave many users without a forecast. We must therefore assess the observing system and DA behaviour in regions of interest before relying on MLWP outputs for operational or impact use.

6.4 | MLWP can learn climatology shortcuts as well as physical relationships

Machine-learned emulators are, by construction, statistical approximators of their training distribution. When training data contain intermittent errors, ideally an emulator could learn to ignore those samples. However, in practice it is likely that the emulator will either learn to reproduce those errors or hedge against them by shifting their predictions toward the climatological mean. Both behaviours are undesirable. Our analysis shows that GraphCast exhibits a small but significant regional warm bias at 0600 UTC consistent with the training distribution, and initial conditions containing the ERA5 error produce slightly different subsequent forecasts. This is evidence that ML models do respond to, and partially internalise, such errors. However, the biases are small. In the context of MLWP, there is substantial discussion as to whether such

models have “learnt the underlying physics”. In our context, we phrase this as, “Can an MLWP model identify that, on certain days, the 2mT field is inconsistent with the remaining 3D temperature fields?” On the whole, our results show that GraphCast has largely ignored the error, despite its prevalence throughout the training data. We note that this is despite the fact that the error is not purely random—it is systematically more likely to occur on days with a smaller diurnal cycle. This hints that GraphCast has been able to identify and ignore this inconsistency to some extent.

6.5 | MLWP models can be valuable for assessing reanalysis quality

An unexpected positive outcome of this work is methodological: we found that comparing MLWP short-lead-time forecasts with their training analysis was an effective way to detect intermittent errors in the reanalysis. Because ML models learn the dominant spatio-temporal structure of their training data, sudden, localised discrepancies between a forecast and the analysis can highlight errors that would otherwise be difficult to detect by eye. We therefore recommend using MLWP verification as a complementary diagnostic when validating reanalysis products, for those for whom the background forecast is not available. While NWP models could also be used to detect such errors, the portability and speed of MLWP models makes them particularly suitable for this task.

6.6 | Practical recommendations for reanalysis practice

Finally, we consider recommendations for improving the analysis and reanalysis in this region, and for other locations in space and time that suffer from sparsity of observations and unusual reporting practices. In the 2D-OI system currently used to produce the 2mT analyses, we suggest that it is important to make better use of observations in time. The natural solution is to produce and use a background forecast that has a high time resolution, such that observations could be compared with the background that is valid at the time that the observation was made, as in 4DVar. In addition, instead of using only the nearest and most recent observation, it would be better (1) to use observations symmetrically in time—for example, using an observation at both 0300 UTC and 0900 UTC, should 0600 UTC be unavailable—and (2) moreover, to consider using all observations available within the 6-h window, suitably weighted according to their distance in time from

the target time. These two changes should enable the system to perform better in regions with a rapidly changing diurnal cycle, as it would no longer be systematically more likely to look ahead than behind to find a suitable observation. In addition, making use of more observations in time will reduce random errors associated with individual observation locations and also reduce jumpy behaviour.

A more fundamental solution would be to let 4D-Var produce the 2mT analyses. This would enable a more extensive and dynamically consistent use of all available 2mT observations within the assimilation window, alongside the large number of other observations that constrain near-surface conditions. Such a development would also support ongoing efforts towards coupled data assimilation. Since 2mT observations are already assimilated in 4D-Var, the approach is expected to be cost-neutral. However, achieving analyses that outperform the current high-resolution 2D-OI system remains challenging. Ongoing research is exploring improved observation-error specification and enhanced 4D-Var resolution to represent near-surface temperature variability better within a fully four-dimensional framework.

ACKNOWLEDGMENTS

H. M. Christensen and B. Antonio were supported through the EERIE project (Grant Agreement No 101081383), funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Climate Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. University of Oxford's contribution to EERIE is funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10049639). H. M. Christensen was also funded through a Leverhulme Trust Research Leadership Award.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The ERA5 datasets used in the current study are available in the Copernicus Climate Change Service, Climate Data Store (2023), for registered users. The following datasets were used: ERA5 hourly data on pressure levels from 1940 to the present (DOI: [10.24381/cds.bd0915c6](https://doi.org/10.24381/cds.bd0915c6)), ERA5 hourly data on single levels from 1940 to the present (DOI: [10.24381/cds.adbb2d47](https://doi.org/10.24381/cds.adbb2d47)). The IFS operational background and analysis datasets are taken from the long-window 4DVar dataset (stream=lwda) in the ECMWF MARS

archive; data for additional machine-learning model forecasts in 2022 are taken from the ECMWF experimental machine-learning datasets (class=ml) in the ECMWF MARS archive (ECMWF, 2018).

ORCID

Hannah M. Christensen  <https://orcid.org/0000-0001-8244-0218>

Jack Barker  <https://orcid.org/0009-0004-5007-3010>

Bobby Antonio  <https://orcid.org/0009-0000-0918-8327>

Massimo Bonavita  <https://orcid.org/0000-0002-0392-284X>

Mohamed Dahoui  <https://orcid.org/0000-0003-4037-7470>

Patricia de Rosnay  <https://orcid.org/0000-0002-7374-3820>

REFERENCES

- Alet, F., Price, I., El-Kadi, A., Masters, D., Markou, S., Andersson, T.R. et al. (2025) Skillful joint probabilistic weather forecasting from marginals. <https://arxiv.org/abs/2506.10772>
- Alexe, M., Boucher, E., Lean, P., Pinnington, E., Laloyaux, P., McNally, A. et al. (2024) Graphdop: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations. <https://arxiv.org/abs/2412.15687>
- Baño-Medina, J., Sengupta, A., Doyle, J.D., Reynolds, C.A., Watson-Parris, D. & Monache, L.D. (2025) Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia. *Climate and Atmospheric Science*, 8(1), 92. Available from: <https://doi.org/10.1038/s41612-025-00949-6>
- Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X. & Tian, Q. (2023) Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970), 533–538. Available from: <https://doi.org/10.1038/s41586-023-06185-3>
- Bonavita, M. (2024) On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12), e2023GL107377. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2023GL107377>
- Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y. et al. (2023) Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *Npj Climate and Atmospheric Science*, 6, 190. Available from: <https://doi.org/10.1038/s41612-023-00512-1>
- Diao, C. & Barnes, E.A. (2025) Assessing MJO tropical-extratropical teleconnections in deep learning weather prediction models. <https://doi.org/10.22541/essoar.174759219.91708232/v1>
- ECMWF. (2018) MARS – the ECMWF meteorological archive. <https://www.ecmwf.int/node/18124>
- ECMWF. (2024a) IFS documentation – Cy49r1: operational implementation, part II: data assimilation. IFS Documentation, operational implementation conducted on 12 November 2024.
- ECMWF. (2024b) IFS Documentation – Cy49r1: Operational implementation, part IV: physical processes. IFS Documentation, operational implementation conducted on 12 November 2024.
- Hakim, G.J. & Masanam, S. (2024) Dynamical Tests of a Deep Learning Weather Prediction Model. <https://doi.org/10.1175/AIES-D-23-0090.1>
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. Available from: <https://doi.org/10.1002/qj.3803>
- Jalan, S. & Sukhatme, J. (2025) Intraseasonal equatorial Kelvin and Rossby Waves in modern AI-ML models. <http://arxiv.org/abs/2507.07952>
- Kang, M.J., Kim, H. & Son, S.W. (2024) Qbo modulation of mjo teleconnections in the north pacific: impact of preceding mjo phases. *Npj Climate and Atmospheric Science*, 7, 12. Available from: <https://doi.org/10.1038/s41612-024-00565-w>
- Keisler, R. (2022) Forecasting global weather with graph neural networks. ArXiv:2202.07575 [physics] <http://arxiv.org/abs/2202.07575>
- Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G. et al. (2024) Neural general circulation models for weather and climate. *Nature*, 632(8027), 1060–1066. Available from: <https://doi.org/10.1038/s41586-024-07744-y>
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirsberger, P., Fortunato, M., Alet, F. et al. (2023) Learning skillful medium-range global weather forecasting. *Science*, 382(6677), 1416–1421. Available from: <https://doi.org/10.1126/science.adi2336>
- Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B. et al. (2024) AIFS – ECMWF's data-driven forecasting system. ArXiv:2406.01465 [physics] <http://arxiv.org/abs/2406.01465>
- Leutbecher, M. (2010) Diagnosis of ensemble forecasting systems. In: *Seminar on diagnosis of forecasting and data assimilation systems, 7–10 September 2009*. Shinfield Park, Reading: ECMWF, pp. 235–266.
- Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S. et al. (Eds.). (2021) *Climate change 2021: the physical science basis*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.
- Meng, Z., Hakim, G.J., Yang, W. & Vecchi, G.A. (2025) Deep Learning Atmospheric Models Reliably Simulate Out-of-Sample Land Heat and Cold Wave Frequencies. <http://arxiv.org/abs/2507.03176>
- Olivetti, L. & Messori, G. (2024) Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-weather, and GraphCast. *Geoscientific Model Development*, 17(21), 7915–7962. Available from: <https://doi.org/10.5194/gmd-17-7915-2024>
- O'Reilly, C.H., Drouard, M., Ayarzagüena, B., Ambaum, M.H.P. & Methven, J. (2024) The role of storm-track dynamics in the intraseasonal variability of the winter enso teleconnection to the north atlantic. *Quarterly Journal of the Royal Meteorological Society*, 150(761), 2069–2086. Available from: <https://doi.org/10.1002/qj.4691> April 2024, Part B.
- Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M. et al. (2022) FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. <https://arxiv.org/abs/2202.11214>
- Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D. et al. (2024) Probabilistic weather forecasting with machine learning. *Nature*, 637(8044), 84–90. Available from: <https://doi.org/10.1038/s41586-024-08252-9>
- Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S. & Thuerey, N. (2020) Weatherbench: a benchmark dataset for

- data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), e2020MS002203. Available from: <https://doi.org/10.1029/2020MS002203>
- Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T. et al. (2024) Weatherbench 2: a benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6), e2023MS004019. Available from: <https://doi.org/10.1029/2023MS004019>
- Roberts, M.J., Reed, K.A., Bao, Q., Barsugli, J.J., Camargo, S.J., Caron, L.P. et al. (2025) High-resolution model intercomparison project phase 2 (highresmp2) towards cmip7. *Geoscientific Model Development*, 18, 1307–1332. Available from: <https://doi.org/10.5194/gmd-18-1307-2025> ©Author(s) 2025. Distributed under the Creative Commons Attribution 4.0 License.
- Rodwell, M.J. & Palmer, T.N. (2007) Using numerical weather prediction to assess climate models, 133(622), 129–146.
- Selz, T. & Craig, G.C. (2023) Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophysical Research Letters*, 50(20), e2023GL105747. Available from: <https://doi.org/10.1029/2023GL105747>
- Strømmen, K., Michel, S.L.L. & Christensen, H.M. (2025) The link between gulf stream precipitation extremes and european blocking in general circulation models and the role of horizontal resolution. *Journal of Climate*, 38(8), 3763–3786. Available from: <https://doi.org/10.1175/JCLI-D-24-0311.1>
- Williams, K.D., Bodas-Salcedo, A., Déqué, M., Fermepin, S., Medeiros, B., Watanabe, M. et al. (2013) The transpose-AMIP II experiment and its application to the understanding of southern ocean cloud biases in climate models, 26(10), 3258–3274. Available from: <https://doi.org/10.1175/JCLI-D-12-00429.1>
- Yuval, J., Langmore, I., Kochkov, D. & Hoyer, S. (2024) Neural general circulation models optimized to predict satellite-based precipitation observations. <https://arxiv.org/abs/2412.11973>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Christensen, H.M., Barker, J., Antonio, B., Bonavita, M., Dahoui, M. & de Rosnay, P. (2026) Error in ERA5 2-m temperature identified using GraphCast. *Quarterly Journal of the Royal Meteorological Society*, e70309. Available from: <https://doi.org/10.1002/qj.70309>

APPENDIX A. BEHAVIOUR OF OTHER MACHINE-LEARNT WEATHER PREDICTION MODELS

To assess whether the behaviour identified for GraphCast is specific to that model or is also exhibited by other MLWP models trained on ERA5, we examined 6-h forecasts from the FuXi (Chen *et al.*, 2023) and Pangu (Bi *et al.*, 2023) models. All forecasts validate at 0600 UTC and were obtained from the MARS archive (ECMWF, 2018). Figure A1 shows time series of area-mean 2-m temperature over the Ethiopian region during 2022. All three MLWP models exhibit behaviour similar to GraphCast: their forecasts do not reproduce the intermittent warm spikes present in ERA5. This suggests that the behaviour identified for GraphCast is also exhibited by these two independently developed MLWP models.

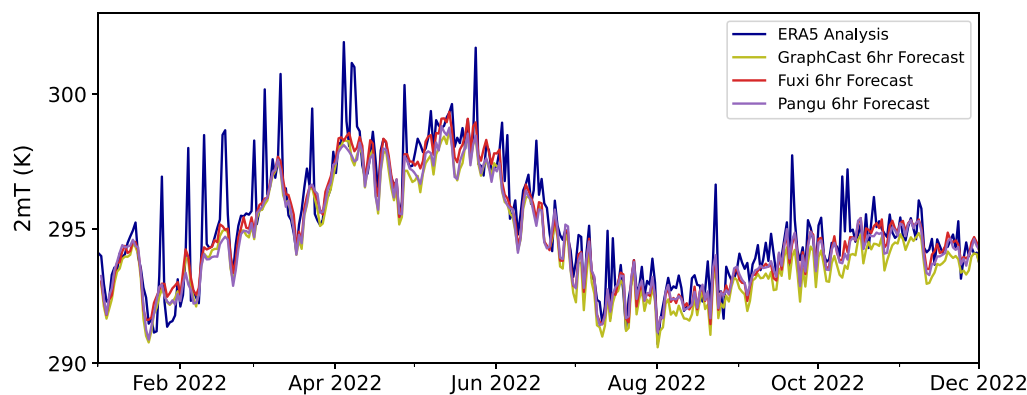


FIGURE A1 Time series of area-mean 2-m temperature 6-h forecasts from three different machine-learned weather prediction models, GraphCast, FuXi, and Pangu, compared with ERA5. Results are averaged over 7.5°N–13.5°N, 35°E–41°E and are valid at 0600 UTC during 2022.