The Rise of Data-Driven Weather Forecasting

A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context

Zied Ben Bouallègue,[a] Mariana C. A. Clare,[b] Linus Magnusson,[a] Estibaliz Gascón,[b] Michael Maier-Gerber,[b] Martin Janoušek,[b] Mark Rodwell,[a] Florian Pinault,[b] Jesper S. Dramsch,[b] Simon T. K. Lang,[a] Baudouin Raoult,[a] Florence Rabier,[a] Matthieu Chevallier,[a] Irina Sandu,[b] Peter Dueben,[b] Matthew Chantry,[a] and Florian Pappenberger[a]

KEYWORDS: Forecast verification/skill; Operational forecasting; Machine learning

ABSTRACT: Data-driven modeling based on machine learning (ML) is showing enormous potential for weather forecasting. Rapid progress has been made with impressive results for some applications. The uptake of ML methods could be a game changer for the incremental progress in traditional numerical weather prediction (NWP) known as the "quiet revolution" of weather forecasting. The computational cost of running a forecast with standard NWP systems greatly hinders the improvements that can be made by increasing model resolution and ensemble sizes. An emerging new generation of ML models, developed using high-quality reanalysis datasets like ERA5 for training, allows forecasts that require much lower computational costs and that are highly competitive in terms of accuracy. Here, we compare for the first time ML-generated forecasts with standard NWP-based forecasts in an operational-like context, initialized from the same initial conditions. Focusing on deterministic forecasts, we apply common forecast verification tools to assess to what extent a data-driven forecast produced with one of the recently developed ML models (PanguWeather) matches the quality and attributes of a forecast from one of the leading global NWP systems (the ECMWF IFS). The results are very promising, with comparable accuracy for both global metrics and extreme events, when verified against both the operational IFS analysis and synoptic observations. Overly smooth forecasts, increasing bias with forecast lead time, and poor performance in predicting tropical cyclone intensity are identified as current drawbacks of ML-based forecasts. A new NWP paradigm is emerging relying on inference from ML models and state-of-the-art analysis and reanalysis datasets for forecast initialization and model training.

SIGNIFICANCE STATEMENT: We compare the traditional approach of generating weather forecasts with a new approach based on machine learning (ML). The traditional approach is based on resolving physical equations with numerical methods and can be quite expensive to run on supercomputers. ML models are trained on a very large dataset that combines observations and physically based models to derive our best estimate of the state of the atmosphere hour after hour, over the past decades. In this work, we assess the forecast performance with statistical tools and a focus on extreme events. Our results suggest that ML models, which are much less expensive to run than standard methods, could have a promising future in numerical weather prediction. Our study also points out weaknesses and limitations that would require further investigations and future ML model improvements.

AFFILIATIONS: [a] European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom; [b] European Centre for Medium-Range Weather Forecasts, Bonn, Germany

1. Introduction

Numerical weather prediction (NWP) is the dominant approach for weather forecasting. A weather forecast is the result of the numerical integration of partial differential equations starting from the best estimate of the current state of the Earth system. The idea that the physical laws of fluid dynamics and thermodynamics can be used to predict the state of the atmosphere dates back to the pioneering works of Abbe (1901) and Bjerknes (1904). In a standard NWP framework, a weather prediction results from a deductive inference: a deterministic forecast is derived using the laws of physics starting from the best possible initial conditions, derived by optimally combining Earth system observations and short-range forecasts through data assimilation. However, our ability to perfectly know the initial conditions and numerically resolve the equations is limited. Hence, ensemble forecasting is used to account for uncertainty in both the initial conditions and the forecasting model, with the resulting ensemble forecast serving as a basis for probabilistic forecasting (Leutbecher and Palmer 2008).

A continuous improvement of the NWP performance has been observed over the last decades, including for the prediction of high-impact weather events (Ben Bouallègue et al. 2019). Skill improvement is achieved through improvements in initial conditions, numerical models, and resolution. At the European Centre for Medium-Range Weather Forecasts (ECMWF), the Integrated Forecasting System (IFS) has been run operationally since 1979 with regular updates of the different components of the forecasting system. The evolution of the IFS accuracy over the last two decades is shown in Fig. 1 (red lines). The steady increase in forecast accuracy, thanks to incremental improvements in numerical modeling, supercomputing, data assimilation and ensemble techniques, observations, and their use in the NWP system, has become known as the "quiet revolution" of weather forecasting (Bauer et al. 2015). However, the computational cost of running a forecast is a major bottleneck that hinders rapid improvements with standard NWP systems. In operational NWP, the computational and timeliness constraints imply finding a balance between increasing model resolution and increasing ensemble size, which are two major factors known to improve the skill of ensemble forecasts (Leutbecher and Ben Bouallègue 2020).

In recent years, data-driven modeling based on machine learning (ML) is showing large potential for weather forecasting applications with the promise to deliver forecasts at a much
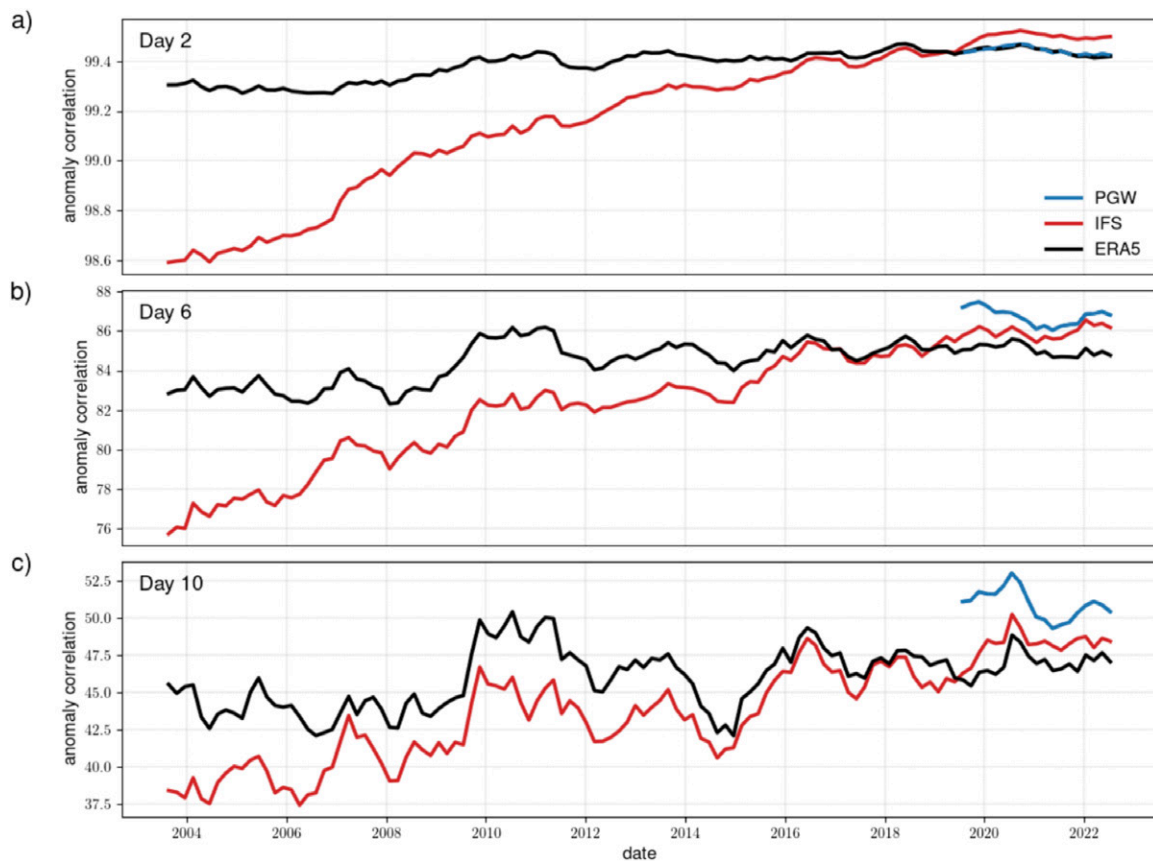
Fig. 1. Forecast accuracy (the larger the better) over the Northern Hemisphere at (a) day 2, (b) day 6, and (c) day 10. Accuracy is measured as the correlation between the forecasts and the verifying analysis for the geopotential height at 500 hPa, expressed as the anomaly with respect to the climatological height. A 1-yr running mean is applied. The constant improvement over the past decades of the IFS forecasts is compared with the performance of the ERA5 forecast (run with the IFS version operational in 2016) and with the performance of the PGW forecast trained over 1979–2018 and verified over 2019–23.

lower computational cost along with potential other benefits such as increased timeliness and potentially increased accuracy (de Burgh-Day and Leeuwenburg 2023). Pioneer works used simple convolutional-based neural networks to predict a small subset of variables using only these variables as predictors (Dueben and Bauer 2018; Weyn et al. 2019). Further developments in global weather forecasting employed more complex neural networks and more variables as predictors, resulting in more accurate machine learning models that were, nevertheless, still considerably less accurate than NWP systems (Weyn et al. 2020; Rasp and Thuerey 2021). However, since 2022, tremendous progress has been made with a series of key works developing machine learning models for weather forecasting, presenting impressive forecast scores for a large number of different weather variables, some of which rival the operational ECMWF deterministic high-resolution (deterministic) forecasts (Keisler 2022; Pathak et al. 2022; Bi et al. 2023; Lam et al. 2022; K. Chen et al. 2023). Concretely, Keisler (2022) uses a graph neural network (GNN) model and claim to produce more accurate forecasts of specific humidity than IFS after day 3; Pathak et al. (2022) leverage Fourier transforms with a transformer and claim to produce comparable accuracy to IFS for 2-m temperature; Bi et al. (2023) use a vision transformer model and claim to produce more accurate forecasts than IFS across numerous variables when both models are verified against reanalysis; Lam et al. (2022) use a GNN and claim more accurate forecasts than IFS on a larger set of atmospheric variables and pressure levels; and finally, K. Chen et al. (2023) use a transformer and claim to improve the scores compared to Lam et al. (2022), especially at longer lead times.

The emergence of data-driven models has been made possible, thanks to the availability of large, high-quality, open, and free meteorological datasets. The aforementioned ML models are trained on ERA5 reanalysis data, which is the fifth-generation ECMWF atmospheric reanalysis, produced by the Copernicus Climate Change service, as one of the European Union Copernicus Programme key deliverables (Hersbach et al. 2020). This dataset is particularly attractive for machine learning problems because it is a continuous weather dataset from 1940 to the present day and it represents the best possible reconstruction of the Earth system state created by blending past observations and short-range forecasts through data assimilation. However, the ML methods presented only train from 1979 because the extensions to 1940 are relatively recent and have lower accuracy due to the very limited availability of satellite data before 1980 (e.g., Hersbach 2023). ERA5 is generated using the operational IFS cycle at the time of production (2016) and is publicly available at a grid resolution of 0.25° (28 km). Hence, ML models are trained on reanalysis at a much lower resolution than that of today's operational forecasts and analyses (28 km instead of 9 km in the case of the ECMWF operational high-resolution forecasts and analyses). Note that, despite this resolution difference, ERA5 "forecasts" are used routinely for forecast verification purposes: the performance of the current IFS is compared with the performance of ERA5 forecasts (10-day forecast initialized from ERA5 at the ERA5 resolution, about 25 km) to help distinguish interannual variability from actual skill improvement due to changes in the forecasting system. This is illustrated in Fig. 1 with ERA5 forecast accuracy represented by black lines.

One approach to data-driven weather prediction would consist of running ML-trained models starting from optimized initial conditions in an operational context. In such a weather prediction system, the forecast inference relies on an ML model rather than the physical model (included, for example, in the IFS). This approach is highly attractive because a forecast can be generated at a speed several orders of magnitude faster than that from conventional methods. At a fundamental level, an ML-based prediction is the result of an inductive rather than a deductive inference. This paradigm shift in terms of logic has implications for the way a weather forecast is interpreted: a forecast becomes a plausible outcome given what has been learned from previous data. However, the mode of inference followed by ML methods can raise concerns, in particular, regarding the ability of such models to predict extreme events unseen in the training dataset. Moreover, the interpretability of ML models is also often questioned when they are perceived as black boxes where the link between the training dataset and the current forecast is difficult to grasp (McGovern et al. 2019). The huge potential benefits and drawbacks of data-driven systems trigger the question of whether ML models can become a component of operational NWP systems.

In this study, we evaluate the performance of data-driven forecasts in an operational-like context. More precisely, the PanguWeather ML model in Bi et al. (2023) (referred to hereafter as PGW), which is open-source for noncommercial use, has been set up to run on the ECMWF computers. For the first time, a forecast generated with an ML model is compared with an operational NWP forecast using the same framework and starting from the same initial conditions, while in Bi et al. (2023), like in previous studies, the ML-based forecasts were initialized from ERA5. We leverage standard verification techniques routinely applied for weather forecast evaluation at ECMWF. Using this methodology, we can assess which aspects of the data-driven forecasts can match the quality of forecasts performed with one of the leading operational NWP systems. This study focuses predominantly on the statistical analysis of forecast performance, but we acknowledge that case studies play a key role in understanding the capability and limitations of ML models in weather forecasting and refer the reader to Magnusson (2023). Also, the practical value of these forecasts would need to be carefully assessed in partnership with experienced forecasters, as discussed in Ebert-Uphoff and Hilburn (2023).

2. Methodology and experiments

Our comparative work is based on implementing PGW in an operational-like setting. PGW uses a vision transformer model architecture (Dosovitskiy et al. 2020) with 3D weather fields as inputs and outputs. Developed by Bi et al. (2023), the network minimizes a loss function defined as the root-mean-square error (RMSE) with a cos-latitude weighting to account for the spherical nature of the Earth as the model is trained on a regular latitude–longitude grid. Like most ML models, PGW uses an iterative method to forecast forward in time. A novelty of their approach, however, is that they choose to minimize the RMSE over a series of fixed short time periods (1, 3, 6, and 24 h) and then achieve weather forecasts at any time using a hierarchical temporal aggregation method. Here, PGW is run for 10 days, using a model time step of 24 h. Although PGW demonstrated reasonable results beyond this time period, only the first 10 days are analyzed here.

For the verification periods under focus in this work, IFS is run operationally at a horizontal grid spacing of 9 km up to 10 days lead time using the operational IFS cycle at the time (43r3 and 45r1 for 2018 and 47r3 for 2022). We also include the publicly available ERA5 forecast in our comparison. ERA5 forecasts start from ERA5 reanalysis and are based on a lower model resolution than the operational IFS forecast (30 km instead of 9 km). Here, we recall that the ERA5 reanalysis and forecasts are produced with a similar setup of IFS as that used for the operational high-resolution forecasts and analysis, but they are produced with the operational cycle at the time when the production of ERA5 started (41r2, in 2016). Thus, ERA5 forecasts and IFS forecasts differ in terms of initial conditions, resolution, and IFS cycle.

We choose to initialize PGW from the same analysis as IFS, namely, the ECMWF operational IFS analysis. This choice appears natural for a fair comparison between PGW and IFS in an operational-like setting. PGW was trained using ERA5 data and has a horizontal grid spacing of 28 km, meaning that initialization from the operational IFS analysis may induce some impact on scores. An optimal configuration of an ML-based forecasting system would likely "fine-tune" (training near convergence) operational IFS analysis. This optimization is outside the scope of this work but is worthy of note. The operational IFS analysis is created by using the current operational data assimilation system, which operates at a higher resolution (9 km) and uses a more recent (and therefore improved) IFS version than ERA5 to construct superior initial conditions.

As a complementary experiment, we also run PGW starting from ERA5 reanalysis (PGW_E5). This experiment shows that PGW starting from the operational high-resolution analysis generally performs better than PGW starting from ERA5 for the first days of the forecast (roughly up to day 4 depending on the variable and domain of interest), as illustrated in Fig. 2.

We also run IFS initialized from the operational IFS analysis at a lower resolution (IFS_LR) close to the PGW resolution to isolate the impact of model horizontal resolution on forecast performance. This impact differs depending on both variable and lead time as illustrated for geopotential at 500 hPa (Z500) and temperature at 850 hPa (T850). For T850 at a lead time of 2 days, a change in horizontal resolution results in a clear degradation of the forecast accuracy in Fig. 2a, but at a 6-day lead time and for Z500, there are only small differences between IFS and IFS_LR errors in our results in Fig. 2c.

As expected, IFS_LR is ranked between ERA5 and IFS in terms of performance. Indeed, IFS and IFS_LR forecasts are better than ERA5 forecasts because they start from the operational IFS analysis. Concerning PGW, which has the same horizontal resolution as IFS_LR, it is interesting to note that both forecasts have similar errors to ERA5 for T850 over the winter period, and these are noticeably larger than with operational IFS forecasts as shown in Fig. 2a. Also, the impact of the model horizontal resolution is smaller at longer lead times, as illustrated in Figs. 2c and 2d.
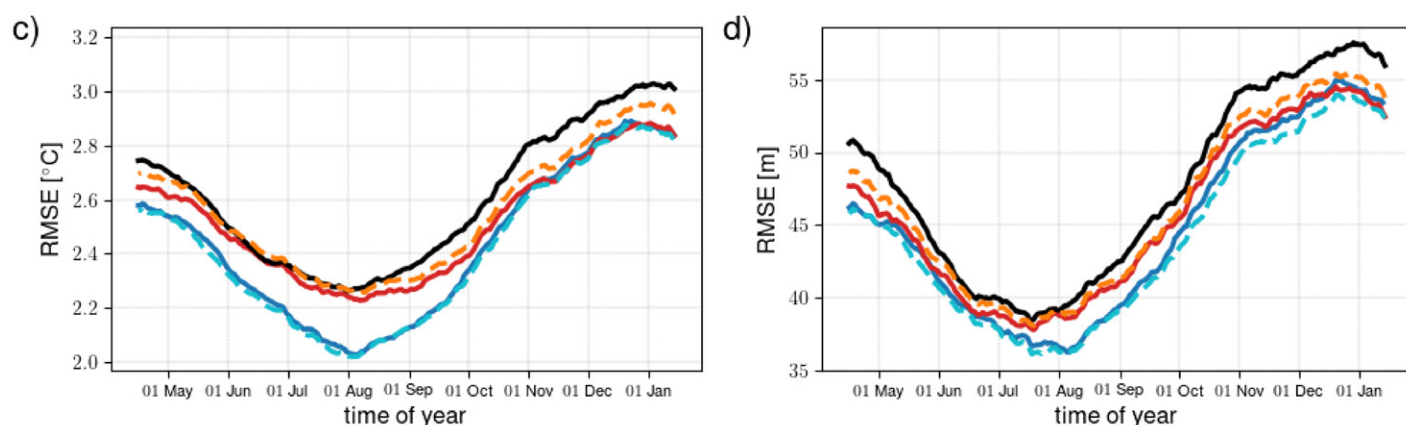
Day 2
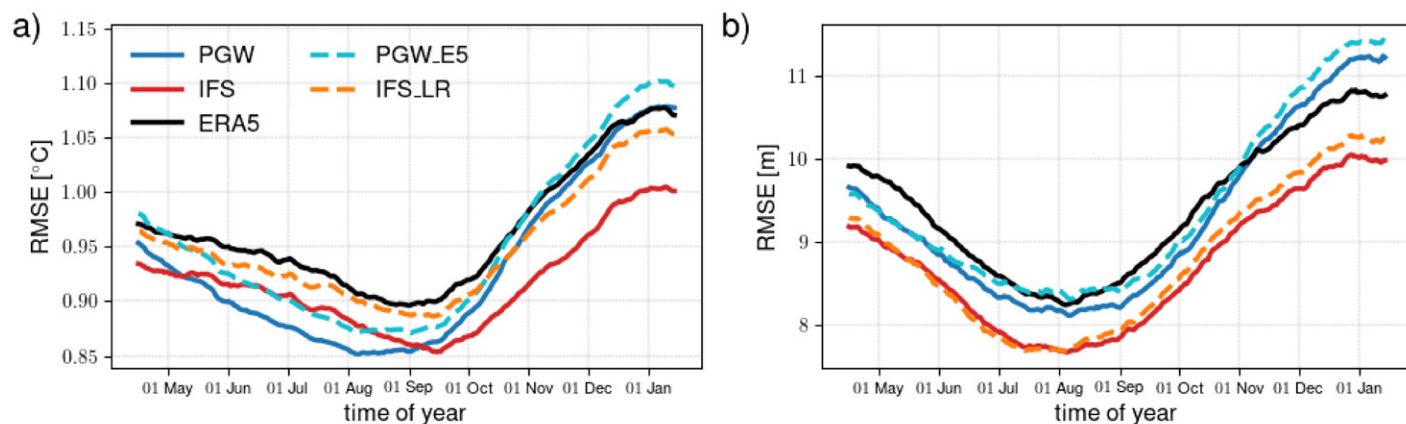
Day 6

Fɪɢ. 2. The seasonal cycle of the RMSE (3-month averaged) at day 2 and day 6 over 1 year covering the period 1 Mar 2022–28 Feb 2023 aggregated over the Northern Hemisphere for (a),(c) T850 and (b),(d) Z500. PGW_E5: PGW initialized with ERA5; IFS_LR: IFS run at a grid resolution of 0.25°. The forecasts are verified against operational IFS analyses on a grid resolution of 1.5°.

Finally, PGW appears performing better than IFS at day 2 for T850 over the summer months (JJA), shown in Fig. 2a, but worse for Z500 throughout the year, shown in Fig. 2b. Further investigations indicate that the RMSE for Z500 and mean sea level pressure in PGW is much worse over the central Arctic than the IFS. These results point to a fast-developing error over the central Arctic in that mass field that manifests at day 2 before other (nonsystematic) errors start to dominate.

In Table 1, we provide an overview of the forecasts analyzed in this study. We also include the ensemble forecast run operationally at ECMWF (ENS) and its ensemble mean (EM), which are discussed in the next sections. The 50-member ensemble has a horizontal grid spacing of 18 km for the period considered here. Please note, however, that the horizontal grid spacing of the ENS was recently reduced to 9 km, consistent with the resolution of the IFS deterministic forecast (Lang et al. 2023).

Tᴀʙʟᴇ 1. List of forecasts investigated in this study.

Name	Grid spacing	Initialization	Short description
IFS	9 km	Operational IFS analysis	Operational IFS high-resolution forecast
PGW	28 km	Operational IFS analysis	Operation-like PanguWeather forecast
IFS_LR	28 km	Operational IFS analysis	IFS forecast run at a lower horizontal resolution
PGW_E5	28 km	ERA5 analysis	PanguWeather forecast initialized with ERA5
ERA5 forecast	28 km	ERA5 analysis	ERA5 hindcast (forecasts for a historical period)
ENS (and EM)	18 km	Ensemble operational analysis	Operational IFS ensemble (end ensemble mean) forecast

3. Data and a case study

We assess the performance of Z500 and T850 forecasts against the operational IFS analysis interpolated to a grid resolution of 1.5° following the World Meteorological Organization (WMO) guideline and aggregated over the Northern Hemisphere. We also assess forecasts of 2-m temperature against surface synoptic (SYNOP) observations over Europe. In addition, a verification of tropical cyclone (TC) forecasts is performed. Details about the verification process are provided in the appendix.

We show the results mainly for two seasons: summer 2022 (1 June–31 August 2022) and winter 2022/23 (1 December 2022–28 February 2023) to allow a focus on both extreme cold and extreme warm temperatures. Both 0000 and 1200 UTC initializations are considered. These two verification periods are independent of the PGW training/validation dataset. Only results for winter 2022 are shown for the upper variables because it is the most dynamically active season in the Northern Hemisphere. Following Bi et al. (2023), the TC verification period covers 2 January–30 November 2018.

A comparison against SYNOP observations helps demonstrate the forecast performance from a user perspective. Nevertheless, in situ observations have their drawbacks: the quality of the measurements is not perfect, the stations are not distributed homogeneously over the verification domain, and measurements can suffer discontinuity at a given station. Also, representativeness is a major concern when comparing the model output with a point observation that might not be representative of the surrounding area. This representativeness issue is partially addressed here by an orography correction applied to the 2-m temperature forecasts (see, e.g., Ingleby 2015).

An illustration of a forecast and observation is provided in Fig. 3. The forecast evolution over consecutive starting times shows that the ensemble spread becomes smaller as we approach the observation date in Fig. 3a. In Sondankylä (Finland), −29°C was observed on that occasion. The PGW forecast has an earlier hint of the event severity than the IFS forecast, but both overestimated the temperature significantly to a similar degree. In the corresponding maps in Figs. 3b–d, PGW forecasts appear smoother than IFS forecasts, deprived of smaller-scale structures, but better capture the cold spell over Scandinavia 6 days ahead of the event. This first subjective assessment of a single case study agrees with the statistical analysis of the forecast performance discussed in the next section.

4. Comparing forecast performance

a. Contextualizing the forecast skill. Results in Fig. 4 (top row) are compelling: for lead times greater than 3 days, PGW forecasts are better than the ERA5 forecasts and as good as the operational IFS forecasts in terms of the RMSE. The EM, the ensemble functional that minimizes the RMSE, is the best performing forecast with this metric. The RMSE is a key indicator of forecast performance, but the RMSE results need to be interpreted in light of other forecast characteristics that also contribute to the quality of a prediction. For example, in terms of realism, the ensemble mean cannot be considered a plausible scenario of the atmospheric state at longer lead times.

In previous studies, there has been a concern that training toward RMSE results in overly smooth forecast fields (see, e.g., the smooth forecasts shown in Keisler 2022). Indeed, the RMSE strongly penalizes large forecast departures from the observations (or analyses), thus discouraging bold forecasts. When comparing the RMSE from different models, it is, therefore, important to check the level of activity of the different forecasts while interpreting the results. The activity of a forecast is here defined as the standard deviation of the forecast anomaly (see the appendix for a formal definition). IFS and ERA5 forecasts have a similar activity to each other and, importantly, similar activity to PGW for both T850 and Z500 (Fig. 4, middle row).
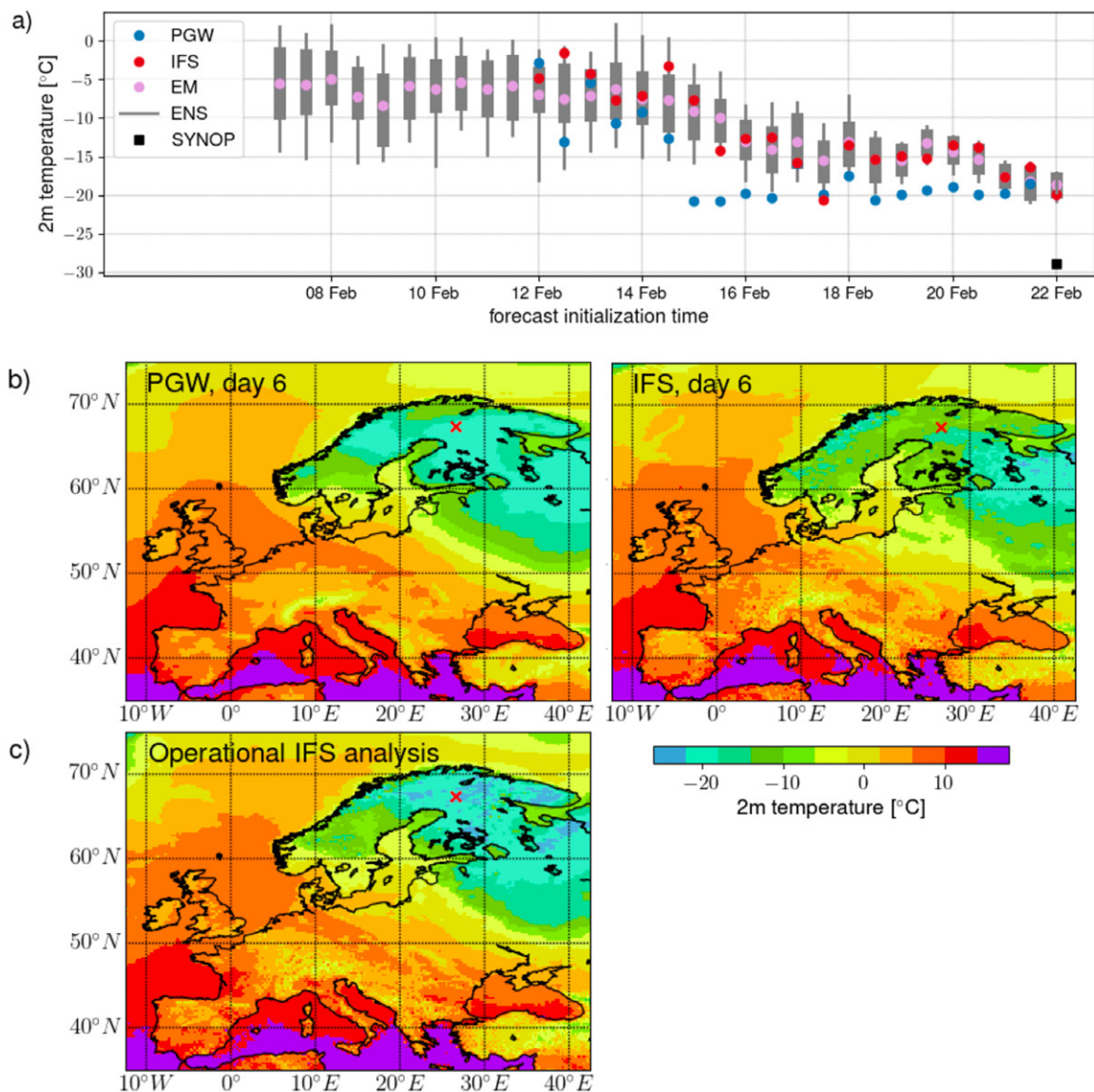
FIG. 3. An example of 2-m temperature forecasts and a corresponding SYNOP observation. (a) Evolution plots showing forecasts for Sodankylä (Finland). All the forecasts are valid for the same day (0000 UTC 22 Feb 2023) but have different initialization times up to 10 days ahead of the event for PGW and IFS. The ensemble forecast is shown in the form of the ensemble mean and quantile forecasts (boxplots showing 5%, 25%, 75%, and 95%) with a maximum lead time of 15 days. (b) PGW (left) and IFS (right) forecasts at day 6, and (c) verifying operational IFS analysis for Europe. The location of the Sodankylä SYNOP station is indicated with a red cross on the maps.

However, a clear smoothing of PGW forecasts at small scales is visible in Fig. 3. This dampening only has a minor contribution to the overall activity because this metric is dominated by larger scales. We note that for Z500, the slight decrease in the activity a with lead time of PGW is not statistically significant and in general, PGW does not become smoother at longer lead times as confirmed by power spectrum analysis (not shown). This is not the case for the EM, which becomes smoother as the forecast uncertainty increases. The lower activity of the EM contributes to this good RMSE performance at longer time ranges. The unpredictable features are filtered out by averaging the ensemble members, but forecast smoothness can be a non-desirable characteristic for some applications.

Finally, Fig. 4 also compares the bias of the different forecasts. Ideally, a forecast should have a bias close to zero. The magnitude of the bias in PGW forecasts grows at a much faster
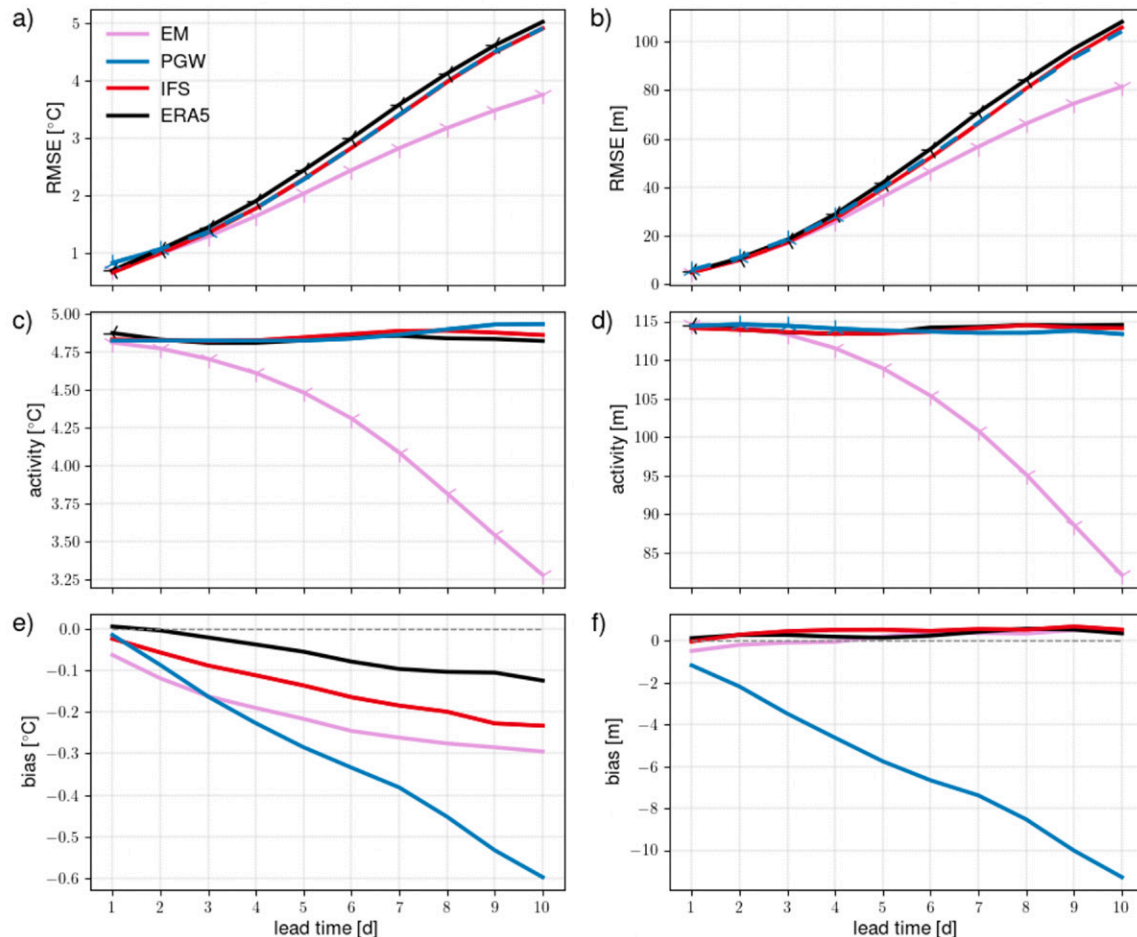
T850   **Z500**

F**IG. 4.** (a),(b) RMSE (the lower the better), (c),(d) forecast activity (the lower the forecast activity the smoother the forecast), and (e),(f) forecast bias (the closer to zero the better), as a function of forecast lead time for (left) T850 and (right) Z500. The forecasts are verified against the operational IFS analyses, and the results are valid for winter 2022/23 over the Northern Hemisphere. A statistically significant difference with respect to the operational IFS forecast is indicated with a marker in (a)–(d).

rate than the bias in IFS or ERA5 forecasts, with the bias drift particularly strong for Z500. Whereas the bias of IFS and ERA5 stabilizes at longer lead times, the incremental bias in PGW forecasts is still present when extending the forecast horizon.

Verification of upper variables against analysis is complemented by the verification of 2-m temperature against SYNOP observations in Fig. 5. We find that verifying against observations shows similar results as against analysis for key metrics: generally good performance of PGW forecasts in terms of RMSE (better than IFS in summer, but worse in summer after day 6), a bias drift for PGW with forecast lead time in summer, and EM outperforming the other forecasts from lead time day 4 onward, both in summer and in winter. Now, verification against observations is used as a framework for a more in-depth analysis of PGW forecast attributes.

b. Checking for statistical consistency. We attempt now to assess the statistical consistency between the deterministic forecasts and the corresponding observations. Here, we try to answer questions such as "Is the forecast able to mimic the observation statistical distribution?", "Is the forecast able to forecast extreme events of the same intensity as the observed ones?", and "Is the forecast systematically offset with respect to the observations?" Statistical

Summer 2022
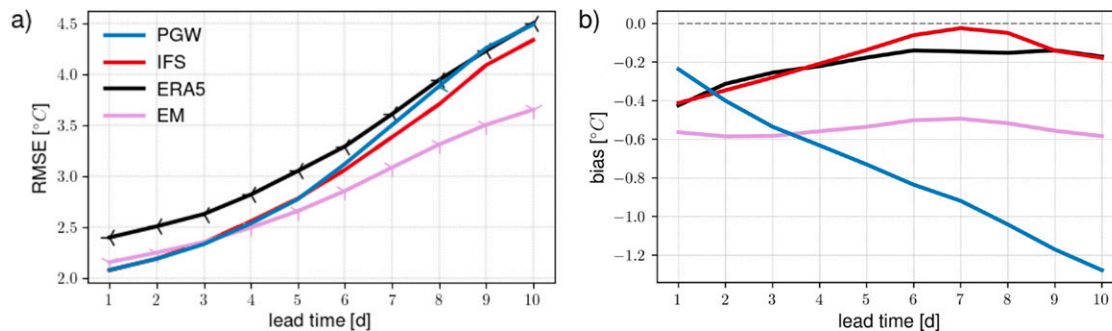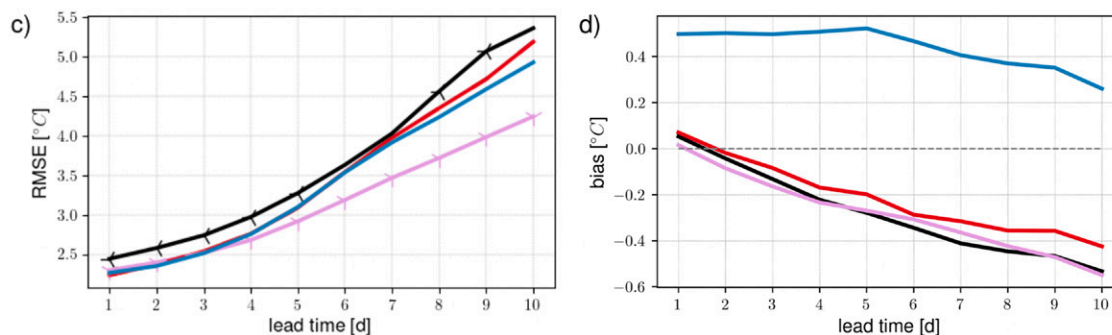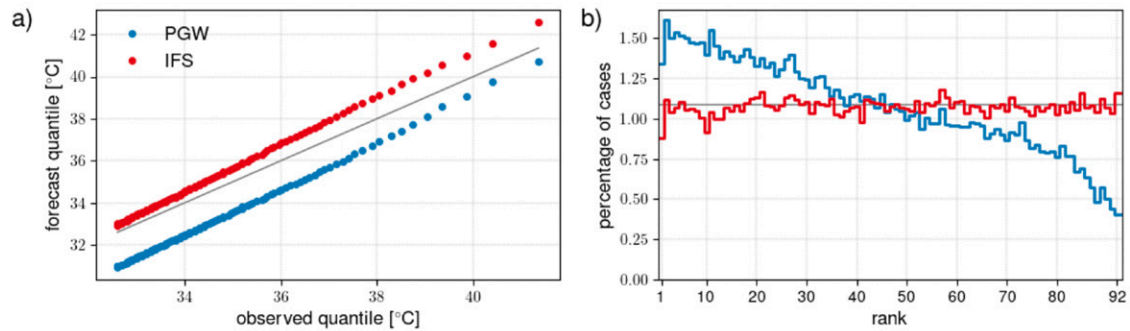

Winter 2022/2023

FIG. 5. Forecast performance for 2-m temperature over Europe during (top) summer 2022 and (bottom) winter 2022/23. (a),(c) RMSE and (b),(d) bias as a function of forecast lead time. Summer 2022 forecasts are initialized at 1200 UTC (valid at midday), while winter 2022/23 forecasts are initialized at 0000 UTC (valid at midnight). The forecasts are verified against SYNOP observations. A statistically significant difference with respect to the operational IFS forecast is indicated with a marker in (a) and (c).

consistency in terms of distribution is analyzed using quantile–quantile (Q–Q) plots and observation rank histograms (a new type of diagnostic described in detail in the appendix). The coherence of the spatial structures in the forecast would require additional diagnostic tools beyond the scope of this study.

Q–Q plots for forecasts at day 6 focus on warm temperatures during summer in Fig. 6a and on cold temperatures during winter in Fig. 6c. For the former, both PGW and IFS forecasts can capture the observed extreme temperatures with PGW displaying a general offset consistent with the PGW bias at day 6 in Fig. 5b. For the latter, extremely low temperatures are not fully captured, neither by PGW nor by the IFS, as already illustrated in the case study in Fig. 3. In northern Europe, very low temperatures are reached closest to the ground during clear-sky nights over snow-covered regions. This cooling is not fully captured by the IFS during the evaluated period (Day et al. 2020).

Observation rank histograms (ORHs) are used to check whether the forecasts cover the observed range at each station separately. With OHRs, we visualize how forecasts fall within the observed empirical distribution during the verification period. This diagnostic includes all stations rather than focusing on the hottest or coldest temperatures in the verification domain as in a Q–Q plot. A flat histogram indicates that the distribution of forecasts and observed temperatures is similar. This is the case for the IFS forecasts over the summer in Fig. 6b while the tilted histogram for PGW reflects a systematic bias in the forecast. During winter, the IFS forecasts tend to be too cold at nighttime over mainland Europe (Sandu et al. 2020) while still not reaching extremely low temperatures in northern Europe, as shown in Fig. 6. The overall negative bias leads to an overpopulated first bin of the IFS histogram in Fig. 6d, while PGW does not fully capture the lowest temperature at each station, leading to underpopulated first bins of the histogram.
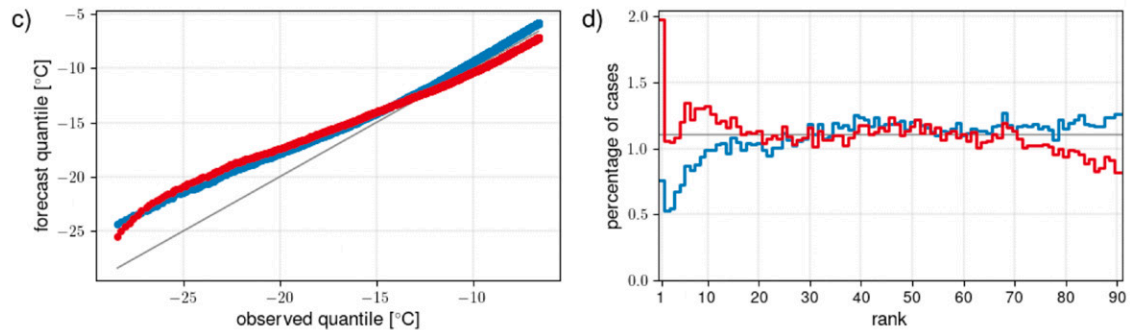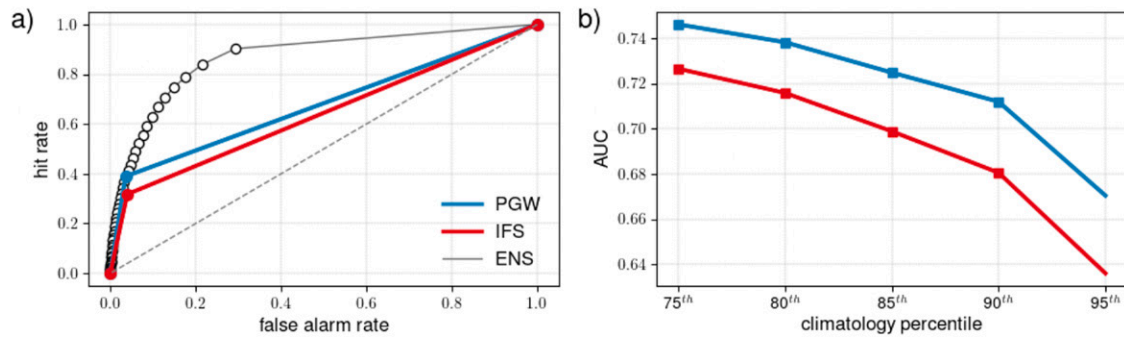
Summer 2022

Winter 2022/2023

FIG. 6. Statistical consistency of 2-m temperature over Europe during (top) summer 2022 and (bottom) winter 2022/23 for IFS and PGW forecasts at day 6. (a),(c) Q–Q plots showing a scatterplot of the empirical forecast quantiles vs the quantiles from the observation distribution at quantile levels 90%, 90.1%, …, 99.9% for the summer period (a) and (c) 0.1%, 0.2%, …, 1.0% for the winter period. (b),(d) Observation rank histograms show the averaged number of forecasts in the bins defined by the sorted observations at each station. For all plots, perfect reliability is indicated by a gray line.

c. Forecasting weather events. The usefulness of a forecast is judged by its ability to predict weather events, often related to extremes. Here, the focus is on the forecast's ability to distinguish between an event and a nonevent. Events are defined as 2-m temperature exceeding a climate percentile. The climatology varies for each station, and the climate percentiles are estimated based on the verification sample for the forecasts and the observations separately, to remove any bias in the forecast because a measure of discrimination ability ought to be independent of the forecast bias. We consider only low-temperature events for the winter period and high-temperature events for the summer period.

The relative operating characteristic (ROC) curve is a popular diagnostic tool in forecast verification. ROC curves plotting the hit rate versus false alarm rate of a high-temperature event in summer and a cold temperature event in winter are shown in Figs. 7a and 7c, respectively. Deterministic forecasts such as PGW and IFS forecasts have only one nontrivial point on the curve. This point is closer to the top-left corner of the plot for PGW than for the IFS, indicating that PGW has better discrimination ability than the IFS for the events under consideration. For the ensemble forecast, the ROC curve is built using one point for each probability issued by ENS using a standard "trapezoidal" approach (Ben Bouallègue and Richardson 2022). The ROC curve of a probabilistic forecast, represented here by the empty circles, covers a much wider area than the curve derived from a single forecast.

A standard measure for discrimination is the area under the ROC curve (AUC). In Figs. 7b and 7d, the AUC of 6-day ahead forecasts is plotted as a function of percentile thresholds. The severity of the event increases as the climate percentiles get closer to 0% in winter and 100% in summer, indicating a rarer event under scrutiny. In general, AUC decreases when focusing on more intense/rare events as it becomes more difficult to predict such events with
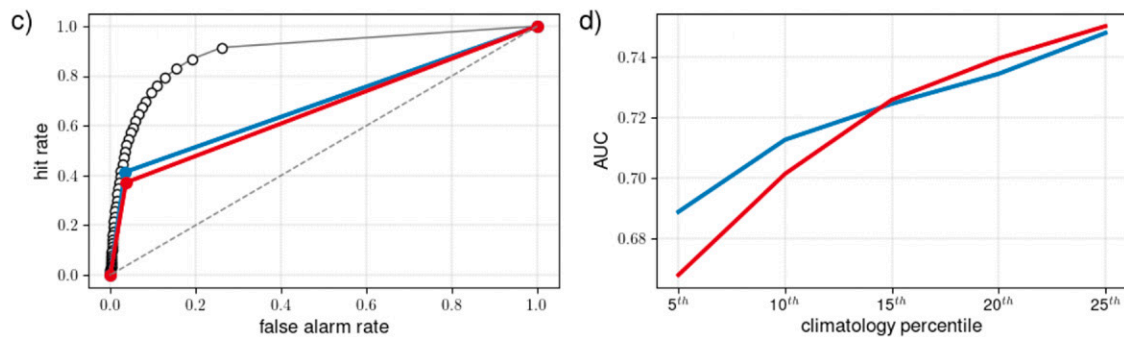
Summer 2022

Winter 2022/2023

Fɪɢ. 7. Performance in forecasting 2-m temperature events defined by climate thresholds for (top) summer 2022 and (bottom) winter 2022/23 over Europe at day 6 lead time. ROC curve (the closer to the top-left corner the better) for an event defined as (a) exceeding the 95% climate percentile in summer and (c) below the 5% climate percentile in winter. The diagonal dashed line is the zero-discrimination line. Results for ENS-derived probability forecast are also shown. (b),(d) Discrimination ability as measured with the AUC (the higher the better) and plotted as a function of climate percentile used to define a weather event. A statistically significant difference between the PGW and IFS results (as estimated by block-bootstrapping) is indicated by a square.

a deterministic forecast. The IFS and PGW have similar levels of performance in winter, while PGW outperforms the IFS in summer with differences statistically significant for percentiles between 75% and 90%, as estimated by block-bootstrapping. Combining the results in Figs. 6 and 7, we see that PGW climatology for summer extremes is less consistent with the analysis climatology, but after accounting for this discrepancy, PGW has more accurate forecasts of summer temperature extremes than IFS.

d. Forecasting tropical cyclones. TCs are a prominent example of extreme weather that has a devastating impact and attracts considerable attention from the public and media. More-over, TCs are characterized by large deviations from the mean state of the atmosphere and are, thus, generally challenging to forecast. Here, we focus on the year 2018 (as is done in Bi et al. 2023), but note that IFS TC forecasts have substantially improved with more recent cycles (Forbes et al. 2021; Majumdar et al. 2023). We assess two key characteristics of TCs: their track position and their intensity (see the appendix for more details). In Fig. 8a, the position error is measured as the distance between the TC position in the forecasts and the observations at a specific time. Larger errors are observed for PGW during the first day com-pared with IFS, but PGW has slightly lower errors for lead times greater than 2 days. This difference is partly explained by the fact that the propagation speed is generally too slow in the IFS (J.-H. Chen et al. 2023) but not in PGW (not shown). Overall, the differences in the position error are small and not statistically significant between models.

Focusing now on the TC intensity, Fig. 8b shows the mean absolute error for TC central pres-sure. Here, we find that PGW clearly underestimates the intensity (i.e., the predicted pressure
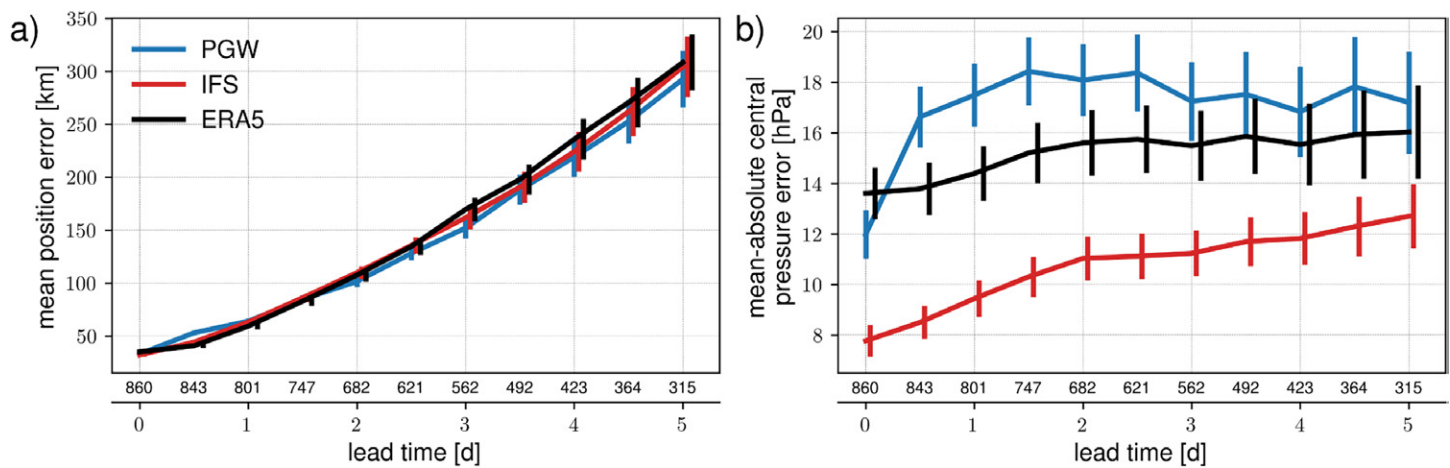
FIG. 8. Tropical cyclone verification results: (a) mean position error and (b) mean absolute central pressure error as a function of lead time for 2018. Forecasts are verified against the IBTrACS dataset and homogenized to have a consistent number of cases between models. For each lead time, the number of cases is displayed directly below the graphs. The vertical bars indicate the 2.5%–97.5% confidence intervals.

is too high). Both IFS and ERA5 perform better than PGW in terms of TC intensity error (except for at 0 day lead time). The large positive bias of PGW in the minimum core pressure results from too-weak gradients and too-weak maximum wind speed, while the IFS more closely resembles the analysis (not shown). The better performance of the IFS compared with ERA5 is mainly explained by its higher resolution (9 vs 28 km) but is also due to improvements in IFS through model development.

The number of unique TCs observed in our verification dataset is 107. We counted 105 TCs in IFS and 95 in Pangu at the initial time; 86 in IFS; and 61 in Pangu at day 5. Among the observed TCs, 51 are of category 1 or above. We count 49 (34) TCs of these categories in the IFS and 46 (33) in Pangu at the initial time (day 5). So the difference in the number of TCs predicted between the IFS and PGW appears to be a function of TC intensity. Further investigations are needed to explore the root causes for the reduced number of low-intensity TCs in the ML-based forecast as well as to closely examine the TC structures and physical consistency between variables.

e. Predicting the forecast error. The day-to-day variability of the error is compared for PGW and IFS forecasts. We aim to identify common patterns in error growth and examine the sensitivity to predictability barriers. For this purpose, we analyze the so-called *predictability barrier plots*: 2D diagrams displaying the forecast error as a function of both the forecast starting time (*x* axis) and the forecast lead time (*y* axis). A more in-depth analysis would involve running different models from different initial conditions as in Magnusson et al. (2019), but this approach is out of scope for this paper.

Examples of predictability barrier plots for PGW and IFS are provided in Figs. 9a and 9b, respectively, focusing on daily scores of Z500 forecasts over Europe. In these plots, a transverse structure indicates rapid error growth leading to a poor forecast at all lead times: in that case, the forecast initialization might be the dominant predictability limiting factor. By contrast, a vertical structure indicates a weather situation difficult to predict for consecutive runs with different initialization, likely to be due to predictability barriers for that specific weather situation.

In general, we see a good agreement between PGW and IFS daily errors. This similarity is even more evident when plotting daily errors for a single lead time (here day 6) for the whole verification period. The correlation coefficient between the two time series is 0.54. Strikingly, Fig. 9c shows the same "bust" in forecasting the weather over Europe for 6 February.
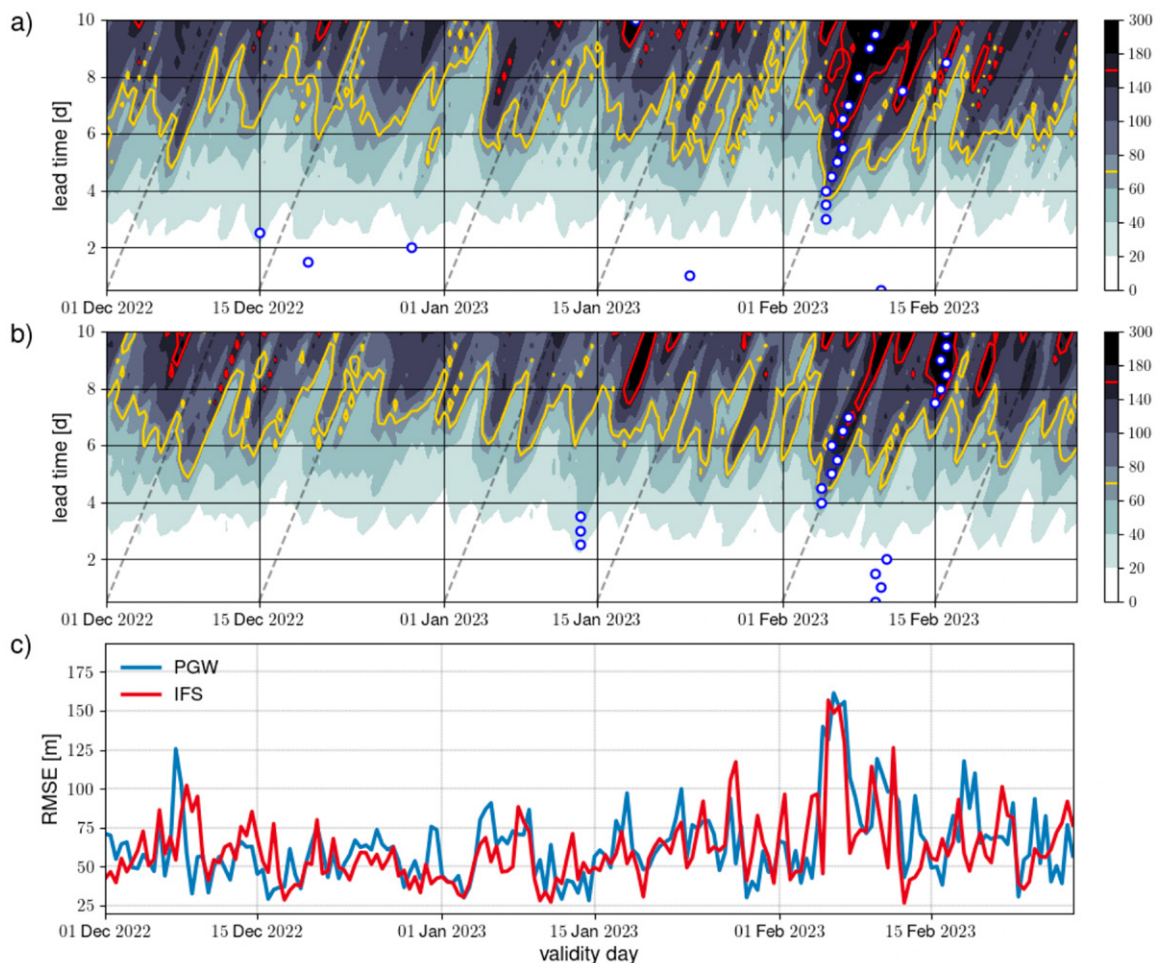
FIG. 9. Predictability barrier plots showing daily RMSE for Z500 for lead times 1–10 days over winter 2022/23 for (a) PGW and (b) IFS. (c) A cross section at day 6 of (a) and (b). In (a) and (b), the shade indicates the score value (m), the vertical lines' intercept scores for all forecasts valid on a given day, the transversal lines' intercept scores for a given forecast run for all lead times, the yellow lines indicate the averaged score for a day 6 forecast, the red lines mark an (arbitrary) large error, and the blue dots indicate the worst score over the period for each lead time.

This flow-dependent nature of the error points toward the need for ensemble forecasting in a similar fashion for ML models like PGW as is common practice for NWP nowadays.

5. Summary and outlook

The results shown here highlight that ML models could have a promising future in numerical weather prediction. To explore the advantages and limitations of data-driven weather forecasts, we have run PanguWeather, an ML model trained on ERA5, initialized with the operational IFS analysis. The PGW forecasts are compared with the operational IFS forecasts to help shape our understanding of the characteristics of both the data-driven forecasts and their errors. Some of the most challenging weather phenomena are linked to rain, but our comparison does not include precipitation because the field is not present in PanguWeather.

Fundamentally, data-driven forecasts show overall good performance comparable with the IFS for both upper-air variables (geopotential height at 500 hPa and temperature at 850 hPa) when verified against operational IFS analysis and for a surface variable (2-m temperature) against observations. While these conclusions are supported by further investigations including other variables such as 10-m wind (not shown here), a number of weaknesses and limitations of PanguWeather are also pointed out.

The data-driven forecast appears smoother than the operational IFS forecast, but the level of smoothness does not seem to increase with the forecast lead time, as we might expect when training toward the RMSE. However, we observe a drift in the bias almost linear with the forecast lead time. In particular, we note a cold bias in surface temperatures over Europe during the summer and a drift in the geopotential height in the Northern Hemisphere during the winter that might originate from poor forecast performance over the Arctic. While statistical postprocessing could offer a means to correct systematic errors in the forecast (Vannitsem et al. 2021), the drifts in the bias should be investigated further and addressed when developing future ML models. A deeper understanding of systematic errors could be achieved by performing conditional verification, for example, focusing on specific physical processes. Moreover, other diagnostic tools could be used to check for physical consistency based, for instance, on multivariate verification that accounts for the correlation between variables.

We also note that there is room for further improvement in the implementation of data-driven weather prediction systems. Like all data-driven approaches presented so far, the ML model used in our study is not trained on the operational IFS analysis and our experiments do not involve fine-tuning. Instead, ERA5 is used as the training dataset, which has so far been the cornerstone of any data-driven approach. For some aspects, ML models appear to directly inherit advantages and drawbacks from the numerical weather prediction system used to generate the training dataset. For example, the model resolution of the training data can in part explain the limitation in forecasting small-scale structures with PanguWeather. However, the forecast initial conditions also play a crucial role: starting a forecast from the operational IFS analysis rather than ERA5 analysis offers an advantage in accuracy also for the medium range. Moreover, the similarities in error growth of a data-driven forecast and a standard NWP forecast indicate similar sensitivities to chaos between ML-based and physically based models.

Good performance of data-driven forecasts compared with IFS is also observed in predicting some extreme events and confirmed by case studies. The results shown here focus first on events defined as climate threshold exceeded at a station location. The performance of ML-based models in forecasting TCs is under scrutiny too, as in Bi et al. (2023). Preliminary investigations indicate that current ML models, due to their lower resolution, tend to predict less weak TCs compared to the IFS, with tracks of similar quality, but IFS better captures their intensity and structure. Additional studies would help to demonstrate the value of data-driven forecasts as well as their strengths and weaknesses in supporting decision-making.

Finally, in this work, we focused on deterministic forecasts but ensemble forecasts are key in providing uncertainty information for decision-making. A Monte Carlo approach for uncertainty quantification has been tested starting ML-derived forecasts from perturbed initial conditions based on the ECMWF ensemble data assimilation and singular vector perturbations. The initial condition perturbation methodology is described in Lang et al. (2021b). The resulting ensemble forecast is showing promising results. In future work, uncertainty in initial conditions will be complemented by mechanisms to account for model uncertainty (see, e.g., Lang et al. 2021a) in a data-driven weather prediction context.

This first assessment of a machine learning-based weather forecast in an operational-like context shows very promising results. The future role of ML models in the context of numerical weather prediction systems, and the ability of this approach to complement physical models remains to be explored. Operational centers should explore the strengths and weaknesses of these models as additional components of their forecasting systems: the ability to run forecasts at a much higher speed and much lower computational cost opens new horizons.

APPENDIX

Verification Process and Scores' Definition

a. Contextualizing the forecast skill. We aim at a quantitative comparison of PGW and IFS forecasts. In general terms, forecast verification consists of measuring the relationship between a forecast and the corresponding observation[A1] (Murphy and Winkler 1987). For this purpose, one can carefully choose from a variety of metrics and diagnostics (see Wilks 2006; Jolliffe and Stephenson 2011). To go beyond the computation of generic scores, it is possible to investigate the properties of the joint distribution of forecasts and observations. The two main forecast attributes are forecast consistency (or calibration) and forecast discrimination ability. Note that these concepts hold also when dealing with probabilistic forecasts (which are not explored here).

[A1] The term *observation* is used in the broad sense of an assumed "truth." It can take the form of an analysis or an observation.

Classical statistic tools involve the computation of summary statistics and scores. Summary statistics include the bias defined as the averaged difference between forecasts and observations, and forecast/observation activity defined as the standard deviation of the forecast/observation anomaly. Scores are metrics measuring the forecast accuracy such as the anomaly correlation shown in Fig. 1 or the forecast error such as the widely used RMSE shown in Fig. 2. Here, we formally define the following quantities:

- The forecast root-mean-square error:

$$\sqrt{\overline{\left(f-o\right)^2}}, \tag{A1}$$

- the forecast mean error (or bias):

$$\overline{f-o}, \tag{A2}$$

- the forecast activity:

$$\sqrt{\overline{\left[\left(f-c\right)-\overline{f-c}\right]^2}}, \tag{A3}$$

- the observation activity:

$$\sqrt{\overline{\left[\left(o-c\right)-\overline{o-c}\right]^2}}, \tag{A4}$$

- the forecast anomaly correlation:

$$\frac{\overline{\left(f-c-\overline{f-c}\right)\left(o-c-\overline{o-c}\right)}}{\sqrt{\overline{\left(f-c-\overline{f-c}\right)^2}}\sqrt{\overline{\left(o-c-\overline{o-c}\right)^2}}}, \tag{A5}$$

with f being the forecast, o being the observation, c being the climatology, and $\overline{}$ being the averaging operator including a latitude weighting.

b. Checking for statistical consistency.

Statistical consistency is tested regionally with Q–Q plots and locally (at the station level) with observation rank histograms. For this exercise, we exclude stations situated at an altitude greater than 1000 m to avoid focusing predominantly on representativeness issues rather than model characteristics. For Q–Q plots, quantiles are estimated from the whole verification sample (Europe, summer 2022 or winter 2022/23) for both observations and forecasts, separately. We restrict our analysis to the warm tail of the distribution in the summer (quantile levels in the range 90%–99.9%) and to the cold tail of the distribution in the winter (quantile levels in the range 0.1%–10%).

As a complementary diagnostic tool, we suggest a new type of plot: the ORH. Inspired by the ensemble rank histogram used to assess the reliability of ensemble forecasts, the ORH is built by ranking the observations from the smallest to the biggest for the whole verification period and individual forecasts, for each station separately. The rank of the forecast for each verification day is registered and populates the histogram. ORH assesses whether forecasts and observations are distributed similarly at a station level. For example, a forecast with a positive bias will lead to a histogram tilted to the right (toward the largest observed values), while a forecast with a negative bias will lead to a histogram tilted to the left (toward the smallest observed values).

c. Forecasting weather events.

Forecasting specific events is at the heart of many weather applications. The ability of a forecast to distinguish between the occurrence and nonoccurrence of an event is called discrimination. Based on local climatology, event thresholds are defined as discussed above. Forecasts and observations are transformed into binary values with respect to a given threshold. A contingency table is populated for each of the dichotomous events. A contingency table is a 2 × 2 table where hits, misses, false alarms, and correct negatives are counted. From this table, it is possible to derive both the hit rate and the false alarm rate, the two components of the ROC curve. The AUC is a common measure of discrimination in weather forecast verification (Mason 1982; Harvey et al. 1992).

Weather events are defined with the help of a local climatology that differs for each station. The same verification setting is used as in Ben Bouallègue et al. (2019) where an event is defined using a percentile of a climatology rather than a fixed absolute value. This approach tries to reflect that user-relevant thresholds are often associated with potential hazards and as such vary from place to place. For example, the 5% percentile of the local temperature climatology corresponds to very different absolute thresholds for say Helsinki and Madrid. Also using a climatology-based threshold allows us to avoid the pitfall of measuring varying climatology rather than actual skill (Hamill and Juras 2006).

A different climatology is defined for (i) the observations and (ii) each forecast, with percentiles directly estimated from the verification sample. This so-called *eigen-climatology* approach corresponds to practically applying an in-sample local bias correction of the forecast as discussed in more detail in Ben Bouallègue et al. (2019). This step is important to

disentangle discrimination from calibration attributes, because the latter can, in principle, be improved by postprocessing. Only stations where measurements are available throughout the full verification period are considered for this exercise.

Finally, we recall that statistical significance is important when comparing competing forecasts (Geer 2016). Here, we assess the chaotic variability of the scores with the use of (block)-bootstrapping. We randomly choose the verification days entering the verification dataset and compute scores for each forecast. Based on a 1000-member block bootstrap sample with blocks of 5 days, statistical significance to the 5% level is estimated.

d. Forecasting tropical cyclones. We also assess performance in forecasting TCs. TCs are tracked in forecasts from PGW, IFS, and ERA5 with the ECMWF operational TC tracker as described in Magnusson et al. (2021). Forecasts up to 5 days are verified here as results for longer lead times are unlikely to be statistically significant. As observations, we use the International Best Track Archive for Climate Stewardship (IBTrACS) database (Knapp et al. 2010, 2018). The verification is based on TCs that are present in the observation database at the forecast initial time. The sample is homogenized to include the same cases for all three models. This homogenization results in a sample size of 860 cases at the analysis time, down to 315 cases at day 5.[A2] An intensity threshold of $17\,\mathrm{m\,s^{-1}}$ is applied to filter the observation dataset for TCs that reach tropical storm strength.

[A2] If only the IFS was validated, the maximum number of cases would be 988 at the forecast initial time and 592 for 5-day forecasts.

References

Abbe, C., 1901: The physical basis of long-range weather forecasts. *Mon. Wea. Rev.*, **29**, 551–561.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Ben Bouallègue, Z., and D. S. Richardson, 2022: On the ROC area of ensemble forecasts for rare events. *Wea. Forecasting*, **37**, 787–796, https://doi.org/10.1175/WAF-D-21-0195.1.

——, L. Magnusson, T. Haiden, and D. S. Richardson, 2019: Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quart. J. Roy. Meteor. Soc.*, **145**, 1741–1755, https://doi.org/10.1002/qj.3523.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bjerknes, V., 1904: Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteor. Z.*, **21**, 1–7.

Chen, J.-H., L. Zhou, L. Magnusson, R. McTaggart-Cowan, and M. Koehler, 2023: Tropical cyclone forecasts in the DIMOSIC project—Medium-range forecast models with common initial conditions. *Earth Space Sci.*, **10**, e2023EA002821, https://doi.org/10.1029/2023EA002821.

Chen, K., and Coauthors, 2023: FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. arXiv, 2304.02948v1, https://doi.org/10.48550/arXiv.2304.02948.

Day, J. J., G. Arduini, I. Sandu, L. Magnusson, A. Beljaars, G. Balsamo, M. Rodwell, and D. Richardson, 2020: Measuring the impact of a new snow model using surface energy budget process relationships. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002144, https://doi.org/10.1029/2020MS002144.

de Burgh-Day, C. O., and T. Leeuwenburg, 2023: Machine learning for numerical weather and climate modelling: A review. *Geosci. Model Dev.*, **16**, 6433–6477, https://doi.org/10.5194/gmd-16-6433-2023.

Dosovitskiy, A., and Coauthors, 2020: An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv, 2010.11929v2, https://doi.org/10.48550/arXiv.2010.11929.

Dueben, P. D., and P. Bauer, 2018: Challenges and design choices for global weather and climate models based on machine learning. *Geosci. Model Dev.*, **11**, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018.

Ebert-Uphoff, I., and K. Hilburn, 2023: The outlook for AI weather prediction. *Nature*, **619**, 473–474, https://doi.org/10.1038/d41586-023-02084-9.

Forbes, R., P. Laloyaux, and M. Rodwell, 2021: IFS upgrade improves moist physics and use of satellite observations. *ECMWF Newsletter*, No. 169, ECMWF, Reading, United Kingdom, 17–24, https://www.ecmwf.int/en/newsletter/169/meteorology/ifs-upgrade-improves-moist-physics-and-use-satellite-observations.

Geer, A. J., 2016: Significance of changes in medium-range forecast scores. *Tellus*, **68A**, 30229, https://doi.org/10.3402/tellusa.v68.30229.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

Harvey, L. O., J. K. Hammond, C. Lusk, and E. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883, https://doi.org/10.1175/1520-0493(1992)120<0863:TAOSDT>2.0.CO;2.

Hersbach, H., 2023: ERA5 reanalysis now available from 1940. *ECMWF Newsletter*, No. 175, ECMWF, Reading, United Kingdom, 10, https://www.ecmwf.int/en/newsletter/175/news/era5-reanalysis-now-available-1940.

——, and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Ingleby, B., 2015: Global assimilation of air temperature, humidity, wind and pressure from surface stations. *Quart. J. Roy. Meteor. Soc.*, **141**, 504–517, https://doi.org/10.1002/qj.2372.

Jolliffe, I. T., and D. B. Stephenson, 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd ed. Academic Press, 627 pp.

Keisler, R., 2022: Forecasting global weather with graph neural networks. arXiv, 2202.07575v1, https://doi.org/10.48550/arXiv.2202.07575.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

——, H. J. Diamond, J. P. Kossin, M. C. Kruk, and C. J. I. Schreck, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) project, version 4. NOAA NCEI, accessed 1 May 2024, https://doi.org/10.25921/82ty-9e16.

Lam, R., and Coauthors, 2022: Graphcast: Learning skillful medium-range global weather forecasting. arXiv, 2212.12794v2, https://doi.org/10.48550/arXiv.2212.12794.

Lang, S., D. Schepers, and M. Rodwell, 2023: IFS upgrade brings many improvements and unifies medium-range resolutions. *ECMWF Newsletter*, No. 176, ECMWF, Reading, United Kingdom, 11, https://www.ecmwf.int/sites/default/files/elibrary/072023/81380-ifs-upgrade-brings-many-improvements-and-unifies-medium-range-resolutions.pdf.

Lang, S. T. K., S.-J. Lock, M. Leutbecher, P. Bechtold, and R. M. Forbes, 2021a: Revision of the stochastically perturbed parametrisations model uncertainty scheme in the integrated forecasting system. *Quart. J. Roy. Meteor. Soc.*, **147**, 1364–1381, https://doi.org/10.1002/qj.3978.

——, and Coauthors, 2021b: More accuracy with less precision. *Quart. J. Roy. Meteor. Soc.*, **147**, 4358–4370, https://doi.org/10.1002/qj.4181.

Leutbecher, M., and T. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

——, and Z. Ben Bouallègue, 2020: On the probabilistic skill of dual-resolution ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **146**, 707–723, https://doi.org/10.1002/qj.3704.

Magnusson, L., 2023: First exploration of forecasts for extreme weather cases with data-driven models at ECMWF. *ECMWF Newsletter*, No. 176, ECMWF, Reading, United Kingdom, 8–9, https://www.ecmwf.int/en/newsletter/176/news/.

——, J.-H. Chen, S.-J. Lin, L. Zhou, and X. Chen, 2019: Dependence on initial conditions versus model formulations for medium-range forecast error variations. *Quart. J. Roy. Meteor. Soc.*, **145**, 2085–2100, https://doi.org/10.1002/qj.3545.

——, and Coauthors, 2021: Tropical cyclone activities at ECMWF. ECMWF Tech. Memo. 888, 140 pp., https://www.ecmwf.int/en/elibrary/81277-tropical-cyclone-activities-ecmwf.

Majumdar, S. J., L. Magnusson, P. Bechtold, J. R. Bidlot, and J. D. Doyle, 2023: Advanced tropical cyclone prediction using the experimental global ECMWF and operational regional COAMPS-TC systems. *Mon. Wea. Rev.*, **151**, 2029–2048, https://doi.org/10.1175/MWR-D-22-0236.1.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338, https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, https://doi.org/10.48550/arXiv.2202.11214.

Rasp, S., and N. Thuerey, 2021: Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: A new model for WeatherBench. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002405, https://doi.org/10.1029/2020MS002405.

Sandu, I., and Coauthors, 2020: Addressing near-surface forecast biases: Outcomes of the ECMWF project 'Understanding uncertainties in surface atmosphere exchange' (USURF). ECMWF Tech. Memo. 875, 43 pp., https://www.ecmwf.int/en/elibrary/81202-addressing-near-surface-forecast-biases-outcomes-ecmwf-project-understanding.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Weyn, J. A., D. R. Durran, and R. Caruana, 2019: Can machines learn to predict weather? Using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *J. Adv. Model. Earth Syst.*, **11**, 2680–2693, https://doi.org/10.1029/2019MS001705.

——, ——, and ——, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002109, https://doi.org/10.1029/2020MS002109.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 627 pp.