Received: 4 March 2024 | Revised: 9 September 2024 | Accepted: 22 September 2024

DOI: 10.1002/qj.4875

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Medium-range predictability of temperature extremes and biases in Rossby-wave amplitude

Onno Doensen[1,2,3] | **Georgios Fragkoulidis**[1,4] | **Linus Magnusson**[5] | **Michael Riemer**[1] | **Volkmar Wirth**[1]

[1]Institute for Atmospheric Physics, Johannes Gutenberg University, Mainz, Germany

[2]Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland

[3]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

[4]Institute for Environmental Research and Sustainable Development, National Observatory of Athens, Athens, Greece

[5]European Centre for Medium-Range Weather Forecasts, Reading, UK

Correspondence
Onno Doensen, Climate and Environmental Physics, Physics Institute, University of Bern, Switzerland.
Email: onno.doensen@unibe.ch

Funding information
Transregional Collaborative Research Center, Grant/Award Number: SFB/TRR165; Deutsche Forschungsgemeinschaft, Grant/Award Number: 445572993; Swiss National Science Foundation, Grant/Award Number: IZCOZ0-205416

Abstract

This study investigates the medium-range predictability of warm and cold extremes in the Northern Hemisphere and the role that upper-tropospheric circulation biases play in this regard. Deterministic ERA5 reforecasts for the period 1979–2019 are evaluated based on the ERA5 reanalysis of the respective period, thus providing a large sample for verification and bias identification. The predictability of temperature extremes at 850 hPa is assessed based on the Gilbert Skill Score and other metrics and is shown to exhibit regional and seasonal variations. Summer is generally characterized by lower forecast skill scores than winter for both warm and cold extremes. Moreover, cold extremes in summer have slightly lower skill scores than warm extremes, while the opposite is true in winter. Biases in the frequency of temperature extremes are, to some extent, consistent with biases in mean temperature and indicate an underestimation in the total amount of extremes for much of the hemisphere in summer. Associated with the latter, biases also emerge in the standard deviation of the daily temperature distribution, with the summer values being largely underestimated over most of the hemisphere. The role of upper-tropospheric circulation in these biases is then assessed by verifying the representation of Rossby-wave packet (RWP) properties. It is found that the amplitude of RWPs is systematically underestimated in most of the hemisphere in summer, while it is overestimated in many parts of the midlatitudes in winter. Overall, the results suggest that the underestimation of RWP amplitude in summer hinders the medium-range predictability of temperature extremes in the explored retrospective and operational forecasts. Although operational European Centre for Medium-Range Weather Forecasts (ECMWF) forecasts gradually improve between 2013 and 2022 in terms of the 850-hPa temperature and 300-hPa RWP amplitude absolute errors, the aforementioned summer biases remain qualitatively similar.

KEYWORDS

predictability, reforecasts, Rossby waves, temperature extremes

1 | INTRODUCTION

The increasing frequency of extreme weather events calls for research efforts to focus on understanding their physical drivers better and improving the quality of their forecasts (Robinson *et al.*, 2021; Sillmann *et al.*, 2017). Early and accurate warnings of these events are crucial in mobilizing the authorities and mitigating societal impacts. Nevertheless, basic aspects of the medium-range predictability of such high-impact weather events remain barely explored.

Utilizing a modern numerical weather prediction (NWP) model version to issue retrospective forecasts (hereinafter referred to as reforecasts) for a sufficiently long period in the past is highly beneficial for both model development and process understanding. On the one hand, monitoring the model's performance on a large set of dates allows robust assessments of its overall skill and biases during typical and atypical weather conditions. On the other hand, evaluating the reforecasts of specific weather events or atmospheric flow configurations may hint at the physical mechanisms at play and unveil variations in practical predictability between regions, seasons, and lead times. Despite the clear potential of reforecast datasets, their availability and utilization in medium-range predictability studies are so far rather limited.

Wulff and Domeisen (2019) investigated reforecasts from several models for the period 1999–2010 and found that summer warm extremes in Europe are characterized by higher medium-range predictability than cold extremes. Lavaysse *et al.* (2019) examined European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble reforecasts for the period 1995–2015 and found that winter cold waves are predicted more accurately than summer heat waves. The emergence of systematic errors in ECMWF forecasts in the Northern Hemisphere was recently investigated in terms of winter jet-stream position in Vitart *et al.* (2022) and summer tropospheric temperatures in Magnusson *et al.* (2022).

Several studies have examined the link between the upper-tropospheric flow properties and weather extremes, and the concept of Rossby-wave packets (RWPs) has proved very useful in this regard (Fragkoulidis *et al.*, 2018; Grazzini *et al.*, 2021; Wirth *et al.*, 2018). The magnitude and duration of temperature extremes, in particular, have been found to be associated with the amplitude and phase speed of RWPs (Fragkoulidis & Wirth, 2020). The correct representation of RWPs in NWP models is thus crucial (Quinting & Vitart, 2019). The nonlinear dynamics governing the evolution of RWPs makes their medium-range prediction highly sensitive to small-scale errors at short lead times. If errors in their properties (e.g., amplitude, phase, phase speed, group velocity) grow substantially, then RWPs effectively transmit erroneous forecast "signals" of synoptic scale downstream (Lojko *et al.*, 2022; Magnusson, 2017). Crucially, part of these errors may be systematic. Recent analyses on operational forecasts have reported negative biases in the amplitude of Rossby waves during winter, possibly attributed to the too coarse model resolution and insufficient representation of physical processes (Gray *et al.*, 2014; Harvey *et al.*, 2018; Martínez-Alvarado *et al.*, 2016). Specifically for the Euro-Atlantic region, Matsueda and Palmer (2018) utilized operational and retrospective forecasts and found that winter exhibits wavy regimes much too frequently. We are aware of no published work on the predictability of Rossby-wave amplitude during summer.

Several questions regarding the representation of temperature extremes and RWPs in modern NWP models remain open. What is the spatiotemporal variability in the medium-range predictability of warm and cold extremes? How does the local temperature distribution evolve with lead time? Does the skill of reforecasts improve over time? Do biases in RWP properties emerge in reforecasts and operational forecasts and do they affect the predictability of temperature extremes? This study aims to address these questions by employing reforecasts from ECMWF and the National Oceanic and Atmospheric Administration (NOAA) spanning the 1979–2019 and 2000–2019 periods, respectively, as well as ECMWF's operational forecasts for the 2013–2022 period.

The remainder of the article is organized as follows. The data and methods employed are presented in Section 2. Section 3 assesses the ECMWF Reanalysis version 5 (ERA5) model performance in forecasting warm and cold extremes and reports on the regional and seasonal variability in a number of skill metrics. Biases in basic properties of the temperature distribution are evaluated in Section 4. Section 5 reports biases in Rossby-wave amplitude and explores their effect on the predictability of temperature extremes. Finally, Section 6 draws the conclusions of the study and provides further remarks. Additional analyses that are related to the main outcomes of this study are provided in the Supporting Information for reference.

2 | DATA AND METHODS

2.1 | Data

Primarily, this study employs deterministic ERA5 reforecast data of temperature (T) at 850 hPa and meridional wind (v) at 300 hPa for the 1979–2019 period. These forecasts have been produced retrospectively by the ECMWF using the ERA5 model (Hersbach *et al.*, 2020), that is,

version Cy41r2 of the Integrated Forecasting System (IFS; operational from March 8–November 22, 2016) and the ERA5 reanalysis as initial conditions. The subset used in this study includes forecasts issued daily at 0000 and 1200 UTC with 12-hourly lead times up to +240 h and stored on a regular latitude–longitude grid of 2° × 2° horizontal resolution. For comparison, we also employ NOAA's Global Ensemble Forecast System Version 12 (GEFSv12) reforecasts issued daily at 0000 UTC for the period 2000–2019 (Hamill *et al.*, 2022), as well as ECMWF's operational forecasts issued daily at 0000 and 1200 UTC for the period 2013–2022. The verification of all forecasts is carried out uniformly based on the corresponding ERA5 reanalysis fields.

2.2 | Definition of extreme temperature events

In order to minimize small-scale features (e.g, land–sea contrasts and small-scale topographic effects) and focus on the synoptic-scale evolution of the temperature field, we identify events of extreme temperature at the 850-hPa isobaric level (Fragkoulidis *et al.*, 2018). As a first step, we compute the climatological annual cycles of the 10th and 90th percentiles of 850-hPa temperature based on reanalysis data for the 1979–2019 period. The annual cycle of every grid point is computed separately for the 0000 and 1200 UTC time series as follows. Each day of the year is represented by a probability distribution that comprises all temperature values in the 21-day windows centred around it in every year. The nth percentile for a given day is then computed robustly based on this distribution, which consists of 861 data points (21 values from each of the 41 available years). After repeating this for every day in the year, the resulting climatological annual cycle is smoothed via a Fourier-series expansion and restriction to frequencies 0–4 year^{-1}.

Time instances of extreme temperature are then defined as follows. *Warm* extremes are defined as those time instances when 850-hPa temperature exceeds the 90th percentile. *Cold* extremes are defined as those time instances when 850-hPa temperature is lower than the 10th percentile. This identification procedure is followed for both reanalysis and forecast data, that is, extreme days in forecasts are defined based on the reanalysis percentile annual cycles.

Given this definition of temperature extremes, each season in the 12-hourly ERA5 reanalysis dataset (daily time instances at 0000 and 1200 UTC) will have around 18 warm and 18 cold extremes on average, out of a sample size of 180. For reference, the sample size of a single season for a given lead time (e.g., Day +5) is also 180 in the ERA5 reforecasts (two forecast initializations per day), while the 41-year total sample size of each season is about 7380. In the case of GEFSv12, the corresponding sample size of a single season is 90 (one forecast initialization per day) and the 20-year total sample size of each season is about 1800. The threshold to define extreme events is subjective in any case, and depends on the objectives of a study (Bouallègue *et al.*, 2019). In this study we opt to label instances in the lower and upper 10% tails of the distribution as extreme events, such that the forecast verification accounts for neither too large nor too small a sample of dates. With the chosen thresholds, the focus remains on high temperature anomalies and the statistical analyses are sufficiently robust.

2.3 | Forecast skill metrics

In the analyses that follow, forecast errors are computed as the deviations of the forecast fields from the reanalysis fields valid at the respective time. Moreover, the Gilbert Skill Score (GSS) is used to evaluate the deterministic skill of the model at correctly predicting the occurrence of observed 850-hPa temperature extremes at a specific lead time (Hogan & Mason, 2011). In particular, the GSS formula is defined as

$$GSS = \frac{h - h_c}{h + m + f - h_c}, \tag{1}$$

where h denotes the number of hits, that is, cases when both the reforecast and reanalysis identify an extreme instance, m denotes the number of misses, that is, cases when an extreme instance is only identified in reanalysis, f denotes the false alarms, that is, cases when an extreme instance is only identified in the reforecast, and

$$h_c = \frac{(h + m)(h + f)}{n}, \tag{2}$$

denotes the correction term that accounts for random hits, given by the number of forecast events times the observed event frequency (n denotes the total number of verified forecast instances). A GSS value of one indicates a perfect model for extremes (i.e., no misses or false alarms), while GSS=0 indicates a model that is equally capable with random predictions.

In addition, the frequency bias score (FBS) is used in order to assess whether and to what extent the occurrence frequency of the events of interest in reforecasts agrees with the one in reanalysis. The FBS is thus given by the ratio of the extreme events count in reforecasts over that in reanalysis:

$$FBS = \frac{h + f}{h + m}, \tag{3}$$

where $FBS = 1$ denotes that the amount of extremes in reforecasts at a specific lead time is the same as in reanalysis, $FBS < 1$ denotes an underforecast event, and $FBS > 1$ denotes an overforecast event.

2.4 | Diagnosis of rossby-wave amplitude

The upper-tropospheric meridional wind (v) field is well-suited for the diagnosis of RWP properties, since the succession of northerlies and southerlies at synoptic scales directly reflects the notion of Rossby waves along the jet (Chang, 1993). The amplitude of these waves is a highly dynamic field that typically exhibits eastward-propagating features of pronounced amplitude, namely wave packets. These RWPs span several thousand kilometres in the zonal direction and their spatially varying amplitude is effectively diagnosed as the two-dimensional "envelope" function (E) of v. In this study, E is computed by applying a Hilbert transform to latitude circles of the v field at 300 hPa, as originally proposed by Zimin *et al.* (2003). This is done in spectral space by zeroing-out the negative wavenumbers, doubling the positive ones, and, finally, performing an inverse Fourier transform to the spatially smoothed v' field (Fragkoulidis & Wirth, 2020). This results in the complex-valued *analytic signal* of v', the modulus of which yields the local in space and time RWP amplitude. Moreover, the argument of the analytic signal can be exploited to diagnose the local phase and phase speed of RWPs as described in Fragkoulidis and Wirth (2020) and Fragkoulidis (2022), results on which are presented in the Supporting Information for reference.

3 | PREDICTABILITY OF TEMPERATURE EXTREMES

Figure 1 shows the GSS values of Northern Hemisphere warm and cold extremes at a forecast lead time of +5 days in December–January–February (DJF) and June–July–August (JJA). Figure 2 shows the evolution of

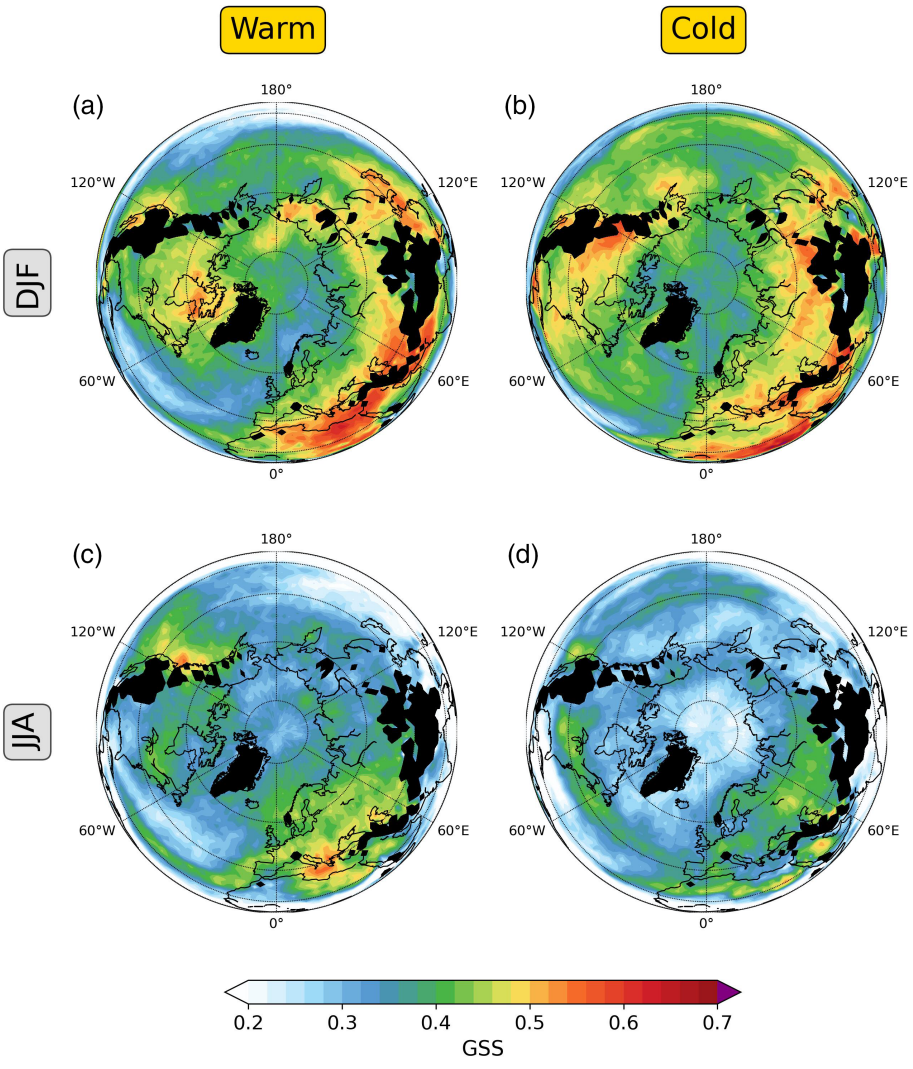

FIGURE 1 Gilbert Skill Score values in ERA5 Day +5 reforecasts of warm (left column) and cold (right column) extremes at 850 hPa in the DJF (upper row) and JJA (lower row) seasons of the 1979–2019 period. Grid points where the multi-annual minimum surface pressure is smaller than 850 hPa (i.e., the 850-hPa isobar is below ground) are masked in black.

midlatitude (30°N –70°N) GSS values with lead time up to Day +10 for both seasons and types of extremes (see also Figures S1 and S2). The expected decrease in the number of hits and increase in the number of misses and false alarms with lead time explain the gradual decrease of GSS values for both event types and seasons. However, pronounced seasonal and regional differences do emerge and are worth reporting. Although winter exhibits higher temperature absolute errors than summer (Figure S8), forecasts of both warm and cold extremes are generally characterized by lower skill (GSS) in summer than in winter. Moreover, cold extremes in summer appear to have slightly lower skill scores than warm extremes, while the opposite is true in winter. These outcomes also apply to the Probability of Detection (i.e., the ratio of hits over the total number of observed events) maps (not shown).

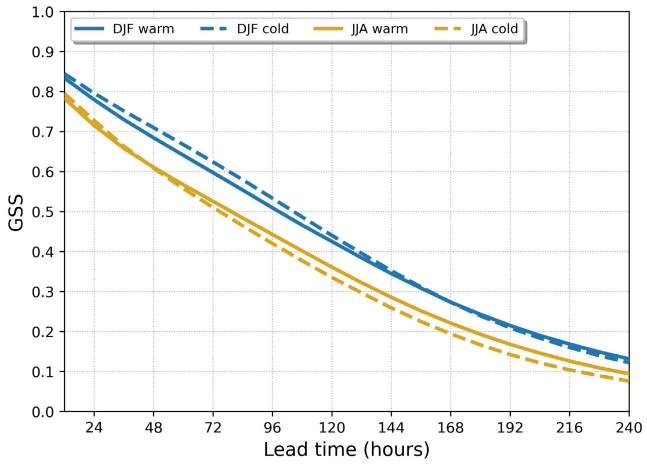

FIGURE 2 Evolution of midlatitude ERA5 reforecast Gilbert Skill Score (GSS) with lead time for warm (solid lines) and cold (dashed lines) extremes at 850 hPa in the DJF (blue) and JJA (orange) seasons of the 1979–2019 period. The GSS values correspond to a weighted average over the 30°N –70°N latitude band.

For reference, throughout the hemisphere the GSS values in GEFSv12 reforecasts are clearly lower than the ones of ERA5 reforecasts presented here (Figure S5). The implied drop in predictability from winter to summer in both models is consistent with the outcomes of Lavaysse *et al.* (2019), whereas the slightly higher scores of warm over cold extremes in summer ERA5 reforecasts is consistent with Wulff and Domeisen (2019). It

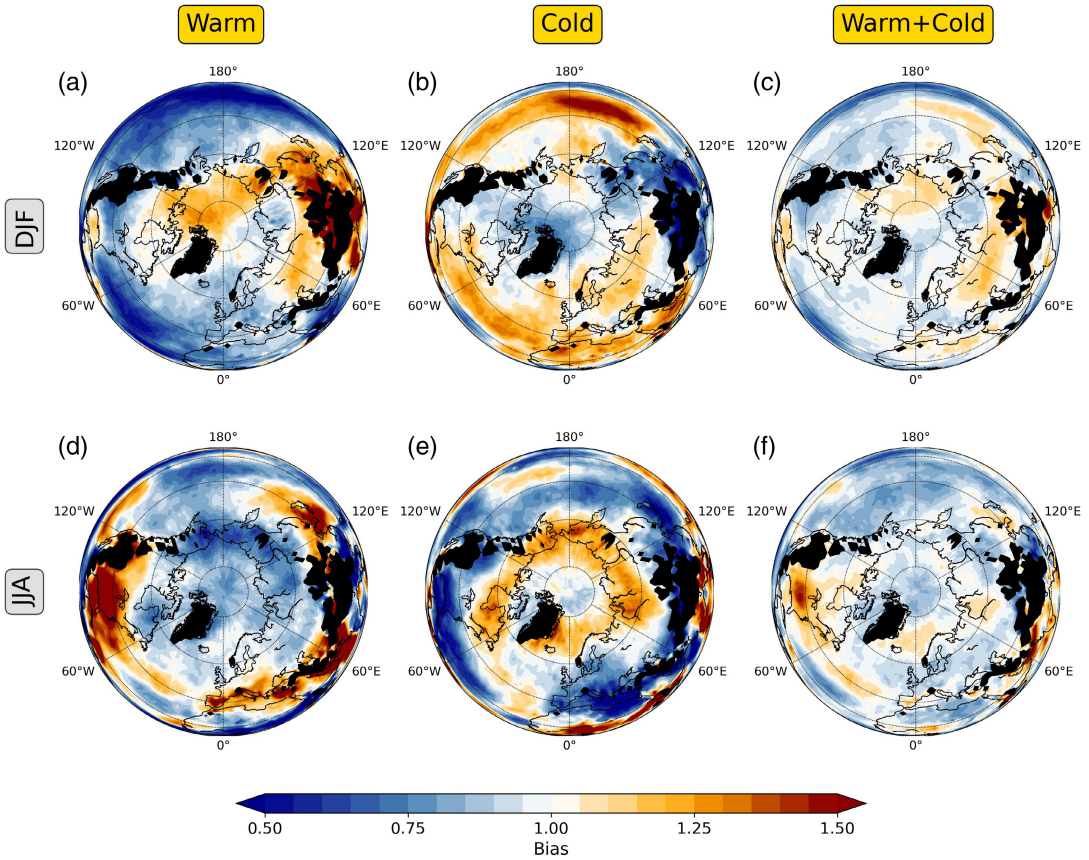

FIGURE 3 Frequency bias scores in ERA5 Day +5 reforecasts for warm (left column), cold (middle column), and all (right columns) extremes at 850 hPa in the DJF (upper row) and JJA (lower row) seasons of the 1979–2019 period.

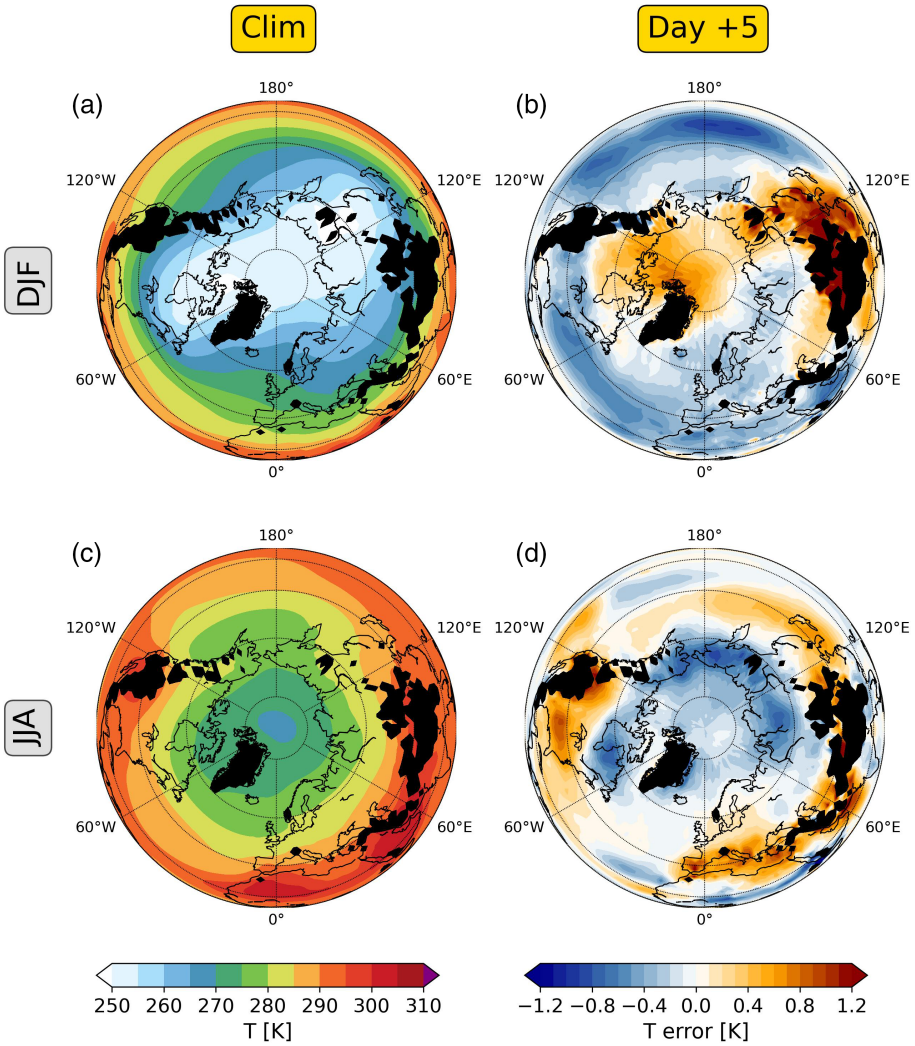

FIGURE 4 Mean 850-hPa temperature (left column) and temperature error of ERA5 Day +5 reforecasts (right column) in the DJF (upper row) and JJA (lower row) seasons of the 1979–2019 period.

should be noted, however, that our analysis is restricted to assessing the occurrence of daily temperature extremes and not the predictability of persistent event properties like their onset, intensity, and duration (e.g. Pyrina & Domeisen, 2023).

Since we define temperature extremes in forecasts based on the climatological distribution of reanalysis data and not the forecast model itself, frequency biases of the extremes potentially emerge. Figure 3 shows the DJF and JJA frequency bias scores of warm, cold, and all (either warm or cold) extremes at a lead time of five days. DJF is characterized by a negative (positive) frequency bias in warm (cold) extremes over the North Pacific and North Atlantic oceans, with the opposite behaviour characterizing parts of Asia. A pronounced pattern that emerges in JJA is the negative (positive) bias of cold extremes in the subtropics (high latitudes) and warm extremes in high latitudes (subtropics). These patterns remain rather steady with lead time, as evidenced in Figures S6 and S7.

The aforementioned frequency biases indicate whether misses or false alarms are the primary factors behind the GSS decrease with lead time. Areas where the forecast model generates too many (few) extremes typically experience an increased number of false alarms (misses). Regarding the bias patterns, it is evident that areas exhibiting cold extreme frequency biases in DJF typically exhibit warm extreme frequency biases of similar magnitude and opposite sign. This suggests that a shift in the temperature distribution toward higher or lower temperatures may be at play. In contrast, the total amount of extremes in JJA is substantially underestimated in much of the Northern Hemisphere, suggesting a narrowing of the temperature distribution. Areas of eastern North America constitute a prominent exception to this, as they exhibit an overestimation of extremes in JJA.

4 | FORECAST BIASES IN THE TEMPERATURE DISTRIBUTION

As already implied, the aforementioned frequency bias patterns may reflect systematic errors in the properties

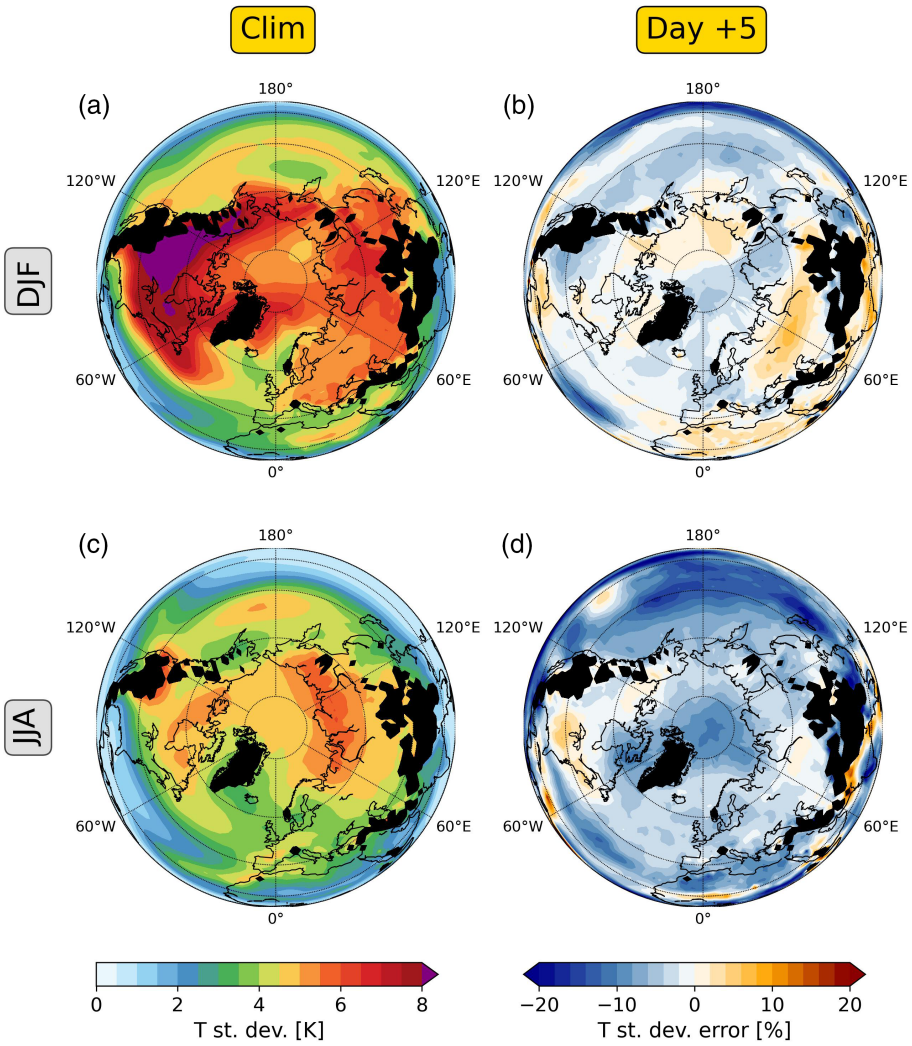

FIGURE 5 Standard deviation of daily 850-hPa temperature in ERA5 reanalysis (left column) and its error in Day +5 ERA5 reforecasts (right column) in the DJF (upper row) and JJA (lower row) seasons of the 1979–2019 period.

of the 850-hPa temperature distribution. In this section we assess the DJF and JJA forecast biases in the mean and standard deviation of the daily temperature distribution and compare the skill of the ERA5 reforecasts with that of GEFSv12 reforecasts and ECMWF operational forecasts.

The results shown in Figure 4 reveal that Day +5 biases in mean 850-hPa temperature are, to some extent, consistent with biases in the occurrence frequency of temperature extremes. An overestimation or underestimation of mean temperature is typically associated with corresponding shifts in the entire temperature distribution and its tails, such that biases in the amount of extremes emerge. For instance, the JJA overestimation of mean temperature in southern Europe is in principle consistent with the reported overestimation of warm extremes and underestimation of cold extremes (Figure 3). Moreover, Figure 4 shows that the circumglobal meridional temperature gradient tends to increase in JJA, due to the underestimation of temperature at higher latitudes and overestimation of temperature in the midlatitudes. This pattern seems to reverse in DJF, albeit with larger zonal asymmetries (see also Figure S8).

Figure 5 shows the relative error of the 850-hPa temperature standard deviation in ERA5 Day +5 reforecasts with respect to the one in reanalysis. A striking outcome in this analysis is that the width of the 850-hPa temperature distribution in JJA already exhibits a widespread underestimation on Day +1 (Figure S10), which then grows with lead time over most of the hemisphere. This effect is more pronounced over the North Pacific ocean and is associated with the aforementioned frequency bias fields that reveal an underestimation in the amount of extremes for many parts of the hemisphere (Figure 3). This is arguably an important factor behind the overall worse skill scores of JJA compared with DJF (Figure 1). Moreover, the underestimation in temperature variability may be one of the drivers behind the aforementioned increase in the meridional temperature gradient (less northward warm-air advection and less southward cold-air advection

lead to warming of low latitudes and cooling of high latitudes). The DJF bias in the width of the 850-hPa temperature distribution is less widespread and less pronounced. No plausible relation to the mean temperature bias can be derived in this case.

The ERA5 reforecast bias in JJA temperature standard deviation is also apparent in GEFSv12 reforecasts for the 2000–2019 period, while no apparent discrepancy arises in DJF (Figure 6). Evidently, the ECMWF operational forecasts of the most recent decade (2013–2022) are also characterized by an underestimation of similar magnitude in the JJA temperature standard deviation. For reference, the seasonal-mean absolute errors in 850-hPa temperature in ERA5 reforecasts (see also Figure S9) are similar to ECMWF operational forecasts between 2013–2019 and systematically smaller than the GEFSv12 reforecasts. Finally, the gradual decrease of the temperature mean absolute error in ERA5 reforecasts is faster in DJF than in JJA, which suggests that the increase in the number and quality of observations over the 1979–2019 period is more influential in DJF forecasts. In contrast, the aforementioned ERA5 reforecast bias in the JJA temperature standard deviation shows no signs of weakening over the years.

5 | FORECAST BIASES IN THE AMPLITUDE OF ROSSBY WAVES AND IMPLICATIONS FOR THE PREDICTABILITY OF TEMPERATURE EXTREMES

The widespread underestimation of temperature variability in JJA is arguably associated, among other things, with model deficiencies regarding dynamical processes that involve the large-scale circulation. In this section we assess this hypothesis by exploring forecast errors in the upper-tropospheric circulation, with a focus on the RWP amplitude (E).

Figure 7 shows the climatological-mean E values in DJF and JJA as well as the Day +5 mean E error in ERA5 reforecasts. The evolution of E errors with lead time is shown in Figure S11. Evidently, an E overestimation in many parts of the midlatitudes (e.g., North Pacific, North Atlantic, Eurasia) emerges in DJF, while E is underestimated in parts of the Arctic and the subtropics. In contrast, a pronounced E underestimation characterizes JJA for most of the Northern Hemisphere. In this case, the JJA E biases are more prominent than the DJF ones in both absolute and relative (to climatology) terms.

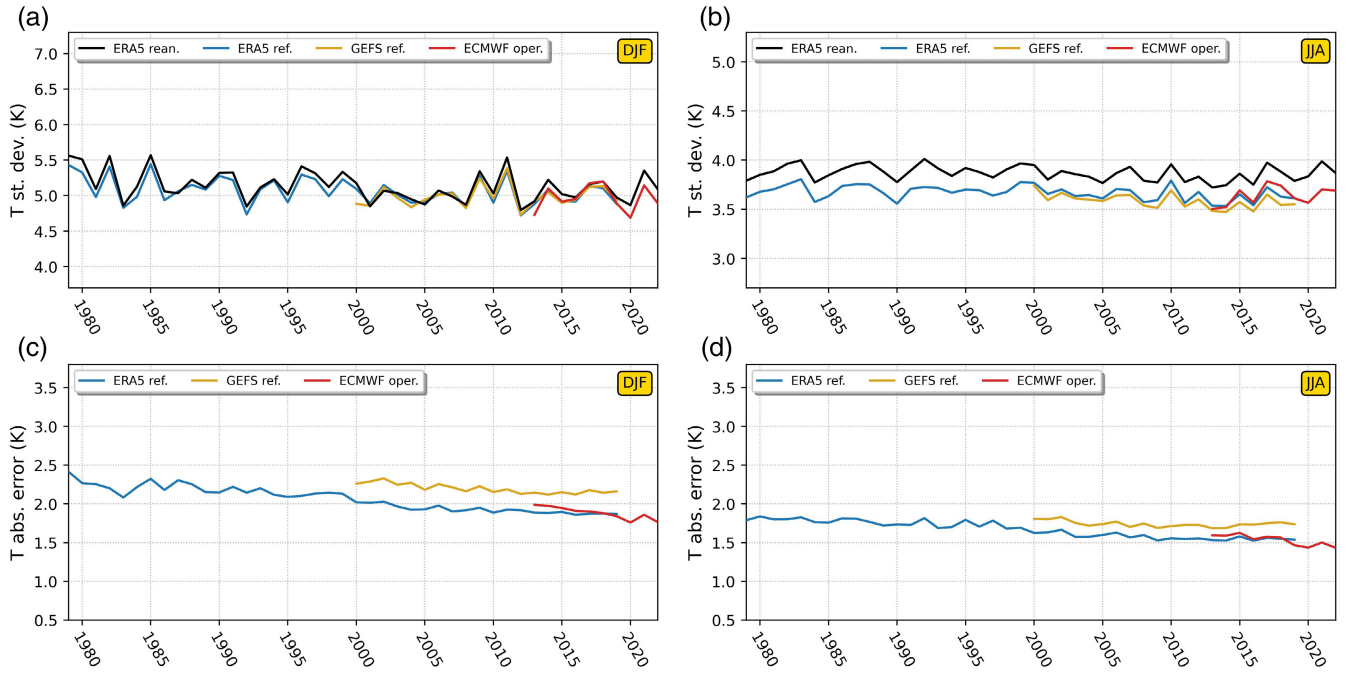

FIGURE 6 Upper row: annual evolution of the midlatitude 850-hPa temperature standard deviation in the ERA5 reanalysis (1979–2022; black) and the corresponding Day +5 forecast value in ERA5 reforecasts (1979–2019; blue), GEFS reforecasts (2000–2019; orange), and ECMWF operational forecasts (2013–2022; red) during (a) DJF and (b) JJA. Lower row: annual evolution of the midlatitude seasonal-mean 850-hPa temperature absolute error of Day +5 ERA5 reforecasts (1979–2019; blue), GEFS reforecasts (2000–2019; orange), and ECMWF operational forecasts (2013–2022; red) during (c) DJF and (d) JJA. All time series correspond to a weighted average over the 30°N –70°N latitude band.

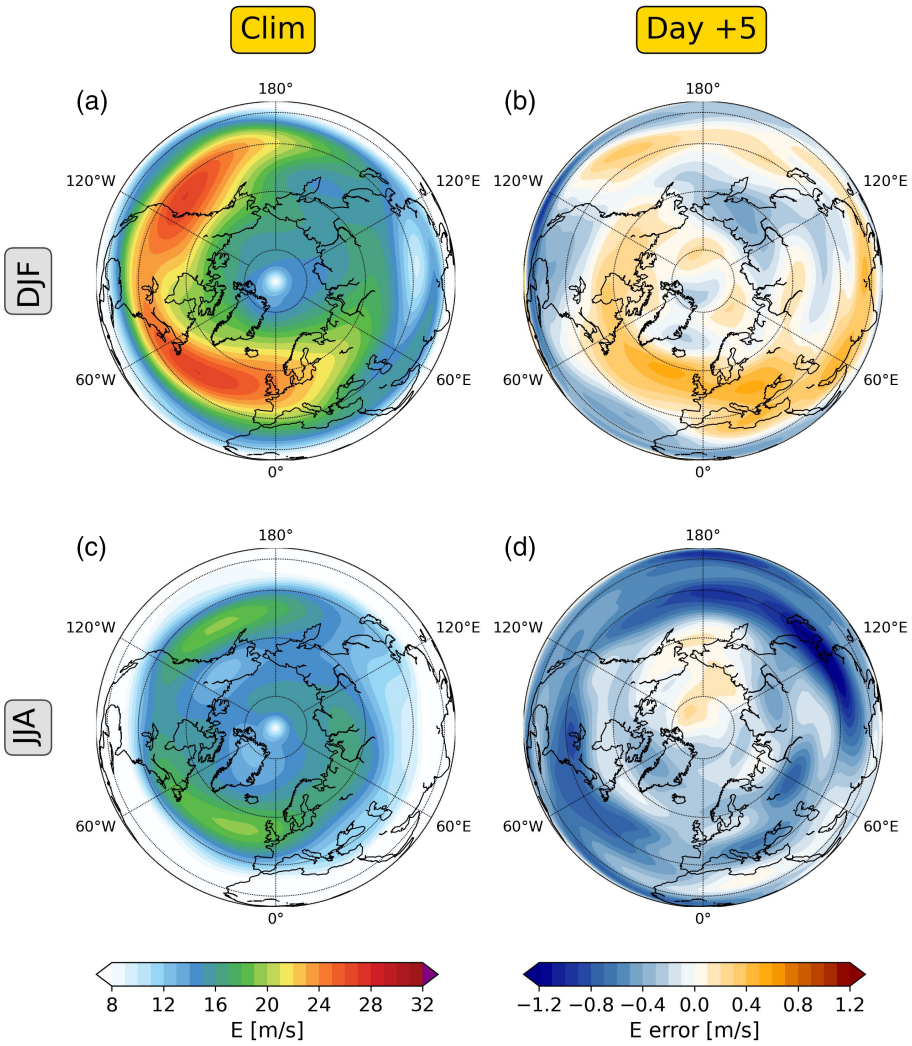

FIGURE 7 Mean 300-hPa E (left column) and E error of ERA5 Day +5 reforecasts (right column) in the DJF (upper row) and JJA (lower row) seasons of the 1979–2019 period.

The JJA E underestimation is already evident on Day +1, grows fast with lead time, and maximizes roughly over areas of the midlatitudes (e.g., North Pacific) where the temperature variability underestimation also maximizes (Figures 7 and 5). Since a causal relation between such forecast bias patterns can hardly be established, we merely hypothesize that one factor behind the pronounced temperature variability underestimation in JJA is the equally pronounced E underestimation.

For reference, Figures S12 and S13 show that, while E absolute errors are clearly larger in DJF, the E absolute errors scaled by the climatological E values are slightly larger in JJA. Interestingly, JJA is also characterized by more pronounced bias patterns than DJF in terms of the phase speed, the absolute phase index (diagnosed following Fragkoulidis (2022)), and the wavenumber of RWPs (Figures S15–S17). When it comes to the zonal wind velocity (Figure S14), the bias pattern of both seasons is more complex and does not indicate any clear relation to the E bias patterns.

Operational ECMWF forecasts and GEFSv12 reforecasts also exhibit an E underestimation in JJA (Figure 8). Although the bias of GEFSv12 reforecasts appears weaker than the ERA5 one, the former are characterized by systematically larger E absolute errors in both seasons. Examining the decadal evolution in the ERA5 reforecast errors and biases shows that the E negative bias in JJA is only slightly reduced between 1979 and 2019, while E absolute errors show a substantial decrease over time. As in the case of the temperature absolute errors (Figure 6), the latter decrease is more pronounced in DJF than JJA. As expected, the reduction in E absolute errors over the years is faster for the operational forecasts than the reforecasts, driven by improvements in both the observations and the model.

The Northern Hemisphere summer season, in particular, exhibits a clear relationship between the upper-tropospheric RWP amplitude and lower-tropospheric temperature anomalies (Fragkoulidis *et al.*, 2018). Motivated by that and the previous findings in this section, we now assess to what extent the systematic E forecast errors may hinder the predictability of 850-hPa temperature extremes. To this end, we compare the GSS values for time instances when E in ERA5 reforecasts is below

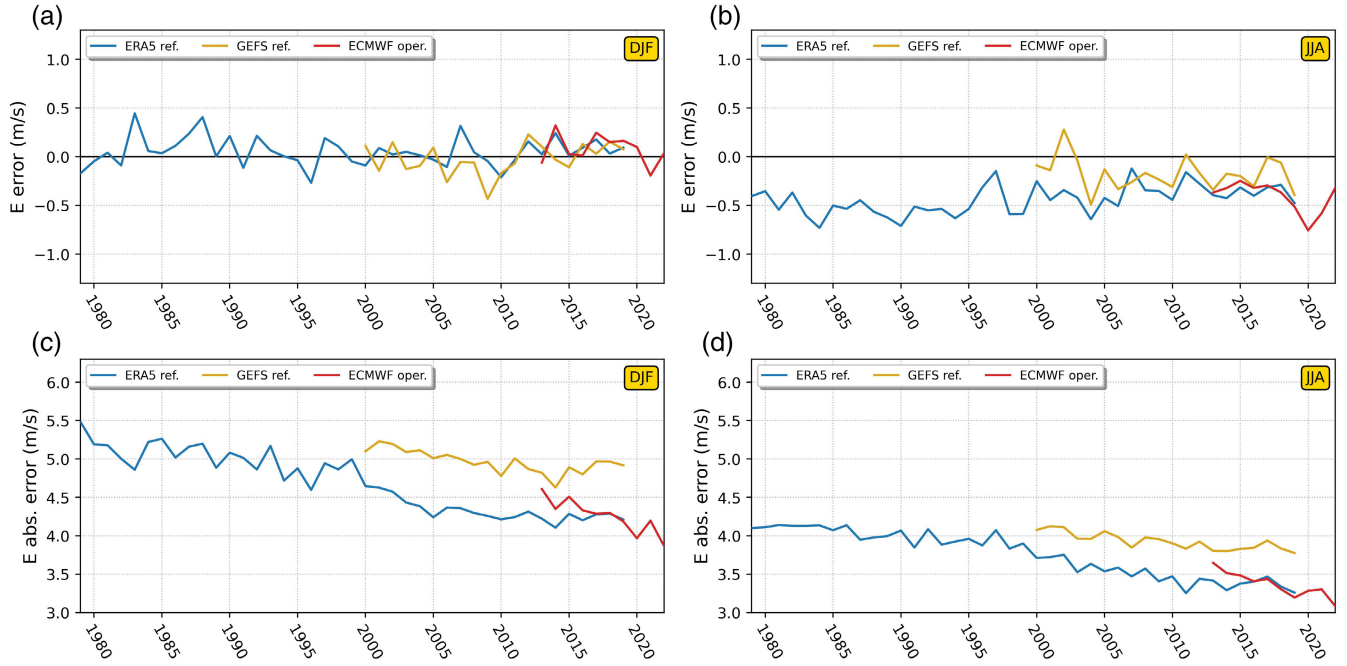

FIGURE 8 Upper row: annual evolution of the midlatitude seasonal-mean 300-hPa *E* error in the Day +5 ERA5 reforecasts (1979–2019; blue), GEFS reforecasts (2000–2019; orange), and ECMWF operational forecasts (2013–2022; red) during (a) DJF and (b) JJA. Lower row: annual evolution of the midlatitude seasonal-mean 300-hPa *E* absolute error of Day +5 ERA5 reforecasts (1979–2019; blue), GEFS reforecasts (2000–2019; orange), and ECMWF operational forecasts (2013–2022; red) during (c) DJF and (d) JJA. All time series correspond to a weighted average over the 30°N –70°N latitude band.

average with those time instances when *E* is above average. In order to compare samples of similar climatological conditions (i.e., the splitting should not be modulated by the fact that the mean *E* values vary within seasons) and equal size, we first compute climatological annual cycles of the 50th percentile of *E* (as described in Section 2 for temperature) for each lead time separately. The 12-hourly time series of *E* forecasts at a given season, lead time, and grid point is then split into two samples depending on whether or not *E* exceeds the corresponding 50th percentile.

Figure 9 shows the difference in Day +5 GSS between the two equally sized *E* samples (Figures S3 and S4 show the evolution with lead time). The results show that the low-*E* sample is characterized by lower GSS values than the high-*E* sample, which provides further support to the hypothesis that the negative *E* bias in JJA hinders the predictability of temperature extremes. The GSS decrease in low-*E* cases appears particularly pronounced for high-latitude warm extremes in DJF and midlatitude cold extremes in JJA. These extremes tend to occur with intense *E* positive anomalies (not shown), which explains the drop in their predictability when the RWP amplitude is relatively low (low-*E* sample).

As a final note, one could repeat this last analysis by comparing the GSS values for time instances when the *E* error is relatively low (i.e., cases of *E* underestimation) with those time instances when the absolute *E* error is relatively low (i.e., cases when *E* is well predicted). Again, GSS values in the first sample are lower than those in the second one in much of the hemisphere, so the same conclusions would be derived (figure not shown).

6 | CONCLUSIONS AND FURTHER REMARKS

In this study we analyzed retrospective and operational forecasts in order to investigate the medium-range predictability of Northern Hemisphere temperature extremes and the role of upper-tropospheric circulation biases in this regard. Evaluating ERA5 reforecasts for the period 1979–2019 allowed us to verify a large sample of warm and cold extremes at the 850-hPa isobaric level and identify forecast errors and biases that may hinder their predictability. Maps of the GSS exhibit pronounced regional and seasonal variability, with summer extremes generally associated with lower skill scores than winter extremes. Moreover, asymmetries emerge between warm and cold extremes; summer cold extremes are characterized by slightly lower predictability than warm extremes, while the opposite is true in winter. Forecast misses and false alarms of these extremes are partly attributed to

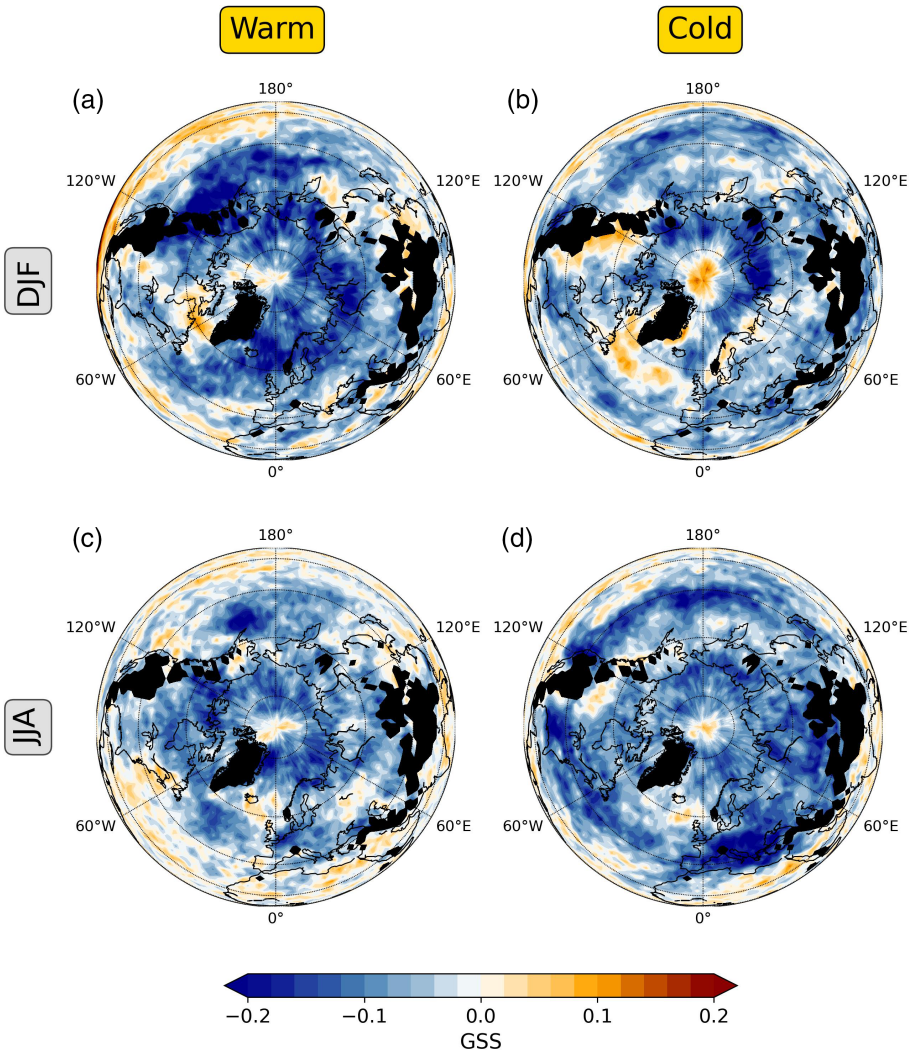

FIGURE 9 Gilbert Skill Score difference in ERA5 Day +5 reforecasts between cases with lower and greater than average E for warm (left column) and cold (right column) extremes at 850 hPa in the DJF (upper row) and JJA (lower row) seasons of the 1979–2019 period.

biases in the mean and standard deviation of the daily temperature distribution. A widespread underestimation of the summer temperature variability that grows with lead time constitutes a striking example of biases in the temperature distribution and is consistent with a negative bias in RWP amplitude at 300 hPa, which also characterizes most of the hemisphere. In contrast, the winter season features mean temperature biases that imply a reduced meridional temperature gradient and a RWP amplitude overestimation in many parts of the midlatitudes. It is then shown that time instances of lower than average RWP amplitude are associated with lower GSS than the ones with higher than average RWP amplitude. Overall, the results suggest that the underestimation of RWP amplitude in summer hinders the medium-range predictability of temperature extremes, while a pronounced seasonal variability in biases and predictability emerges. Finally, broadly similar biases to the ERA5 ones are also found in GEFSv12 reforecasts of the 2000–2019 period and ECMWF operational forecasts of the 2013–2022 period.

Tracing the root of the reported biases in mean temperature and RWP amplitude was beyond the scope of this study. The aim was rather to detect these biases, examine their spatiotemporal variability, and assess hypotheses about their role in the predictability of temperature extremes. Building on the results of this study, an interesting next step would be to explore the causes behind the pronounced negative RWP amplitude bias in JJA that seems to remain largely unaffected in recent years, while absolute errors in this respect gradually decrease. The baroclinically less unstable but convectively more unstable summer season is characterized by weather systems and flow configurations of smaller scale, while physical processes at the subgrid level become more important than in the winter season. As an example, the notoriously hard to resolve and predict summer deep convection can have far-reaching effects that grow with lead time by inducing strong divergent outflow that effectively perturbs the upper-tropospheric flow. If this process is misrepresented in NWP models, an underestimation of

Rossby-wave amplitude is to be expected (Schemm, 2023; Teubler & Riemer, 2021).

ACKNOWLEDGEMENTS

We acknowledge ECMWF and NOAA for freely providing the reanalysis and reforecast data used in this study and Johannes Gutenberg University Mainz for granting computing time on the supercomputer Mogon II (https://hpc.uni-mainz.de/, last access: 23 October 2023).

FUNDING INFORMATION

The research leading to these results has been performed within the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" funded by the German Science Foundation (DFG). O. Doensen has also been supported by the Swiss National Science Foundation under grant number IZCOZ0-205416 and G. Fragkoulidis by the DFG under grant number 445572993.

DATA AVAILABILITY STATEMENT

The ERA5 reanalysis and GEFSv12 reforecast data used in this study are freely available online. The ECMWF operational forecasts and ERA5 reforecasts are made available by ECMWF under license to authorized users. Processed data and code employed in the analyses presented can be provided by the authors upon request.

ORCID

Onno Doensen https://orcid.org/0000-0003-1281-3044
Georgios Fragkoulidis https://orcid.org/0000-0002-1767-4189
Linus Magnusson https://orcid.org/0000-0003-4707-2231
Michael Riemer https://orcid.org/0000-0001-6431-9537
Volkmar Wirth https://orcid.org/0000-0001-5611-8786

REFERENCES

Bouallègue, Z.B., Magnusson, L., Haiden, T. & Richardson, D.S. (2019) Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quarterly Journal of the Royal Meteorological Society*, 145, 1741–1755.

Chang, E.K.M. (1993) Downstream development of baroclinic waves as inferred from regression analysis. *Journal of the Atmospheric Sciences*, 50, 2038–2053.

Fragkoulidis, G. (2022) Decadal variability and trends in extratropical Rossby wave packet amplitude, phase, and phase speed. *Weather Clim. Dynam*, 3, 1381–1398. Available from: https://doi.org/10.5194/wcd-3-1381-2022

Fragkoulidis, G. & Wirth, V. (2020) Local Rossby wave packet amplitude, phase speed, and group velocity: Seasonal variability and their role in temperature extremes. *Journal of Climate*, 33, 8767–8787.

Fragkoulidis, G., Wirth, V., Bossmann, P. & Fink, A.H. (2018) Linking northern hemisphere temperature extremes to Rossby wave packets. *Quarterly Journal of the Royal Meteorological Society*, 144, 553–566.

Gray, S.L., Dunning, C.M., Methven, J., Masato, G. & Chagnon, J.M. (2014) Systematic model forecast error in Rossby wave structure. *Geophysical Research Letters*, 41, 2979–2987.

Grazzini, F., Fragkoulidis, G., Teubler, F., Wirth, V. & Craig, G.C. (2021) Extreme precipitation events over northern Italy. Part II: Dynamical precursors. *Quarterly Journal of the Royal Meteorological Society*, 147, 1237–1257.

Hamill, T.M., Whitaker, J.S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P. et al. (2022) The reanalysis for the global ensemble forecast system, version 12. *Monthly Weather Review*, 150, 59–79.

Harvey, B., Methven, J. & Ambaum, M.H. (2018) An adiabatic mechanism for the reduction of jet meander amplitude by potential vorticity filamentation. *Journal of the Atmospheric Sciences*, 75, 4091–4106. Available from: http://journals.ametsoc.org/doi/10.1175/JAS-D-18-0136.1

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, 1999–2049.

Hogan, R.J. & Mason, I.B. (2011) Deterministic forecasts of binary events. *Forecast Verification*, 31–59. Portico. http://doi.org/10.1002/9781119960003.ch3

Lavaysse, C., Naumann, G., Alfieri, L., Salamon, P. & Vogt, J. (2019) Predictability of the European heat and cold waves. *Climate Dynamics*, 52, 2481–2495.

Lojko, A., Payne, A. & Jablonowski, C. (2022) The remote role of North-American mesoscale convective systems on the forecast of a Rossby wave packet: A multi-model ensemble case-study. *Journal of Geophysical Research: Atmospheres*, 127, e2022JD037171.

Magnusson, L. (2017) Diagnostic methods for understanding the origin of forecast errors. *Quarterly Journal of the Royal Meteorological Society*, 143, 2129–2142.

Magnusson, L., Alonso-Balmaseda, M., Dahoui, M., Forbes, R., Haiden, T., Lavers, D. et al. (2022) Summary of the UGROW subproject on tropospheric temperature bias during JJA over the northern hemisphere. *ECMWF Technical Memoranda*, 891, 1–15.

Martínez-Alvarado, O., Madonna, E., Gray, S.L. & Joos, H. (2016) A route to systematic error in forecasts of Rossby waves. *Quarterly Journal of the Royal Meteorological Society*, 142, 196–210.

Matsueda, M. & Palmer, T.N. (2018) Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. *Quarterly Journal of the Royal Meteorological Society*, 144, 1012–1027.

Pyrina, M. & Domeisen, D.I.V. (2023) Subseasonal predictability of onset, duration, and intensity of European heat extremes. *Quarterly Journal of the Royal Meteorological Society*, 149, 84–101.

Quinting, J.F. & Vitart, F. (2019) Representation of synoptic-scale Rossby wave packets and blocking in the S2S prediction project database. *Geophysical Research Letters*, 46, 1070–1078.

Robinson, A., Lehmann, J., Barriopedro, D., Rahmstorf, S. & Coumou, D. (2021) Increasing heat and rainfall extremes now far outside the historical climate. *npj Climate and Atmospheric Science*, 4, 45.

Schemm, S. (2023) Toward eliminating the decades-old "too zonal and too equatorward" storm-track bias in climate models. *Journal of Advances in Modeling Earth Systems*, 15, e2022MS003482.

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L.V., Hegerl, G. et al. (2017) Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and Climate Extremes*, 18, 65–74.

Teubler, F. & Riemer, M. (2021) Potential-vorticity dynamics of troughs and ridges within Rossby wave packets during a 40-year reanalysis period. *Weather and Climate Dynamics*, 2, 535–559.

Vitart, F., Emerton, R., Rodwell, M., Alonso-Balmaseda, M., Haiden, T., Johnson, S. et al. (2022) Investigating biases in the representation of the pacific sub-tropical jet stream and associated teleconnections (a UGROW sub-project). *ECMWF Technical Memoranda*, 889, 1–20.

Wirth, V., Riemer, M., Chang, E.K.M. & Martius, O. (2018) Rossby wave packets on the midlatitude waveguide — a review. *Monthly Weather Review*, 146, 1965–2001.

Wulff, C. & Domeisen, D. (2019) Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers. *Geophysical Research Letters*, 46, 11520–11529.

Zimin, A.V., Szunyogh, I., Patil, D.J., Hunt, B.R. & Ott, E. (2003) Extracting envelopes of Rossby wave packets. *Monthly Weather Review*, 131, 1011–1017.

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: Doensen, O., Fragkoulidis, G., Magnusson, L., Riemer, M. & Wirth, V. (2024) Medium-range predictability of temperature extremes and biases in Rossby-wave amplitude. *Quarterly Journal of the Royal Meteorological Society*, 150(765), 5390–5402. Available from: https://doi.org/10.1002/qj.4875