Received: 17 November 2023 | Revised: 31 May 2024 | Accepted: 5 June 2024

DOI: 10.1002/qj.4801

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Investigating the medium-range predictability of European heatwave onsets in relation to weather regimes using ensemble reforecasts

Alexander Lemburg | **Andreas H. Fink**

Institute of Meteorology and Climate Research, Karlsruhe Institute ofTechnology (KIT), Karlsruhe, Germany

Correspondence
Alexander Lemburg, Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany.
Email: alexander.lemburg@kit.edu

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: SFB/TRR 165

Abstract

In this study, the medium-range predictability of heatwave (HW) onsets in four midlatitude European regions is investigated statistically with the help of ensemble reforecasts for the period 2001–2018. The concept of Euro-Atlantic weather regimes is adopted to characterise HWs (about 50 in each region) and to study whether forecast skill may depend on the large-scale dynamical setup. HW onsets over the British Isles and Scandinavia are mainly associated with Scandinavian and European blocking regimes, whereas the "no regime" case is observed more frequently for Central Europe. Stratified by weather regime, the predictability of heatwave onsets is then studied by means of a multiple metric-based analysis of European Centre for Medium-Range Weather Forecasts (ECMWF) and Global Ensemble Forecast System Version 12 (GEFSv12) ensemble reforecasts. For two of the regions considered, Central Europe and the British Isles, a conclusive picture is obtained: medium-range predictive skill is significantly higher for HW onsets associated with Scandinavian or European blocking compared with cases with no pronounced regime. This skill advantage mostly concerns the large-scale flow and, to some extent, 850-hPa temperatures, but is generally not reflected in the correct prediction of near-surface temperatures. Finally, we investigate for two regions how exceptionally good or poor forecasts are related to the atmospheric state during or shortly after forecast initialisation. At 10 days lead time, poor large-scale flow predictive skill for Central European HW onsets is linked to anomalously high baroclinicity further upstream and an intensified North Atlantic jet stream, whereas good forecasts on average feature an initial state close to climatology. Forecast skill for near-surface temperatures is not affected by such dynamical precursors, but rather by pre-existing soil-moisture anomalies. For the British region, exceptionally good forecasts of both large-scale flow and near-surface temperatures are associated with an already established continental blocking. In contrast to Central Europe, pre-existing soil-moisture anomalies play less of a role there.

KEYWORDS

dynamical meteorology, Europe, heatwaves, numerical weather prediction

1 | INTRODUCTION

Over the past 30 years, Europe has seen average temperatures rising at more than twice the pace of the global average's increase (WMO, 2022). In line with that trend, unprecedented extreme heatwaves (HWs) have been observed, which are projected to increase further in frequency and intensity in all parts of the continent (Collins *et al.*, 2013). Among all weather- and climate-related risks, HWs pose a major and often underestimated one (Hughes *et al.*, 2016). The HW of August 2003 with an estimated number of 70,000 excess deaths (Fink *et al.*, 2004; García-Herrera *et al.*, 2010), the extremely long-lasting heat (and accompanying drought) in western Russia in 2010 (Dole *et al.*, 2011; Trenberth & Fasullo, 2012), and the recurring hot and dry extremes affecting multiple European regions in both 2018 and 2019 (Kueh & Lin, 2020; Rousi *et al.*, 2023; Sousa *et al.*, 2020) are recent examples of the potentially devastating impact of HWs. Since adequate short-term countermeasures or adaption techniques exist, a timely and accurate prediction of HWs may greatly alleviate the toll on human life and society (Ebi *et al.*, 2004; Mücke & Litvinovitch, 2020).

Processes leading to HWs are manifold and depend strongly on the time-scale considered. Therefore, the challenges of predicting HWs and the very meaning of predictability itself vary considerably with forecast time (Domeisen *et al.*, 2023). At lead times of up to three days, weather models usually capture the evolution of the large-scale circulation reasonably well (Haiden *et al.*, 2021). The accurate prediction of near-surface temperatures then depends mostly on rather small-scale details of the flow field and the adequate simulation of cloudiness and near-surface diabatic heating (Lemburg & Fink, 2022). Beyond 14 days, probabilistic extended-range forecasts may—more skilfully than for any other types of weather—indicate an increased likelihood of HWs occurring by exploiting some predictability that is inherent to slowly varying boundary conditions, such as local or remote soil-moisture anomalies (Teng *et al.*, 2019; Wulff & Domeisen, 2019), as well as tropical modes of intraseasonal variability (Rouges *et al.*, 2023).

This study concentrates on what lies in between and what is still mostly an initial value problem: medium-range weather forecasts ranging from 5–12 days. Prior studies of European HWs at such lead times, based on probabilistic evaluation of reforecast ensembles, pointed out that the exact prediction of the onset is most challenging (Lavaysse *et al.*, 2019) and that region-specific differences in predictability may exist (Pyrina & Domeisen, 2023). In this article, we therefore aim to extend the existing literature, focusing on the predictability of the onset of HWs in different European regions. In contrast to most prior studies, our assessment not only considers maximum near-surface temperature forecasts, but especially takes into account the large-scale dynamics aspect within a multi-metric-based approach.

Indeed, at our lead times of interest, the successful prediction of weather extremes relies strongly on the adequate representation of large-scale Rossby-wave dynamics (Fragkoulidis & Wirth, 2020; Grazzini & Vitart, 2015; Wirth *et al.*, 2018). In the northern midlatitude regions of Europe, lasting HWs are often associated with a substantial blocking of the large-scale atmospheric flow due to amplified and/or breaking Rossby waves (Kautz *et al.*, 2022; Kueh & Lin, 2020; Pfahl & Wernli, 2012; Schaller *et al.*, 2018). Although state-of-the-art numerical weather prediction (NWP) models nowadays generally do considerably well in predicting the large-scale flow evolution up to lead times of around six days, severe forecast busts still occur (Rodwell *et al.*, 2013). Especially, failures in predicting the formation of a blocking weather regime can often be traced back to a common pathway of upscale and downstream error propagation (Baumgart *et al.*, 2019).

The initial source of uncertainty is frequently linked to errors in upstream diabatic processes stemming from the usually unresolved convective-scale and/or cloud microphysics. For spring and autumn cases of forecast busts, Grams *et al.* (2018) pointed to an important role of the release of latent energy in warm conveyor belts for the amplification of atmospheric flow waviness and the possible growth and propagation of forecast errors. With the help of numerical simulations, Steinfeld *et al.* (2020) provided further evidence for the significant contribution of latent heating to the generation of anticyclonic circulation anomalies over the Euro-Atlantic sector. For a summer HW case in 2017, Lojko *et al.* (2022) present evidence for the high sensitivity of ridge amplification over Europe to activity of mesoscale convective systems over the contiguous United States.

While these studies point at probable common causes of poor forecasts of blocked weather regimes, the complexity of downstream error growth makes for a large case-to-case variability. We therefore do not study such error growth mechanisms in detail. Instead we opt for a more statistical approach by considering a large number of HW cases (around 50 each) in the period 2001–2018 in four different midlatitude regions of Europe, which are depicted in Figure 1. Before we address the medium-range predictability of HW onsets, we investigate in detail the dynamical regimes under which HWs form in different European regions. To do so, we make use of the concept of year-round Euro-Atlantic weather regimes. Based on empirical orthogonal function (EOF) analysis and subsequent *k*-means clustering, this metric reduces the complexity of the atmospheric flow field by projecting it onto

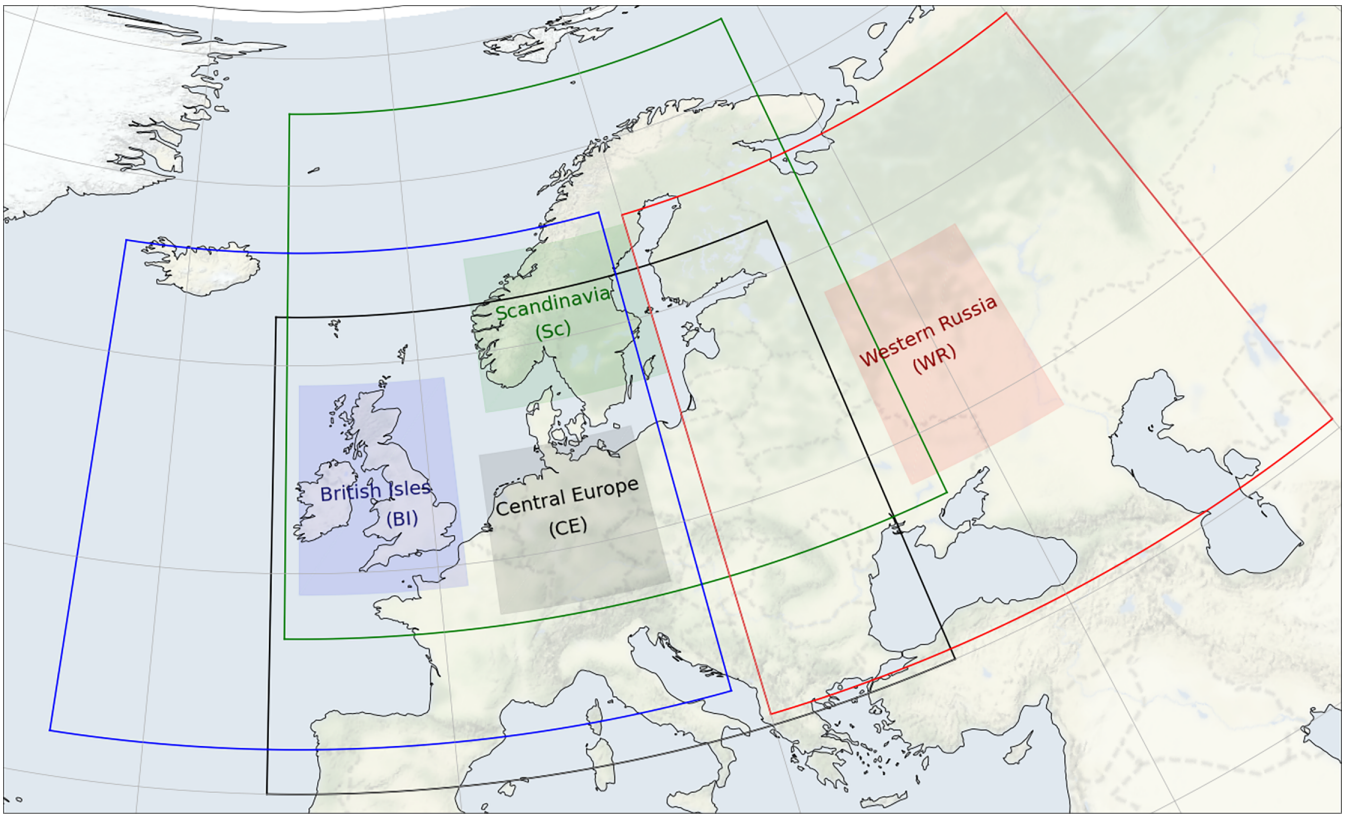

FIGURE 1 Research domain of this study. The small boxes with colour shadings represent the actual regions of interest for which HWs were evaluated. Moreover, these boxes denote the area over which all temperature-based metrics such as the nRMSE of 850-hPa temperatures and the relative error in Tmax-EFI are calculated. The corresponding larger boxes depict the areas over which large-scale synoptic metrics such as the ACC and nRMSE for 500-hPa geopotential are computed for the respective region. Latitude/longitude coordinates of the depicted boxes are provided in Table S1 in the Supporting Information. [Colour figure can be viewed at wileyonlinelibrary.com]

the seven main modes of synoptic-scale variability in this domain (Grams *et al.*, 2017). Year-round climatologies of Euro-Atlantic weather regimes are available and the forecast skill of NWP and extended-range models has also been assessed extensively in the context of these regimes previously (Büeler *et al.*, 2021; Osman *et al.*, 2023), including the summer season.

In the present study, we therefore take a complementary approach, in which we will not focus on the predictability of these weather regimes per se. Instead, we are more interested in the intricate details of the predicted flow fields that may be of particular importance for adequately predicting HW onset. Using reforecast ensembles of the two state-of-the art NWP models European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) and Global Ensemble Forecast System Version 12 (GEFSv12), we will study multiple aspects of forecast quality in detail in relation to HW onsets over a range of lead times from 3–12 days for the four above-mentioned European regions. Multiple commonly used evaluation metrics, such as 500-hPa geopotential anomaly correlation coefficients and root-mean-square errors, are employed to answer the question of whether HWs associated with a certain weather regime are predicted better than those forming under a different regime.

In a second part of this article, we further explore the possible causes of exceptionally good or poor predictability of HW onsets, this time with a focus on Central Europe and the British Isles. Our aim is to robustly identify anomalies in the atmospheric state during or shortly after forecast initialisation that may render a medium-range forecast of HW onset more or less successful. Of particular interest are the characteristics of the flow further upstream (e.g., baroclinicity or the state of the North Atlantic jet) and the antecedent Euro-Atlantic weather regimes, as well as pre-existing soil-moisture anomalies.

For the most part, in this article the term *predictability* refers to the so-called *practical predictability* or, more precisely, the respective predictive skill of the two models used. However, if both models show substantially enhanced predictive skill for a certain regime or with respect to a particular initial state, this may also suggest higher *inherent predictability*—but not necessarily so. We will therefore refrain from using such more specific terms

and use the word *predictability* in a rather broad sense throughout the article.

This article is structured as follows. In Section 2, we introduce the forecast data used and the concept of weather regimes, and describe the objective detection of HWs further, as well as the applied forecast evaluation metrics. Section 3 features the multi-metric statistical evaluation of HW-related forecast quality in four different European regions, stratified by weather regime during HW onset. Section 4 demonstrates, for Central European and British HW onsets, how the predictability of both the large-scale flow evolution and regional near-surface temperature during HW onsets might be linked to anomalies in the initial state of the atmosphere or soils. Section 5 contains a summary and discussion of the findings of this study and puts them into the context of previous, current, and suggested further research.

2 | DATA AND METHODS

2.1 | 2001–2018 reforecasts

As NWP models are frequently updated, their forecast skill will change over time and the use of operational forecasts is therefore not suited for investigating predictability over longer time periods. Reforecasts (also known as hindcasts) circumvent this problem by using a fixed NWP model version for the generation of a reforecast ensemble covering the entire time period investigated, which makes it possible to systemically compare forecasts for the August 2003 HW with those for the July 2018 HW. To study the predictability of HW onsets, we therefore use two reforecast ensembles of two state-of-the art weather forecast models: primarily the ensemble reforecasts from ECMWF and ensemble reforecasts generated at National Centers for Environmental Prediction (NCEP) based on the GEFSv12 model.

2.1.1 | ECMWF reforecasts

The reforecasts from ECMWF are published as part of the Subseasonal to Seasonal Research (S2S) project (Vitart *et al.*, 2017) and have been evaluated (Vitart *et al.*, 2019) and used to assess predictability of HWs and many other phenomena before (Ferranti *et al.*, 2018; Vitart & Robertson, 2018). Whereas the intended use case is obviously probabilistic weather prediction beyond the deterministic time-scale, there is no reason not to use ECMWF reforecasts for investigating medium-range predictability. Reforecasts at the ECMWF are produced on the fly: twice a week, a reforecast ensemble for the last 20 years is run, with the current date as the respective initialisation date for each year. Each individual reforecast simulation features 11 ensemble members and is integrated until forecast day 46. Up to forecast day 15 (we do not consider longer lead times in this study), the model is run at a spectral resolution of T639 ($\approx$ 16 km) with 91 vertical levels. Output is available at a spatial resolution of 1.5° and a temporal resolution of 24 h for 0000 UTC for instantaneous variables and 0000–2400 UTC for accumulated or maximum variables.

For our use case, a downside of this product is the limitation to two initialisations per week. Therefore, for any kind of HW onset date, one would have available only a limited range of lead times. Thus, there is no way to generate a homogeneous dataset for evaluating predictability in daily lead-time steps. This problem can be circumvented by using multiple annual iterations of the reforecast data. We are aware that obviously this contradicts our initial argument for using a reforecast ensemble—which is homogeneity in terms of the model itself. This choice is nonetheless made legitimate by the fact that only small changes occurred in the model development cycle over the years 2015–2021. The only major update is the change from ERA-Interim to ERA-5 in the initialisation of the reforecasts. We therefore trade sample size and having a daily initialisation date for some model homogeneity and merge the annual iterations from the years 2016, 2017, 2019, 2020, and 2021. From the 2020 and 2021 iterations, we both include a twice-a-week reforecast and then consequently fill the gap by using missing initialisation dates from the previous years. Doing so, we finally obtain a common dataset containing daily initialisation dates for the common time period of 2001–2015. To extend the sample of HWs, this period was then extended further to 2018. In those last three years, there are therefore obviously some initialisation dates missing, which has, however, no substantial impact on the validity and interpretability of results.

2.1.2 | GEFSv12 reforecasts

For a more robust evaluation of HW onset predictability, we consider, in the form of the NCEP GEFSv12, a second reforecast dataset with high spatial resolution and reasonable medium-range forecast skill (Guan *et al.*, 2022). In contrast to the reforecasts within the S2S project, the GEFSv12 reforecasts are targeted at medium-range or early extended-range forecasts with a maximum lead time of 16 days. Compared with ECMWF, this reforecast dataset comes with the advantage of daily initialisation dates but with the downside of only featuring five ensemble members. The model has a spatial resolution of 25 km with

64 vertical levels and output is provided on a 0.5° grid in six-hourly time steps.

2.2 | Weather regimes

As mentioned in the Introduction, the concept of Euro-Atlantic weather regimes allows a practical and meaningful stratification of HW onsets by the prevalent large-scale synoptic setup. The seven Euro-Atlantic weather regimes used in this study were first introduced in Grams *et al.* (2017), which also features a more detailed section about the technical realisation. In order to assign any instantaneous flow field to a certain weather regime, first climatological mean weather regime patterns are identified on the basis of ERA5 1979–2021 data. To do so, six-hourly 500-hPa geopotential height anomalies are computed with respect to a 91-day moving window climatology. Thereafter, the anomalies are filtered with a 10-day low-pass filter and normalised seasonally. EOF analysis restricted to the domain 80°W–40°E, 30°–90°N and subsequent *k*-means clustering then finally yield an anomaly pattern of an optimal number of seven weather regimes, which together represent 70% of the spatio-temporal variability in the 500-hPa geopotential field. Now, each instantaneous geopotential height anomaly field can be projected onto each of these seven patterns, which yields a so-called instantaneous weather regime index (IWR), a non-dimensional quantity that can be thought of as very similar to a principal component of an EOF analysis (see Michel & Rivière, 2011 for more details). If one particular regime's IWR is larger than those of any other regime and if the IWR value is furthermore above 1.0 for at least five consecutive days, the respective day's large-scale weather pattern is then assigned to that particular weather regime. Thereby, the $Z500$ field of any given day can be allocated to either one of the seven Euro-Atlantic weather regimes or none of the seven regimes at all, which is true for about 30% of all days in summer ("no-regime case"). Throughout this article, we will usually use abbreviations for the weather regimes, which are listed in Table 1 along with abbreviations used for the four European regions of interest. A detailed overview of the average synoptic patterns associated with Euro-Atlantic weather regimes as well as the year-round climatology of their relative frequencies can be found in Büeler *et al.* (2021).

TABLE 1 Abbreviations used in this study for the four European regions of interest and the seven Euro-Atlantic weather regimes.

European region	CE	Central Europe
	BI	British Isles
	Sc	Scandinavia
	WR	Western Russia
Euro-Atlantic weather regime	AtlTr	Atlantic Trough
	zonal	zonal regime
	ScTr	Scandinavian Trough
	AtlRi	Atlantic Ridge
	EuBL	European blocking
	ScBL	Scandinavian blocking
	GrBL	Greenland blocking
	NoReg	no regime

2.3 | Heatwave detection

With the help of an objective algorithm, HWs of a duration of three or more days are detected for all four European regions for the months May–September over the period 2001–2018. We use hourly ERA-5 data (Hersbach *et al.*, 2020) at 1° spatial resolution and determine the maximum 2-m temperature value for each day. After detrending the data series at each grid point individually, the local 90th percentile of the climatological reference period 2000–2019 is calculated for each grid point individually. For all grid points where the 90th percentile is surpassed, the standardised anomalies are then summed up over the domain of interest, for example, the Central European domain depicted with black shading in Figure 1. In the case in which this regional sum also exceeds the 90th percentile for at least three consecutive days, that particular event is considered as a HW in the framework of this article. In contrast to some earlier studies, the HW detection algorithm used here may be a bit less stringent, due to the choice of the 90th percentile as one of the key criteria. For the scope of this study, we are willing to shift more towards a higher sample size, with the trade-off of including some less intense HWs. Table 2 provides an overview of the characteristics of the identified heatwaves, such as the total number for each region (and per month), as well as statistics with respect to length and intensity. Intensity is represented by the maximum temperature Extreme Forecast Index, which will be introduced later in Section 2.6.

2.4 | Selection of non-HW weather episodes

Throughout this article, the predictability of weather patterns or temperature fields during HW onsets will occasionally be compared with the predictability of

TABLE 2 Overview of the HWs detected in the four European regions of interest. 25th and 75th denote the respective 25th and 75th percentiles of the HW length or the analysed Tmax-EFI value, whereas min/max denote the minimum or maximum values (minimum omitted for HW length, as it is always the selected minimum threshold of 3 days).

	Total number of HWs and per month						HW length statistics				Tmax-EFI statistics				
Region	total	May	Jun	Jul	Aug	Sep	25th	med	75th	max	min	25th	med	75th	max
CE	49	13	10	8	11	7	4	5	6	13	0.38	0.52	0.58	0.66	0.74
BI	47	13	7	11	5	11	4	4	6	17	0.31	0.41	0.51	0.60	0.73
Sc	48	13	6	14	7	8	4	5	7	15	0.25	0.46	0.55	0.60	0.73
WR	40	10	8	10	8	4	4	5	7	35	0.28	0.48	0.53	0.61	0.70

Note: The Tmax-EFI values presented here are spatial means calculated over the respective regions (see shaded areas in Figure 1) based on gridpoint-wise calculated values based on the ECMWF 11-member ensemble reforecasts for maximum 2-m temperature at a lead time of 24 h (quasi-analysis). Furthermore, the values were temporally averaged over the first 3 days of the respective HW.

weather outside HWs. To do so, we compile a list of all May–September dates in the period 2001–2018 and remove (region-specific) all HW days and the 7 days before onset and the 7 days after decay. Moreover, we do not want to evaluate forecasts for a sequence of consecutive days, since they may feature very similar weather. For this reason, and to allow better comparison with HW onset events, which are rather rare events temporally separated by at least multiple days, we further demand that all "onset days of non-HW weather episodes" are at least 5 days apart. Doing so, the ratio of non-HW to HW weather episodes also reflects the actual ratio of individual HW to non-HW days to a good approximation.

2.5 | Variables of interest

For the evaluation of predictability, we will mostly refer to typical variables that are commonly used for forecast verification, such as the 500-hPa geopotential or the temperature at 850 hPa. The exact way in which these variables will be used to quantify forecast quality will be outlined in the next and final part of this section. Beyond these commonly used quantities, we will further use a variety of basic or more advanced quantities that are useful in the context of HW predictability, such as Eady growth rate, soil moisture, or upper-level jet speed. All these additional variables are listed in Table 3, along with their acronyms used in this article. This table also serves as an overview of the datasets and forecast evaluation metrics used.

On many occasions throughout this article, fields of these variables are not presented in absolute terms but instead as anomalies with respect to a running-mean background climatology. For this, we define as climatology a running 21-day average centred around the day of interest, computed over all 18 years of available reforecast data.

2.6 | Evaluation of predictive skill

For the scope of this article, we decided not to evaluate HW onset predictability in the form of temperature-threshold-based binary skill measures, which are also often used in their probabilistic form (e.g., Brier skill score). Instead, we employ multiple rather classic evaluation metrics used in forecast verification, such as the 500-hPa geopotential anomaly correlation coefficient. This way of evaluation allows us to compare the quality of forecasts at times of HW onset directly with those of summertime weather episodes outside HWs, an advantage which we deem to be of interest. Moreover, the comparison of forecast performance with regard to different metrics may allow a better understanding of the underlying reason for a bad forecast.

The first three forecast skill metrics employed are deterministic ones, meaning they are calculated individually for each ensemble member of the reforecast. For reasons of physical consistency and to exclude general model biases, the forecast evaluation is always done with respect to the (quasi-)analysis of the given reforecast model, not against ERA5. For instance, the respective forecast errors are computed by simply subtracting from the predicted absolute fields either the model's analysed state for the forecast's valid day (lead time 0 h; for most fields) or the predicted values from a short 24-h integration (for Tmax). To evaluate the predictive skill with respect to the large-scale circulation, we use the following two metrics, which are computed over the enlarged regional domain denoted by the contours in Figure 1, respectively (see Table S1 for coordinates of these boxes):

1. 500-hPa geopotential anomaly correlation coefficient (**Z500-ACC**);
2. 500-hPa geopotential root-mean-square error normalised by the field mean (**Z500-nRMSE**).

TABLE 3 Overview of the data used in this study and the variables of interest, as well as the evaluation metrics used.

	Primary Forecast dataset	**Secondary Forecast dataset**	**Primary evaluation dataset**	**Auxiliary reanalysis data**
Data	ECMWF reforecasts multiple iterations (2016,2017 2019,2020,2021; see Section 2.1.1)	GEFSv12 reforecasts	respective forecast datasets' analysis state of the control run or short integration 0–24 h	ERA5 (just for HW detection & weather regime assignment)
Time period	MJJAS 2001–2018	MJJAS 2001–2018	MJJAS 2001–2018	MJJAS 2001–2018
Init. dates	daily due to merging procedure (with some dates missing 2016–2018)	daily	–	–
Model resolution	T639 (≈ 16 km)	C384 (≈ 25 km)	–	31 km
Data resolution	1.5°	0.5°	–	1°
Ensemble size	11 (10 dist. + ctrl)	5 (4 dist. + ctrl)	1 (control run)	–
	Acronym	**Variable**	**Vertical level**	**Time/integration time**
Variables of interest	Tmax	maximum temperature	2 m	max(0000–2400 UTC)
	*T*850	temperature	850 hPa	0000 UTC
	*Z*500	geopotential	500 hPa	0000 UTC
	SMO20	soil moisture	top 20 cm	0000 UTC
	u250	zonal wind	250 hPa	0000 UTC
	IWP	integrated water vapour	int. over trop.	0000 UTC
	Eady-GR	Eady growth rate	925 h–500 hPa	0000 UTC
	Acronym	**Description (*S* = spatial domain, *E* = ensemble space)**		
Evaluation metrics	Z500-ACC	500-hPa geopotential anomaly correlation coefficient (*S*)		
	Z500-nRMSE	root-mean-square error in *Z*500 (*S*)		
	*T*850-nRMSE	root-mean-square error in *T*850 (*S*)		
	Tmax EFI-RE	Tmax extreme forecast index relative error (*S*,*E*), see Eq. (1)		

Note: In the last part of the table, the descriptions of the forecast evaluation metrics used are supplemented by a letter in parentheses. *S* means that the evaluation metric is computed over the spatial domain (mostly the case here), whereas an additional *E* denotes that the metric also integrates over the ensemble space of each individual forecast. Tmax-EFI is, strictly speaking, a grid-point-based measure, but we compute a spatial mean afterwards.

Errors in the prediction of temperature fields are quantified by the following metric, which is—in contrast to the *Z*500-based metrics—computed over the smaller regional domains depicted by the shaded boxes in Figure 1:

3. 850-hPa temperature root-mean-square error normalised by the field mean (*T*850-**nRMSE**).

The above metrics correspond to the standard typically used in forecast evaluation. A proper computation of the anomaly correlation coefficient (ACC) requires the subtraction of a climatology. For this, we use the 21-day running climatology as mentioned in Section 2.5, which is also calculated lead-time specific for the reforecasts. The computations are mostly done with the help

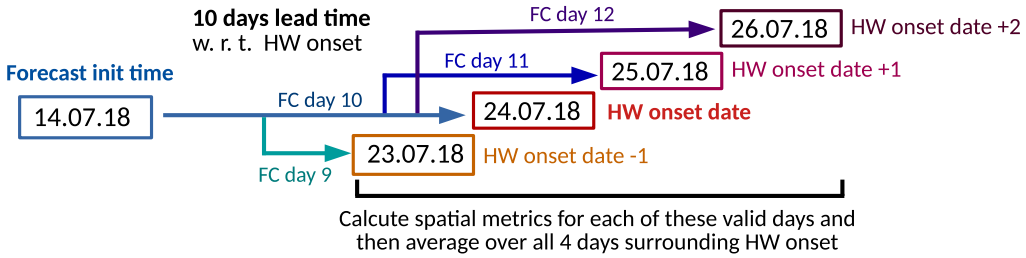

FIGURE 2 Depiction of the evaluation scheme used to assess predictive skill as a function of a given lead time for the metrics $Z500$-ACC, $Z500$-nRMSE, and $T850$-nRMSE. This evaluation scheme is applied individually to each ensemble member of the reforecast dataset and for all daily lead time steps between 3 and 12 days. As an example, the evaluation of 10 heatwave onset cases will therefore yield 110 individual $Z500$-ACC scores per lead time in the case of the ECMWF reforecast ensemble. [Colour figure can be viewed at wileyonlinelibrary.com]

of the command-line tool Climate Data Operators (CDO: Schulzweida, 2022), which incorporates proper weighing of grid cells by their area.

To assess robustly whether the model predicts the temporal evolution of the synoptic-scale flow during HW onset skilfully—for example, the development of a stationary ridge over multiple days—the above-mentioned deterministic skill scores are not only applied to a single temporal snapshot during HW onset. Instead we compute an average score over four days around HW onset as follows. For a 10-day lead time forecast of a HW onset, we not only calculate the spatial metrics $Z500$-ACC and $Z500/T850$-nRMSE for forecast day 10, but also compare forecast day 9 with the analysis of one day prior to onset. In the same way, we also compute the metrics for forecast days 11 and 12 (see also Figure 2). The metrics for all four days around HW onset are then averaged. For the ACC, we perform a Fisher-z transformation before calculating the average. We are aware that this evaluation procedure renders the reference to lead time for any given forecast a bit vague. Throughout this article, we will still consider any forecast initialised 10 days before HW onset a *10-day lead-time forecast*, although fields up to forecast day 12 are included in the evaluation.

Finally, we employ a fourth skill metric to assess how well the extremeness of temperature is represented across the entire forecast ensemble:

4. relative error in the Extreme Forecast Index for maximum 2-m temperature (TMAX-EFI-RE).

In general, the exteme forecast index (EFI) is a unitless ensemble-integrated measure, which describes how strongly the predicted cumulative density function of a variable deviates from a lead-time-dependent model climatology, the so-called M-climate (Zsótér, 2006). Per definition, the EFI ranges between −1 and 1 and values between 0.5 and 0.8 are generally regarded as unusual, whereas values above 0.8 designate an extreme event.

For this article, Tmax-EFI is calculated very similarly to the operational product at the ECMWF, with the only difference being a slightly altered climatological period as well as the choice of a 21-day running climatology window. The computation is only done for ECMWF and not for GEFSv12 because of its very low number of ensemble members. Although somewhat counterintuitive due to the expected narrow distribution, Tmax-EFI may also be calculated for a quasi-analysis, that is, for the 0–18 h maximum-temperature forecasts of the ensemble. Averaged spatially over the respective domain of interest, this analysed EFI will not only be used to compute forecast errors. In addition, it will also be used as a practical tool to assess and compare the intensities of HWs. On average, the HWs considered in this article feature a domain-averaged value of around 0.55 (see Table 2).

For our forecast evaluation, we will use the Tmax-EFI metric in a less stringent way than the more common metrics discussed above. The intention here is to evaluate whether the model was at least somewhat successful in suggesting unusually high temperatures, even though the duration of the hot spell and/or the timing may have been off by one or two days. For the first three days of the HW, we therefore compute the spatial mean Tmax-EFI for each day individually and then pick the maximum predicted value. From this we then subtract the analysed Tmax-EFI averaged over the first three days of the HW. Finally, the Tmax-EFI error is then also normalised by dividing it by the observed three-day mean value:

$$Tmax - \mathrm{EFI}_{predmax} - \mathrm{RE} = \frac{3dmax(EFI_{\mathrm{pred}}) - 3davg(EFI_{\mathrm{ana}})}{3davg(EFI_{\mathrm{ana}})}. \quad (1)$$

For better readability, we will on most occasions refer to this metric simply as the relative error in Tmax-EFI. Similar formulations of this EFI-based metric were tested, which do not include the normalisation by the analysed value or which also use a three-day average instead of the

maximum for the predicted EFI values. It was found that our main results are insensitive to this subjective choice of an evaluation metric.

3 | RESULTS: PREDICTABILITY OF HWs 2001–2018 IN RELATION TO WEATHER REGIMES

3.1 | Overview: regimes during HW onset

Before we review the medium-range predictability of HW onsets by means of multiple evaluation metrics, we first provide an overview over the Euro-Atlantic weather regimes that are typically associated with HW onsets. The weather regime data are based on ERA5 reanalysis and a heatwave onset only gets assigned to a certain regime when, on two of the three first HW days, the same weather regime was registered (i.e., the respective weather regime index surpasses the criteria specified in Section 2.2). Figure 3 compares, for each European region considered, the relative fraction of each weather regime during the HW onset (upper pie chart) against the mix of weather regimes found in all other summertime weather episodes outside HWs (bottom pie charts). Additional information about the number of HW onsets per weather regime and basic statistics concerning length and intensity are provided in Table 4.

Not surprisingly, HW onset is mainly associated with the development of classic continental blocking regimes, namely Scandinavian (ScBL) and European blocking (EuBL). This is especially true for the British Isles and Scandinavia regions, where these blocking regimes occur during HW onset in 66% and 75% of all cases, respectively. Interestingly, Central European HWs do often, in about 30% of cases, occur in a dynamical setup that cannot be robustly attributed to one of the given weather regimes (NoReg). Western Russian being located at the eastern boundary of the European domain renders the assignment of Euro-Atlantic weather regimes less meaningful compared with other regions. The NoReg case dominates during HW onset, but the relative fractions are generally close to summer climatology.

For a better understanding, Figure 4 provides—exemplarily only for Central European HWs—composite means of anomalous analysed 500-hPa geopotential fields during HW onset, stratified by the associated weather regime. Here stippling denotes a significant difference from the sample of all 49 Central European HWs. As evident from the figure, the "no regime" composite still features substantial positive anomalies, that is, a ridging over Central Europe. This is not unexpected, as the Euro-Atlantic weather regime classification adopted represents rather stable large-scale flow anomalies with a characteristic time-scale that exceeds those of most synoptic-scale phenomena (e.g., a passing cyclone). A short-lived ridge enabling a heatwave over Central Europe may therefore not surpass the threshold needed to be assigned to one of the canonical blocking regimes. HW onsets related to ScBL are distinguished from those associated with EuBL not only in the more northward extent of the block but by the existence of a pronounced trough upstream over the Atlantic.

For the three other European regions, the 500-hPa geopotential composite-mean anomaly fields during HW onset are presented in the Supporting Information (Figures S1–S3). For further comparison, Figure S4 also provides composites of the three weather regimes of interest for non-HW cases.

How well are European HW onsets predicted and is there increased or reduced predictability compared with regular summer weather? Are there differences among the different regions, and finally, are forecasts better or worse depending on which weather regime the HW onset is associated with? In the upcoming sections, we tackle these questions, focusing on the reforecasts from ECMWF, also in terms of the figures presented. Differences between ECMWF and the secondary reforecast dataset GEFSv12 will be discussed briefly and additional material will be provided in the Supporting Information. For a first quick overview with regard to the aforementioned questions about HW predictability, we again point to Figure 3. The numbers written into the pie chart denote the average lead time until the $Z500$-ACC first drops consistently below the value of 0.6 (a value usually chosen to characterise useful forecasts; e.g., Simmons, 1986). For the rest of this article, we refer to it as a large-scale forecast skill horizon.

3.2 | Predictability of central European HW onsets

In ECMWF reforecasts, Central European HW onsets associated with either Scandinavian or European blocking exhibit an additional day of adequate large-scale forecast skill compared with those HWs evolving in conditions without a distinct weather regime. Moreover, HW onsets under such classic blocking regimes also show better scores compared with non-HW weather episodes associated with the same blocking regimes, which is not the case in times of "no regime".

A more detailed presentation of ECMWF reforecast skill scores including $Z500$-ACC and other metrics as a function of lead time is presented in Figure 5. Here, it

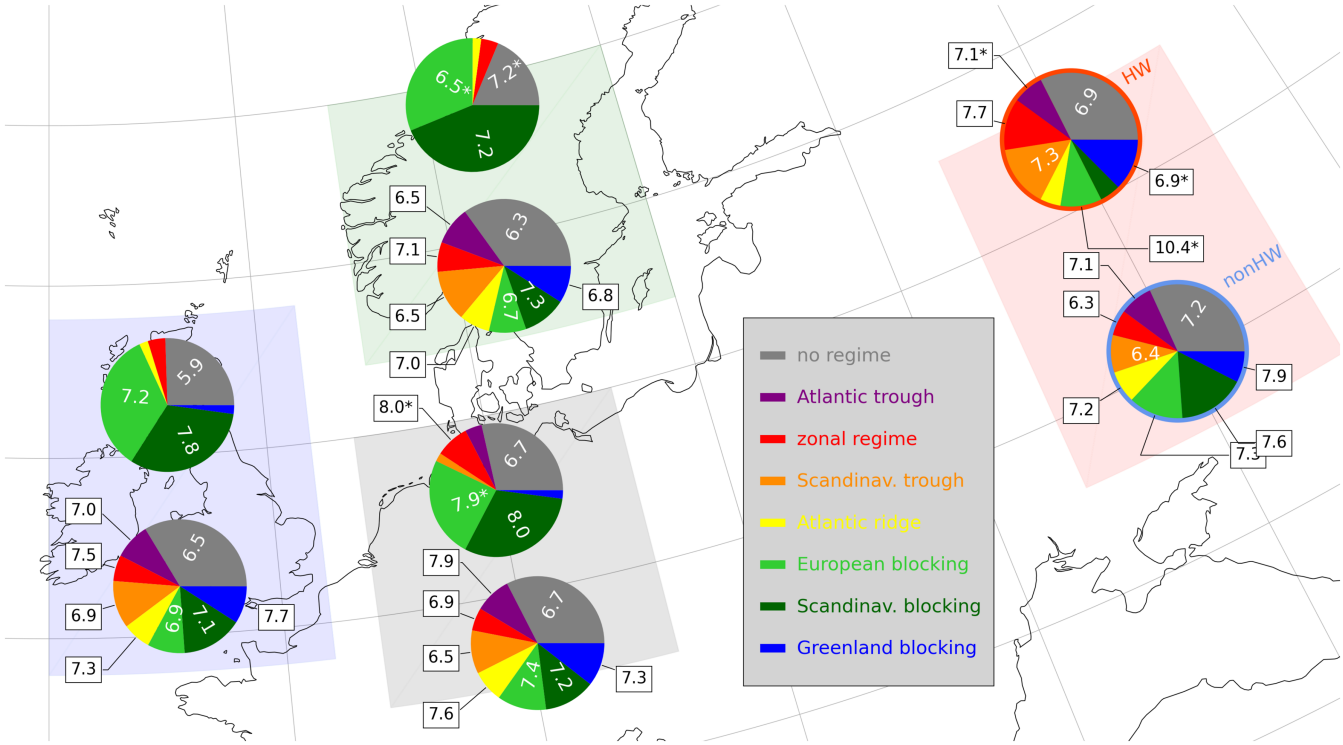

FIGURE 3 Comparison of regime fraction and average large-scale forecast skill horizon between HW and non-HW days. The pie charts depict, for each region of interest, the relative fraction of each of the seven Euro-Atlantic regimes during HW onset (upper pie charts) and for all non-HW periods in the investigated time period of 2001–2018. The numbers denote the average forecast skill horizon in terms of the adequate representation of the large-scale synoptic weather patterns (average lead time until 500-hPa geopotential ACC drops below 0.6). The ACC scores were calculated individually for each member of the ECMWF reforecast ensemble using the method described in Section 2.6. To highlight the more frequent weather regimes during HW onset, the forecast skill horizon is written directly onto the pie charts when the relative fraction exceeds 12.5% (also for non-HW days for better comparability). A star behind the number denotes that the 2.5%–97.5% confidence range w.r.t. the forecast skill horizon spans more than half a day. If fewer than four HWs of a given regime exist, the forecast skill horizon is not provided, due to too low sample size. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Number and characteristics of HW onsets associated with one of the three dominant weather regimes for the four European regions considered.

Region	Regime	Number	Length	Tmax-EFI
CE	ScBL	15	5	0.57
	EuBL	12	4.5	0.55
	NoReg	14	5	0.57
BI	ScBL	15	4	0.58
	EuBL	16	6.5	0.47
	NoReg	12	4	0.57
Sc	ScBL	21	6	0.54
	EuBL	15	4	0.55
	NoReg	9	5	0.6
WR	ScTr	6	6.5	0.44
	GrBL	5	5	0.54
	NoReg	13	5	0.53

Note: For each region and dominant regime, the median length as well as the median analysed Tmax-EFI are presented.

becomes evident that $Z500$-ACC drops significantly below non-HW levels already after day 6 for NoReg HW cases (Figure 5c), whereas, in association with classic blocking regimes, predictability of the synoptic-scale patterns remains higher for most lead times (Figure 5a,b). Up until forecast day 7, this skill advantage reaches statistical significance at the 5% level when compared with non-HW cases of the same regime (line segments printed thicker). Moreover, around 7 days lead time, HW onsets associated with ScBL or EuBL are also significantly better predicted than HW onsets of all regimes (hexagons filled with a "+" symbol). One has to be careful, though, when interpreting $Z500$-ACC, as this score may benefit from large, spatially coherent anomalies relative to the climatology (Andersson *et al.*, 2015). This means that, in times of a European blocking, $Z500$-ACC may be an exceptionally forgiving metric, whereas the no-regime case will naturally produce worse scores due to a higher likelihood of phase errors. Despite the call for caution, the overall assessment of predictive skill does not change substantially when additional metrics are considered. The

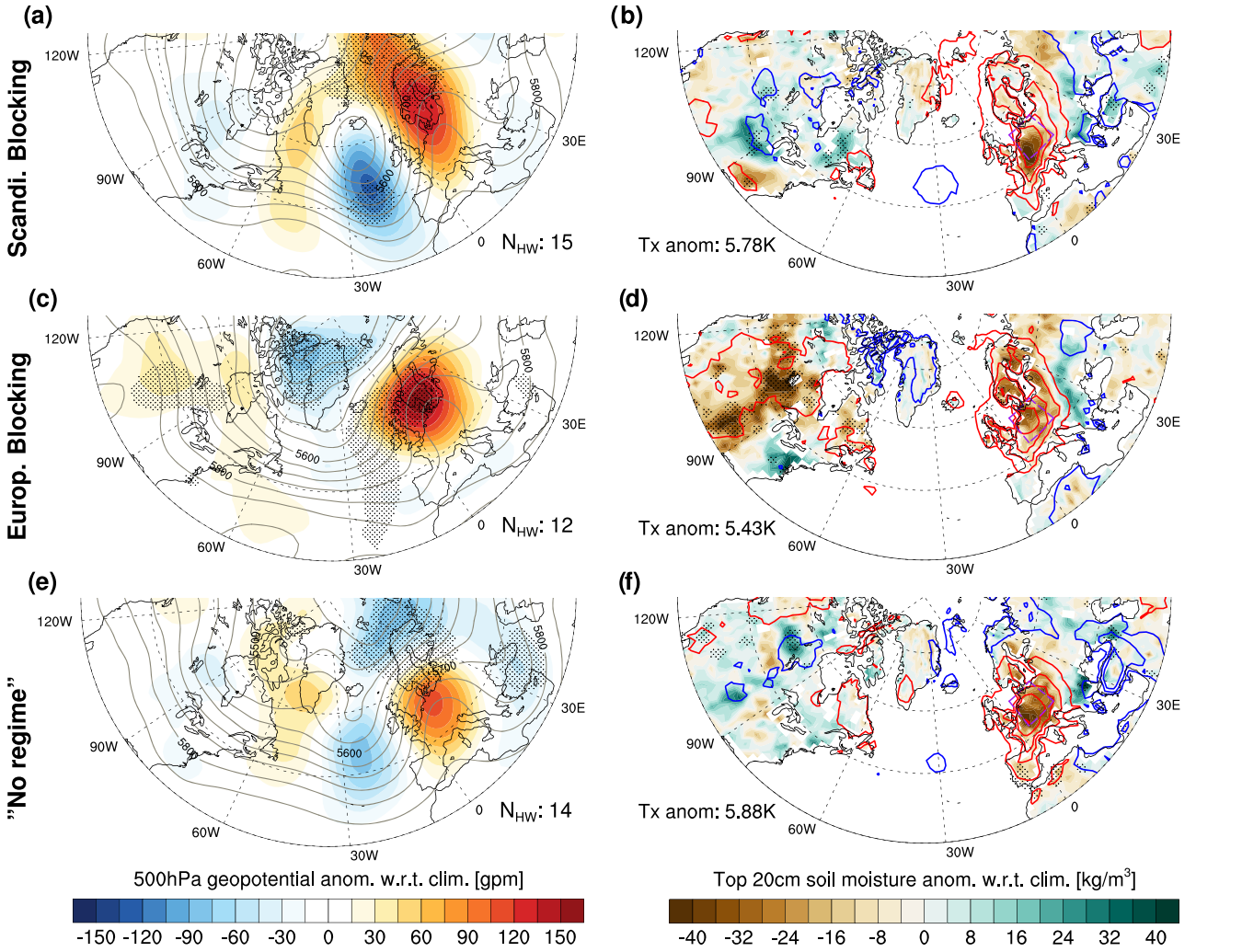

FIGURE 4 Overview of the 500-hPa geopotential, soil moisture, and Tmax anomalies associated with Central European HW onsets, stratified by three dominant weather regimes during HW onsets. The shadings depict the composite-mean 500-hPa geopotential anomaly (left panel)/top 20-cm soil-moisture anomaly (right panel), always with respect to the respective HW day's running three-week summer climatology, for (a,b) Scandinavian blocking type, (c,d) European blocking type, and (e,f) "no regime" type HWs. Hatching highlights regions in which the respective anomaly differs significantly from the sample of all 49 Central European HWs. In the left panels, the absolute composite-mean geopotential fields are shown in contours, whereas, in the right panels, coloured contour lines denote anomalies of maximum 2-m temperature with respect to the running summer climatology with a contour-line interval of 1.5 K. The labels in the centre denote the number of HW onsets registered in association with the respective weather regime, as well as the spatially averaged maximum temperature deviation in the Central Europe domain (only land points considered). [Colour figure can be viewed at wileyonlinelibrary.com]

normalised root-mean-square error (nRMSE) for 500-hPa geopotential draws a very similar picture (Figure 5d–f). This metric, which now also accounts for amplitude errors and domain-averaged biases, supports the hypothesis that Central European HW onsets associated with one of the two blocking regimes are predicted better in terms of the large-scale flow field. This is particularly true for lead times of 6 days or longer, for which the differences also reach statistical significance at the 5% level. We want to note that the elevated skill metrics do not necessarily imply that the onset of the heatwave-enabling blocking itself is also predicted well. In some cases, blocking may have been established already during the time of forecast initialisation (we will later see this indeed being important, particularly for well-predicted British heatwaves). Given such a scenario, forecast skill in terms of $Z500$ is expected to be increased due to models generally being more skilful in capturing the persistence rather than the onset of a blocking (Ferranti *et al.*, 2015; Matsueda & Palmer, 2018; Pelly & Hoskins, 2003).

850-hPa temperature fields are generally harder to predict in times of HW onsets. Particularly for longer lead times, the large deviation from the normal temperature range usually leads to considerable biases, resulting in

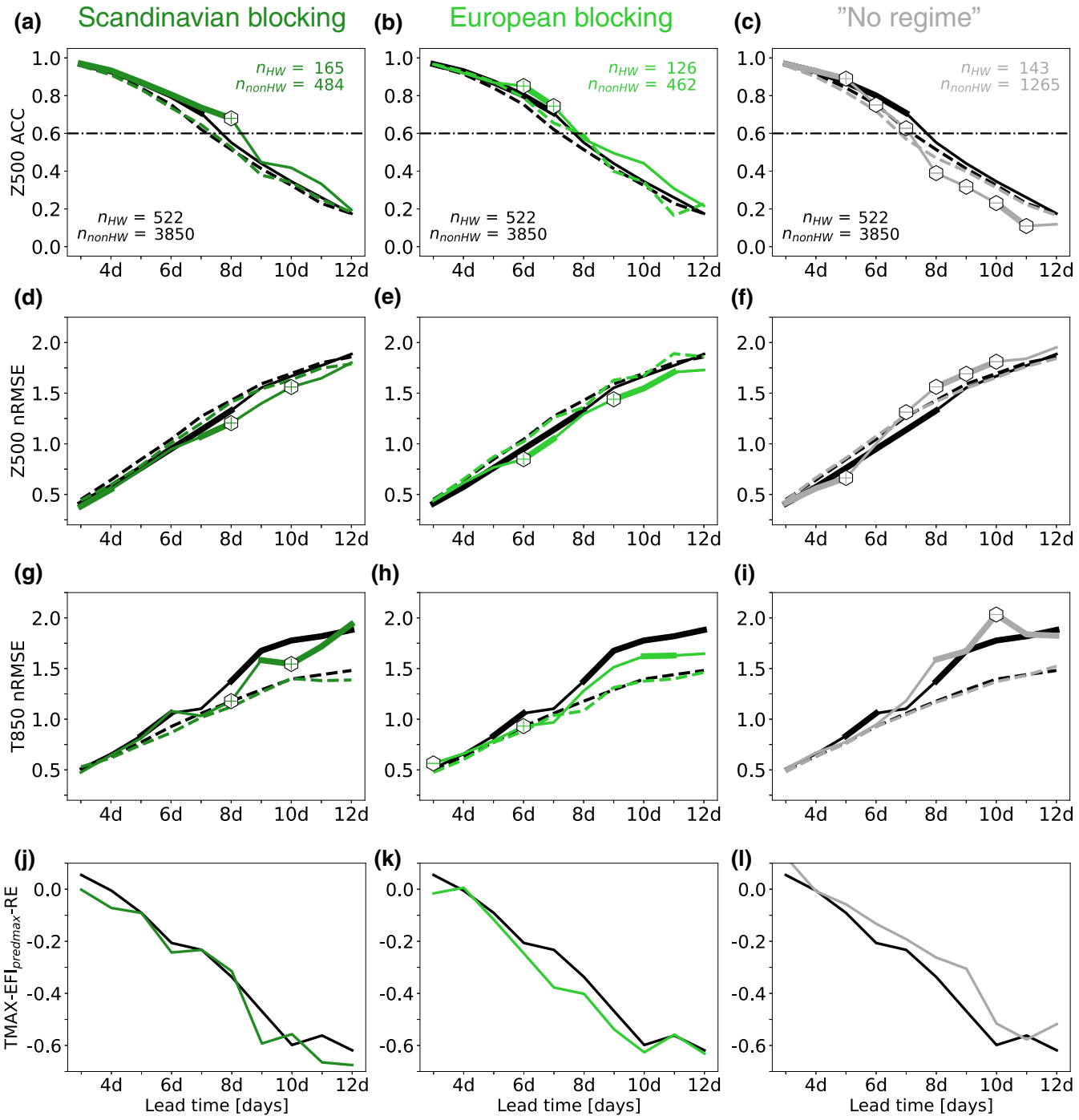

FIGURE 5 Comparison of ECMWF reforecast's predictive skill for Central European HW onsets versus non-HW weather episodes, stratified by ERA5-based analysed weather regimes during HW onset. Each plot depicts a skill metric (from top to bottom: *Z*500-ACC, *Z*500-nRMSE, *T*850-nRMSE, Tmax-EFI-RE) as a function of lead time. The median of the HW onset subsample is depicted with solid lines, whereas the median for non-HW episodes is shown with dashed lines (omitted for Tmax-EFI). From left to right, the columns depict skill metrics for ScBL-type HW/non-HW onsets (dark green), EuBL-type HW/non-HW onsets (light green), and NoReg-type HW/non-HW onsets (grey). For better comparison, each column also features in black the skill metric for the HW/non-HW sample containing all weather regimes. Line segments are printed thicker for lead times at which HW onsets of a specific regime show either significantly lower or higher predictive skill compared with the non-HW sample for the same regime. To denote significantly higher (lower) predictive skill for HW onsets of a given regime compared with all HW onsets, hexagons filled with a "+" ("−") symbol are used. Significance is tested at the 5% level via bootstrapping (30,000 iterations). The printed numbers are the respective sample sizes (number of individual ensemble member forecasts). [Colour figure can be viewed at wileyonlinelibrary.com]

nRMSE scores that are expected to be worse. Nonetheless, EuBL and ScBL-type HWs exhibit predictability comparable with regular summer weather episodes up until forecast 7 and only slightly worse performance thereafter (Figure 5g,h). In contrast, NoReg-type HWs suffer from considerably worse scores from forecast day 8 onwards (Figure 5i). Since the domain-wide 850-hPa temperature biases are largely similar across HWs of all regimes (Figure S5), the poorer nRMSE score is therefore mostly related to phase errors.

Finally, we evaluate how the well the ensemble forecasts predict the extremeness of near-surface temperatures by means of the Tmax-EFI-RE score (Figure 5j–l). In contrast to a root-mean-square error, this metric does not punish phase errors as long as the anomalous heat occurs somewhere in the domain, and further allows for some temporal mismatch in the exact occurrence of extreme temperatures. Interestingly, the previously mentioned differences in forecast quality between regimes are not reflected at all in this metric. For instance, the underestimation of temperature extremeness is slightly more pronounced for the EuBL-type HW onsets at forecast day 7, although the nRMSE for 850-hPa temperatures is the lowest of all regimes at that lead time. Similarly, at forecast day 9, both ScBL and EuBL-type HWs exhibit the worst Tmax-EFI scores; the differences from NoReg are not statically significant, however. We suppose that the disagreement with results from metrics discussed earlier may be an indicator of the intricacies of near-surface temperature prediction in situations with little synoptic forcing. Under a blocking high, the correct prediction of Tmax might be influenced more and more by more local boundary-layer processes (Gómez *et al.*, 2019; Imran *et al.*, 2018; Lemburg & Fink, 2022).

Are the presented findings about HW onset predictability in Central Europe robust or rather model-dependent? Using the same metrics, we find no major disagreement between the two models. In GEFSv12, $Z500$-ACC is again the metric for which HW onset predictions exhibit elevated skill up to lead times of 9 days compared with non-HW weather episodes of the same regime (Figure S9). When comparing HW onsets of different weather regime types, the most robust finding is again the overall lower $Z500$ predictive skill for "no regime" cases. However, compared with the ECMWF reforecasts, the $Z500$ skill advantage for EuBL and ScBL does not come out as clearly in the GEFSv12 reforecasts, likely also due to the lower sample size. Moreover, in contrast to ECMWF, there are no longer substantial differences in predictive skill between HW onsets and non-HW weather episodes given the presence of the same regime. For instance, no-regime HW onsets have shown a substantially lower $Z500$-ACC score beyond 7 days lead time in ECMWF reforecasts, whereas the same metric in GEFSv12 generally attests low skill to "no regime" cases regardless of whether the target day considered belongs to a HW onset or not. With regard to the quality of temperature forecasts, the $T850$-nRMSE scores show significantly better predictability for EuBL and ScBL-type HWs up to forecast day 7. In contrast to ECMWF, NoReg HW onsets already show significantly lower $T850$ predictability on forecast day 6. For longer lead times beyond 8 days, however, the quality of $T850$ forecasts is no longer correlated with the weather regime during HW onset. As Tmax-EFI scores were not calculated for GEFSv12 because of the low ensemble size, the median maximum temperature bias is used instead. For this metric, the overall picture is close to ECMWF, with EuBL HWs featuring a comparably high Tmax bias for most lead times despite the rather good large-scale predictability.

3.3 | Predictability of British Isles HW onsets

Overall, HW onsets over the British Isles show somewhat comparable predictability characteristics compared with Central European HWs in ECMWF reforecasts. In terms of the adequate prediction of large-scale Rossby-wave patterns, the differences are even more pronounced. ScBL-type HWs show a particularly long large-scale predictability horizon of nearly 8 days on average, which is followed by HWs forming under the EuBL regime (7.2 days). NoReg-type HW onsets clearly come in last, with an average $Z500$-ACC forecast skill horizon of about 6 days. For a more detailed view, we discuss again the multiple-metric view of predictive skill as a function of lead time. HW onsets of the ScBL type consistently exhibit elevated $Z500$-ACC scores up until forecast day 9, with predictive skill often being significantly higher compared with both non-HW weather episodes under the same regime and HW onsets of all regimes. EuBL-type HWs display normal predictive skill in $Z500$-ACC for lead times of up to 8 days, but then catch up and overtake ScBl-type HWs at forecast day 9 (Figure 6a,b). Very similar to Central Europe, HW onsets associated with the "no regime" case show significantly worse $Z500$-ACC for nearly all lead times (Figure 6c). Interestingly, the comparably good predictive skill in terms of $Z500$-ACC is barely reflected by the $Z500$-nRMSE metric. At short lead times of 4 and 5 days, the EuBL-type HWs exhibit significantly better scores (Figure 6e). However, from forecast day 7 onwards, $Z500$-nRMSE suggests equal or worse predictability compared with non-HW episodes, irrespective of the regime. Only at forecast 6 do the ScBL- and EuBL-type HWs clearly outperform NoReg HWs, which already see a sharp

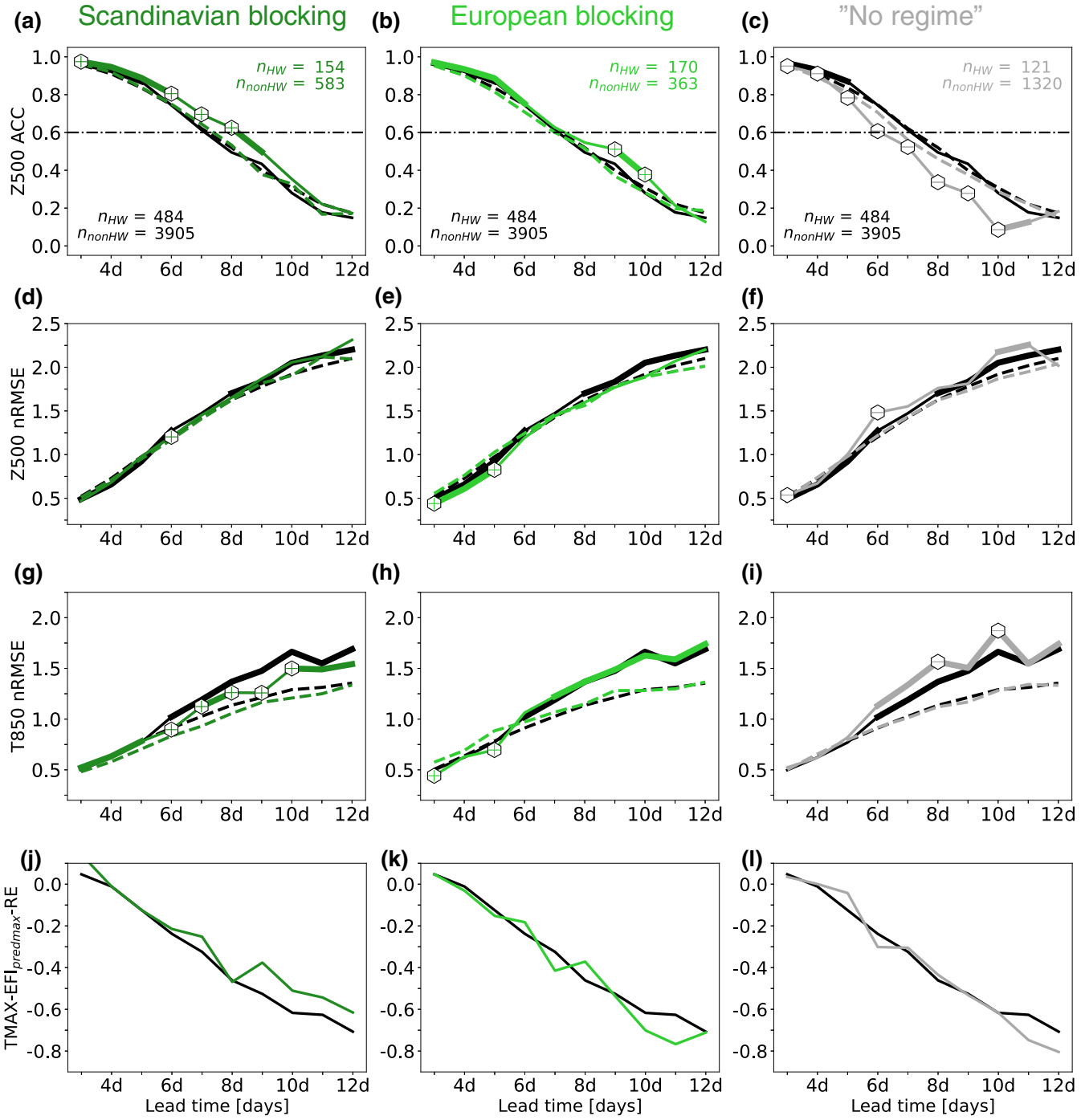

FIGURE 6 Same as Figure 5, but for the British Isles region. [Colour figure can be viewed at wileyonlinelibrary.com]

decline in *Z*500-nRMSE at this rather short lead time (Figure 6d–f).

The predictive skill for the 850-hPa temperature fields reflects quite well the regime-specific differences in large-scale predictability as evaluated by *Z*500-ACC. Up until forecast day 5, *T*850-nRMSE scores do not differ much between HW onsets and non-HW episodes, with EuBL-type HW onsets featuring slightly but statistically significantly better *T*850-nRMSE scores (Figure 6g-i). At lead times of about one week, the nRMSE generally becomes larger during HW onsets compared with non-HW weather episodes, which is, as in the Central European case, likely explained by a growing cold bias (see Figure S6). A notable exception is ScBL-type HW onsets, which—from forecast day 6 onwards—feature consistently better predictive skill in *T*850 compared with HW onsets under different regimes. This finding is partly attributable to an overall lower bias in *T*850 and may

also agree well with the advantage in predictive skill in $Z500$-ACC. Significantly poorer forecast quality in $T850$ is again found for NoReg HW onsets, very similar to the Central European case.

As was the case for Central European HWs, the Tmax-EFI-based metric does not reveal any significant differences between HWs of the three regime types (Figure 6j–l). This suggests again that, for the prediction of extreme near-surface temperatures, an adequate representation of the large-scale circulation might be necessary, but not sufficient.

In our secondary forecast dataset, the GEFSv12 reforecasts, predictability of HW onsets over the British Isles displays overall similar characteristics (Figure S10). For ScBL-associated HW onsets, the ACC score is both better than in the non-HW case and better than for HW onsets under other regimes up until forecast 9, albeit not always significantly. NoReg-related HW onsets also show significantly worse large-scale predictive skill for most lead times. In terms of the predictability of 850-hPa temperature fields, GEFSv12 mostly agrees well with ECMWF, with EuBL-type HWs exhibiting significantly better $T850$-nRMSE scores up to forecast day 5 and a slight advantage for ScBL-type HWs for longer lead times. Minor differences between both models considered are found with regard to the underestimation of near-surface maximum temperatures. Whereas ECMWF does not show consistent differences in the median Tmax bias between regimes (Figure S6), GEFSv12 features a significantly lower bias up to forecast 8 for ScBL-type HWs.

3.4 | Predictability of Scandinavian HW onsets

Moving on to Scandinavian HW onsets and their predictability in the ECMWF, we find an unexpected deviation from the aforementioned pattern. A clearly extended large-scale forecast skill horizon is no longer found for the classic blocking regimes. Instead, the NoReg-type HW onsets consistently display $Z500$-ACCs of more than 0.6 up to lead times of 7 days on average, which is either on par with or even better than for EuBL- and ScBL-type HWs. In the more detailed breakdown provided in Figure 7, NoReg actually exhibits the best $Z500$-ACC scores for many of the lead times considered, particularly beyond forecast day 8. EuBL-type HW onsets feature the worst $Z500$-ACC scores over the entire lead-time range, whereas ScBL-type HWs display significantly elevated predictive skill at least up to forecast day 7 (Figure 7a–c). However, when we consider the nRMSE of $Z500$ instead of the ACC, NoReg HWs no longer feature a comparably good score. Instead, $Z500$-nRMSE shows significantly poorer forecast performance for nearly all lead times in comparison with HWs associated with classic blocking regimes (Figure 7d–f). This hints at large-amplitude errors or systematic biases, but may also be a somewhat coincidental occurrence due to the rather low number of NoReg HW cases. Nevertheless, it complicates the interpretability of Scandinavian HW predictability.

Moreover, the $Z500$-ACC scores are not reflected in the temperature-field-based metrics (Figure 7g–i). The most prominent contrast is found for the NoReg case at a lead time of around one week: while the $Z500$-ACC scores suggest comparably good large-scale predictability, the $T850$-nRMSE indicates significantly poorer forecast quality compared with other regimes, which is attributable to a comparably much larger bias in $T850$ (Figure S7). Similar to HW onsets over Central Europe and the British Isles, it is again the ScBL regime that robustly exhibits the highest predictive skill in $T850$. The consideration of the Tmax-EFI score gives more merit to the hypothesis that NoReg HWs are not better predicted despite showing significantly higher $Z500$-ACC scores. At lead times of 6–8 days, the Tmax-EFI underestimation is highest for NoReg HWs (Figure 7j–l). The differences among regimes are again not statistically significant, however.

In the GEFSv12 reforecasts, the predictability of Scandinavian HW onsets is found to be rather similar. As in the ECMWF reforecasts, the $Z500$-ACC scores exhibit the strongest boost compared with non-HW cases for the NoReg onset type (Figure S11). This increased large-scale forecast quality is again not found to be present in the temperature-based metrics, however. Already on forecast day 5 and 6 the NoReg HW onsets show significantly worse $T850$-nRMSE and Tmax bias scores compared with both non-HW episodes and other HW onsets under a different weather regime. Very similarly to ECMWF, GEFSv12 also attests that ScBL-type HW onsets have the best predictive skill in $T850$.

3.5 | Predictability of Western Russia HW onsets

In the Western Russia region, ECMWF reforecasts predict 500-hPa geopotential fields better for HW onsets than for non-HW weather periods up until a lead time of about 9 days (Figure 8a,b). Unsurprisingly, this is again not entirely reflected in the 850-hPa temperature fields. However, up until forecast day 6, $T850$-nRMSE exhibits significantly better values (Figure 8c), a feature that sets apart Western Russia HWs from British HWs, for instance. A likely explanation is the distance from the ocean, such that minor errors in the predicted direction of flow will certainly have less impact on the temperature forecast than

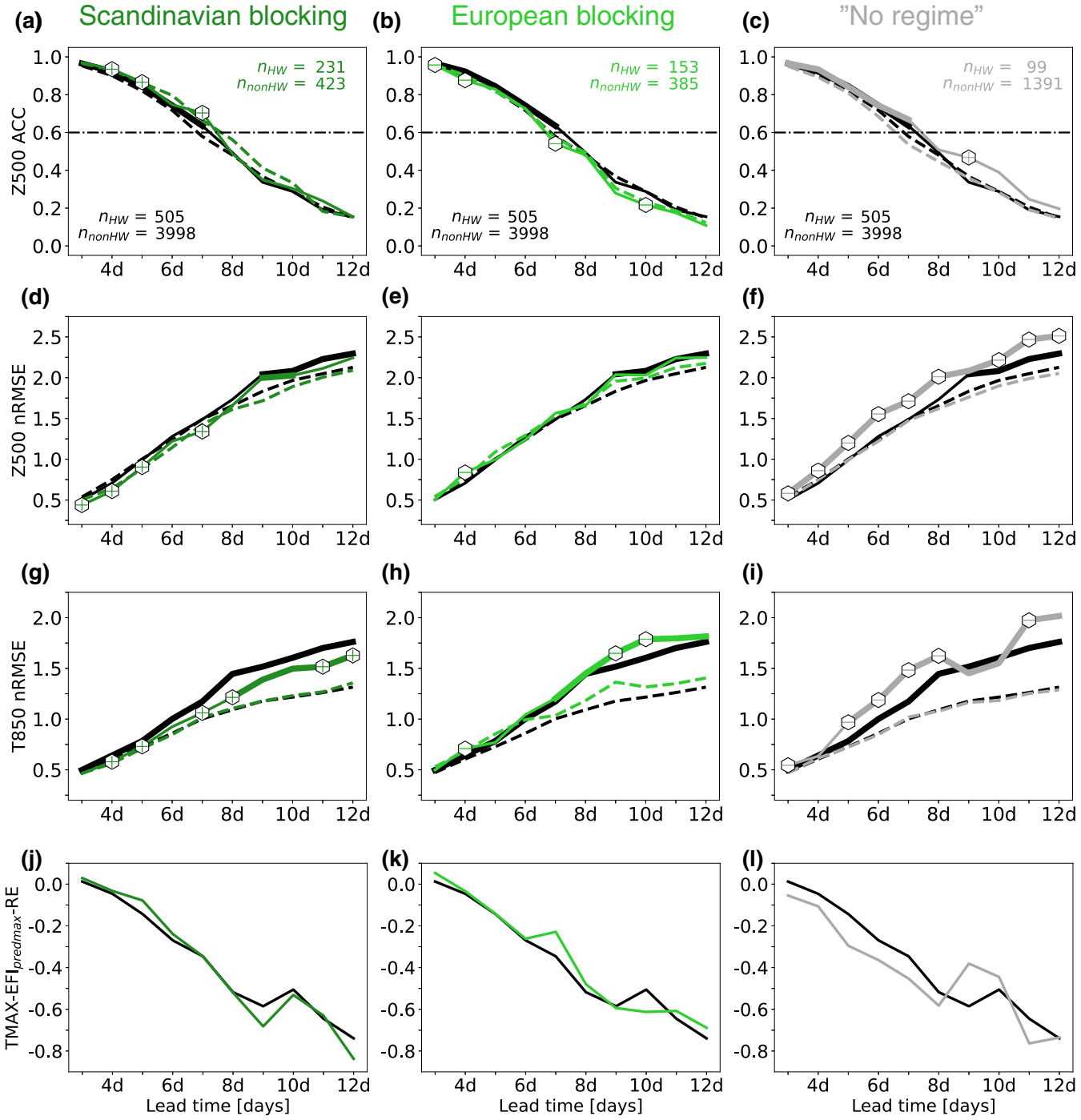

FIGURE 7 Same as Figure 5, but for Scandinavia. [Colour figure can be viewed at wileyonlinelibrary.com]

over the British Isles. As previously mentioned, a separation of forecast skill by weather regimes during HW onset is less useful for Western Russia than for the other European regions. Therefore, we only compare the "no regime" case against all HW cases, since it offers a sufficient sample size. Very similar to the Central European and British HWs, we find reduced skill in multiple metrics when the actual HW onset is associated with "no regime". The clearest decrease in skill is present in $Z500$-ACC for nearly all lead times, whereas the nRMSE of $Z500$ shows a somewhat less pronounced, but mostly also significant, distinction compared with all HWs.

Up until forecast day 8, the predictive skill for 850-hPa temperature fields is the same for NoReg HWs compared with all HWs, although a substantial decrease in $Z500$-ACC skill is already present at those lead times. A significant reduction in $T850$-nRMSE is only present for lead times of 9 and 10 days, which is likely associated with

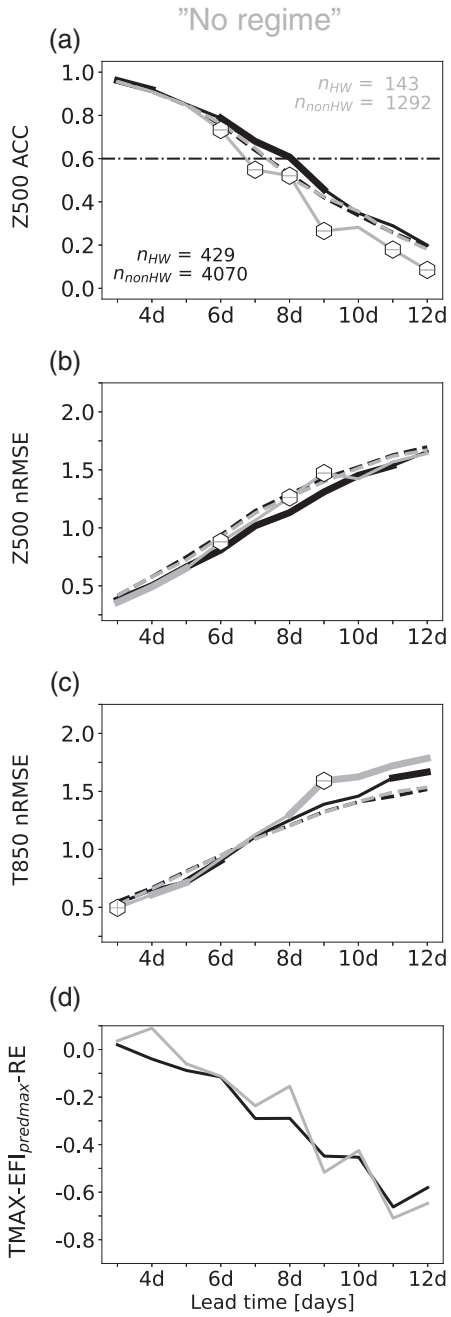

FIGURE 8 Same as Figure 5, but for the Western Russia region, and only for "no regime" cases.

a sudden increase in the negative *T*850 bias (Figure S8). However, the Tmax-EFI score shows no considerable differences between the NoReg case and all HW cases (Figure 8d).

GEFSv12 reforecasts draw a similar picture in terms of both the predictability of large-scale weather patterns and regional temperature fields (Figure S12). Compared with ECMWF, the decrease in predictive skill during "no regime" cases is less pronounced and mostly not statistically significant, but there are no major disagreements. Regarding 850-hPa and near-surface temperature forecasts, both models mostly agree well up to forecast day 8, with only small disagreement for longer lead times. GEFSv12 suggests no significant disadvantage for NoReg cases, whereas ECMWF displayed significantly worse *T*850-nRMSE scores for lead times of 9 days.

4 | RESULTS: THE ROLE OF THE INITIAL STATE

We will now investigate whether and to what extent the medium-range predictability of HW onsets may depend on the initial state during and shortly after forecast initialisation. Here, "initial state" is a broad term used for preceding weather regimes, upstream (thermo)dynamical precursors, or large-scale anomalies in soil moisture. In other words, we aim to understand possible causes for both forecast busts and windows of opportunity for enhanced HW onset predictability. The investigation is focused on HWs over Central Europe and the British Isles and lead times between 9 and 11 days. In contrast to the article's first part, we will no longer a priori stratify HW onsets by the associated Euro-Atlantic weather regime. Instead, we rather intuitively divide the sample of HW onset cases into those with the overall (ensemble-integrated) best predictive skill and those with the worst metrics. It is in our view crucial to consider more than one target metric. Heatwave onset cases may be well predicted in terms of the evolution of the synoptic-scale flow into some blocking, but the extremeness of temperatures may be underestimated considerably—or the other way around. We therefore consider two selection metrics: ensemble-median *Z*500-ACC and the relative error in Tmax-EFI, aggregated for each HW respectively over forecasts with 9, 10, and 11 days lead time (or 6, 7, and 8 days for the 6–8 day lead-time case). For both of these metrics, we then select the respective HWs with forecast performance above the upper quartile (best-predicted; 12 cases for Central Europe, 11 for British Isles) or below the lower quartile (worst-predicted).

4.1 | What affects the predictability of Central European HW onsets?

Table 5 provides an overview of the best- and worst-predicted Central European HW onsets in ECMWF reforecasts at 9–11 days lead time, with respect to both the aforementioned selection metrics. Overall, well and poorly predicted HWs share similar characteristics in terms of median length and median intensity. Notably, the intersection of both selection metric groups turns out

TABLE 5 Overview of the 12 best- and worst-predicted Central European HWs in ECMWF reforecasts at 9–11 days lead time according to the two different selection metrics *Z*500-ACC and relative error in Tmax-EFI. The length and the analysed Tmax-EFI (averaged over first three HW days) as a measure of extremeness are provided as basic information for each HW presented. Ensemble-median *Z*500-ACC and the maximum predicted Tmax-EFI are presented as an indicator of how well the individual HW onsets were predicted by ECMWF. Finally, the analysed regime during the first three days is given. The font colours denote intersections within the groups, with green colours indicating that a HW is found among the best predicted in both the *Z*500-ACC and the Tmax-EFI group. Red denotes intersection in these two selection metric groups for the worst-predicted HWs. Purple denotes cross-group intersection, where the best-predicted HW with respect to one selection metric is found in the group of worst-predicted HWs with respect to the other selection metric, or vice versa. HW onsets of the respective group which are also found in the same group for lead times of 6–8 days are marked with an asterisk.

12 best-predicted HWs w.r.t *Z*500-**ACC**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
*	2018-07-24	5	0.61	0.69	0.35	EuBL
*	2004-08-05	6	0.43	0.68	0.10	ScBL
	2006-07-04	3	0.59	0.65	0.31	EuBL
	2006-07-17	12	0.56	0.64	0.53	EuBL
	2016-09-10	6	0.73	0.63	0.65	ScBL
	2018-07-30	6	0.66	0.61	0.61	ScBL
*	2007-06-07	5	0.57	0.61	0.38	ScBL
*	2003-05-04	5	0.58	0.60	0.27	Zonal
*	2015-08-06	9	0.61	0.58	0.34	ScBL
*	2008-05-08	7	0.48	0.57	0.35	NoReg
*	2002-07-29	3	0.46	0.57	0.31	EuBL
*	2003-06-10	5	0.58	0.56	0.46	AtlTr
	median	5	0.58	0.61	0.35	

12 best-predicted HWs w.r.t **Tmax-EFI**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
*	2015-08-28	5	0.41	0.29	0.59	NoReg
	2006-07-17	12	0.56	0.64	0.53	EuBL
	2018-07-30	6	0.66	0.61	0.61	ScBL
*	2016-09-10	6	0.73	0.63	0.65	ScBL
*	2013-08-02	7	0.51	0.33	0.42	ScBL
	2003-06-10	5	0.58	0.56	0.46	AtlTr
*	2010-07-08	5	0.68	0.23	0.53	NoReg
	2005-09-05	6	0.56	0.44	0.44	NoReg
	2002-07-29	3	0.46	0.57	0.35	EuBL
	2003-09-17	6	0.63	0.07	0.48	NoReg
	2003-08-02	13	0.53	0.29	0.40	EuBL
	2015-07-01	5	0.71	0.44	0.53	ScBL
	median	6	0.57	0.44	0.505	

12 worst-predicted HWs w.r.t *Z*500-**ACC**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
	2018-05-26	7	0.64	−0.18	0.10	EuBL
	2016-05-07	5	0.55	−0.16	0.08	ScBL
*	2011-08-21	6	0.50	−0.13	0.18	ScBL
*	2003-05-30	7	0.47	−0.06	0.04	NoReg
	2013-06-17	4	0.64	−0.05	−0.11	Zonal
*	2018-08-06	4	0.70	0.00	0.32	NoReg
*	2003-09-17	6	0.63	0.07	0.48	NoReg
	2004-06-08	3	0.47	0.09	0.15	NoReg
*	2016-08-24	5	0.68	0.11	0.14	ScBL
*	2005-06-20	5	0.53	0.12	0.22	NoReg
	2006-06-11	4	0.50	0.13	0.13	EuBL
	2018-09-17	5	0.72	0.16	0.26	Zonal
	median	5	0.59	0.035	0.145	

12 worst-predicted HWs w.r.t **Tmax-EFI**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
*	2013-06-17	4	0.64	−0.05	−0.11	Zonal
*	2005-05-26	5	0.73	0.19	−0.08	NoReg
	2017-05-27	4	0.68	0.23	−0.05	NoReg
*	2002-08-16	5	0.38	0.46	−0.02	ScBL
*	2014-06-07	5	0.72	0.41	0.06	ScBL
	2003-05-30	7	0.47	−0.06	0.04	NoReg
*	2012-05-21	3	0.54	0.20	0.07	EuBL
	2016-05-07	5	0.55	−0.16	0.08	ScBL
*	2016-06-23	3	0.52	0.25	0.07	EuBL
	2018-05-26	7	0.64	−0.18	0.10	EuBL
*	2016-08-24	5	0.68	0.11	0.14	ScBL
	2004-08-05	6	0.43	0.68	0.10	ScBL
	median	5	0.595	0.195	0.065	

to be rather small, with less than 50% of the best predicted HWs in terms of *Z*500-ACC also being present in the group of those with the smallest Tmax-EFI error. A similar intersection fraction is also found in the group of the worst-predicted HWs. Most remarkable is the existence of cross-group intersections: for instance, one of the cases, a minor HW in early August 2004, was predicted well in terms of *Z*500-ACC (ensemble median 0.68), but the ensemble forecast failed to suggest anomalously high near-surface temperatures. Nearly 50% of the

worst-predicted HWs according to $Z500$-ACC are assigned to "no regime", whereas 75% of the best-predicted ones are categorised as either Scandinavian or European blocking. A similar distinction is not found for the Tmax-EFI-based selection metric. All in all, these findings regarding the outer predictability quartiles of HWs are consistent with the regime-dependent statistical evaluation in the first part of this article.

4.1.1 | The dominant role of the upstream dynamical setup

First we explore what may possibly impact the medium-range predictability of the large-scale flow patterns enabling HW onset over Central Europe. To do so, we use the ECMWF ensemble-median $Z500$-ACC at lead times of 9–11 days as our evaluation metric, in order to identify the HWs with either the best or the poorest large-scale predictability in the medium-range. The role of the initial atmospheric state shortly after forecast initialisation is visualised in Figure 9. In this figure we compare, for a number of atmospheric key quantities, the respective analysed anomalies of the atmospheric state one week before HW onset, so about 3 days into the forecast. Anomalies are presented in the form of composite-averaged deviations from the running 21-day climatology. Doing so, we exclude seasonal effects and instead highlight anomalies of the atmospheric state on synoptic to subseasonal time-scales.

Looking at the anomalous 500-hPa geopotential one week prior to HW onset, only a few significant deviations from summer climatology are detected for both the best-predicted and the most poorly predicted HW onsets (Figure 9a,b). However, considering more dynamic quantities, there is robust statistical evidence that the large-scale predictability of Central European HW onsets may be affected by certain characteristics of the atmospheric flow further upstream over the Western Atlantic. More precisely, we find that forecast busts in terms of the $Z500$-ACC score are on average associated with significant anomalies in both baroclinicity and, consistent with that, the intensity of the jet stream over the West Atlantic. The former, the increased baroclinicity, is represented by the anomalous Eady growth rate (Figure 9e,f). An extended zone of significantly increased Eady growth rates is found spanning from Newfoundland to the central Atlantic. We do not detect an accompanying significant signal in sea-surface temperatures off the coast of Eastern North America. This suggests that the increased baroclinicity is not related to an existing anomaly in the sea-surface temperature gradient in the region. Instead, we find that the increased meridional temperature gradient is caused by anomalously cold air masses over the Sea of Labrador (blue contours in Figure 9d). Since we exclude seasonal effects, we assume that the reason for this is southward advection of polar air masses in association with a synoptic-scale disturbance. In line with this, the composite mean of the worst-predicted HWs also exhibits anomalously low geopotential near the southern tip of Greenland (blue contours in Figure 9b).

Along with the differences in baroclinic instability, we may also expect differences in the state of the North Atlantic jet stream (NA jet). Indeed, for the group of the 12 worst-predicted HWs we find a robust sharpening of the jet stream. Zonal wind speed at 250 hPa is found to be significantly intensified around its climatological mean position, alongside a substantial weakening towards the north and to the south (Figure 9d).

In contrast to the forecast bust cases, well-predicted Central European HWs are not characterised by any significant atmospheric anomalies a week before HW onset (Figure 9a,c,e). There is therefore little evidence for a window of opportunity for enhanced predictability in terms of $Z500$-ACC. For both, well and poorly predicted Central European HWs, we further find no statistically robust indication that the large-scale predictability of HWs is associated with the preceding Euro-Atlantic weather regimes some 7–10 days before HW onset.

It has further been shown that moist diabatic processes may substantially impact the jet stream structure and thereby also downstream error growth (Grams *et al.*, 2018; Lojko *et al.*, 2022). In our understanding, a rather simple composite-mean approach like the one presented here is not sufficient to unravel the intricate details of the interaction of convection and large-scale dynamics. This is further exacerbated by the very low sample size we are naturally dealing with. Nonetheless, we have also looked into composite mean fields of quantities related to moist processes. For the best-predicted HWs, we again find a picture that is close to the running climatology (Figure S13). For the worst-predicted HWs, we detect anomalously humid air masses along the Atlantic coast of the United States, with a significant increase in integrated water vapour by about 6 kg/m^2. Such a substantial increase may not necessarily imply an important role of moist processes in impacting forecast quality. For precipitation, we indeed found a much less clear signal without any coherent significant anomalies. Hence, a substantial role of moist processes in downstream HW predictability over Central Europe can neither be confirmed nor ruled out in the scope of this study.

So far, we have only focused on lead times of 9–11 days and distinguished good and poor predictability via the median $Z500$-ACC. To extend this investigation, we

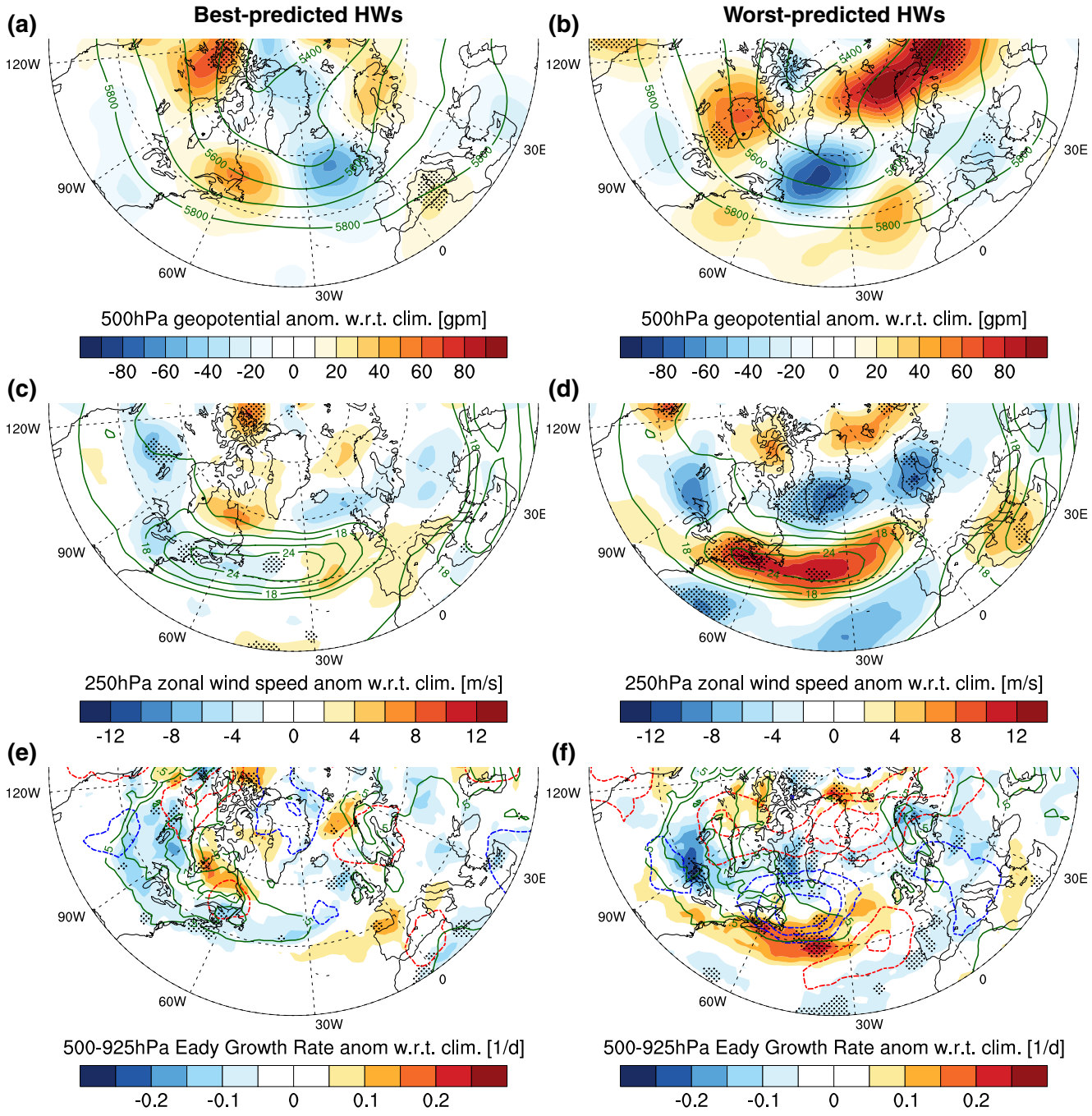

FIGURE 9 Comparison of analysed anomalies of the atmospheric state a week before HW onset between best- and worst-predicted Central European HWs at 9–11 days lead time according to the ECMWF reforecasts. Predictive skill for each HW onset is here evaluated by means of the median $Z500$-ACC. The pre-HW analysed atmospheric anomalies, depicted via shadings, are always calculated w. r. t. the running summer climatology and are then averaged over day 8, 7, and 6 before the respective HW onset. The left panels depict the pre-HW anomalies for the group of the 12 best-predicted HWs, whereas the right panels display anomalies for the 12 most poorly predicted HWs. The first row displays pre-HW 500-hPa geopotential anomalies (green lines always denote mean of all HW cases). In the second row, the pre-HW state of the North Atlantic jet stream is displayed by means of 250-hPa zonal wind anomalies. The last row visualises anomalies in baroclinicity in the form of the anomalous Eady growth rate. For additional information, the coloured contours denote anomalies in 700-hPa temperature (contour interval 1 K). Dots denote statistical difference of the best/worst sample-mean with respect to the mean of all 49 Central European HWs at the 5% level, as tested via a bootstrapping routine with 10,000 repetitions. [Colour figure can be viewed at wileyonlinelibrary.com]

make use of the fact that the upstream anomalies are most pronounced in the form of the state of the NA jet. We therefore measure the state of the NA jet by calculating the pattern correlation between the respective 250-hPa zonal wind-speed anomalies and the two dominant EOF modes of the same quantity, with the first pattern representing a northward-shifted jet and the second one featuring an intensified jet.

For forecasts with lead times of 9–11 days, we see results that are in line with the previously shown composite mean: around 8 days before HW onset, so two days after forecast initialisation, the worst-predicted HWs display significantly higher than normal spatial correlations with the intensified jet EOF pattern (Figure 10c). For lead time of 6–8 days, we again detect the strongest signals around 8 to 6 days before HW onset (Figure 10a). This time, this would mean there is already an existing anomaly during forecast initialisation. However, the differences in the state of the jet stream are less pronounced and not statistically significant. This is attributable to the fact that only 50% of the heatwaves with the worst $Z500$-ACC predictive skill at 9–11 days lead time are also found among the worst-predicted ones at 6–8 days lead time (see heatwaves marked with an asterisk in Table 5 or Table S2 in the Supporting Information for the list of best and worst-predicted Central European HWs at 6–8 days lead time).

When the quality of a forecast is decided by the relative error in Tmax-EFI rather than $Z500$-ACC, it is probably not surprising that we do not find any significant role of the North Atlantic jet (Figure 10b,d). Anomalies in the upstream dynamical state may seem to affect large-scale predictability, but do not necessarily lead to a better temperature forecast at the medium-range lead times considered.

Are the findings about the possible impact of upstream atmospheric conditions on HW onset predictability in Central Europe robust or specific to the ECMWF reforecasts? In the context of this investigation, this question is equivalent to asking whether the GEFSv12 generally agrees with ECMWF on which HWs are better or less predictable. Regarding the best-predicted HWs according to the $Z500$-ACC metric, the intersection is quite high. For our focus lead time of 9–11 days, we find that 83% of HWs are found to be the best-predicted in both datasets. Much greater disagreement exists for the worst-predicted HWs, with an intersection of only 42%. Therefore, when we repeat the same comparison of the initial atmospheric state, this time using the predictive skill of the GEFSv12 forecast dataset, we find a signal that is less robust. Worst-predicted HWs again show some intensification of the jet in its climatological mean latitude, but the intensification is narrower in terms of longitudinal extent and shifted much further to the east.

However, if we consider the fact that GEFSv12 features a slightly lower model-intrinsic predictability (not shown explicitly here), we may repeat this analysis, using 8–10 days lead time for evaluation instead of 9–11 days. Then the results become much closer compared with ECMWF (Figure S14). The intensified and narrowed NA jet structure is still found a bit further east over the Atlantic. Moreover, we also detect again a zone of significantly increased Eady growth rate in parts of the western Atlantic. We therefore conclude that the presence of upstream anomalies—an anomalously high meridional temperature gradient as well as an intensified NA jet—may indeed be a somewhat robust precursor for poor large-scale predictability of HW onsets over Central Europe.

4.1.2 | The role of pre-existing soil-moisture anomalies

Finally, we briefly explore whether pre-existing soil-moisture anomalies are associated with anomalous predictability of HW onsets over Central Europe. Figure 11 presents in the form of box plots, in a very similar manner to Figure 10, how the best- and worst-predicted HWs in ECMWF reforecasts differ in terms of analysed soil-moisture anomalies some days ahead of HW onset. We distinguish between regional soil-moisture anomalies (boxes without fill pattern) and supraregional soil-moisture anomalies, for which we spatially average over a much larger European area (boxes with star fill pattern). Pre-existing regional soil-moisture anomalies are not associated with better or worse predictive skill with regard to the large-scale circulation patterns as measured by $Z500$-ACC. However, some correlation with supraregional soil-moisture anomalies may exist. For forecasts with 6–8 days lead time, we find that the worst forecasts with respect to $Z500$-ACC are those for which the large-scale European soil moisture is significantly elevated above climatological levels some 7–10 days before HW onset (Figure 11a). For lead times of 9–11 days, we again find differences between well and poorly predicted HWs in a similar fashion. This time, it is the well-predicted HWs that feature significantly drier than normal soils 9 and 10 days before HW onset (Figure 11c). One could speculate that this finding would suggest that large-scale moisture deficits may not only impact local near-surface temperatures, but also exert some control over the large-scale circulation and its predictability. For instance, a blocking anticyclone could preferably be initiated and maintained over a region with anomalous sensible heating due to desiccated soils (Miller *et al.*, 2021). Similarly, Martius *et al.* (2021) also found an effect of

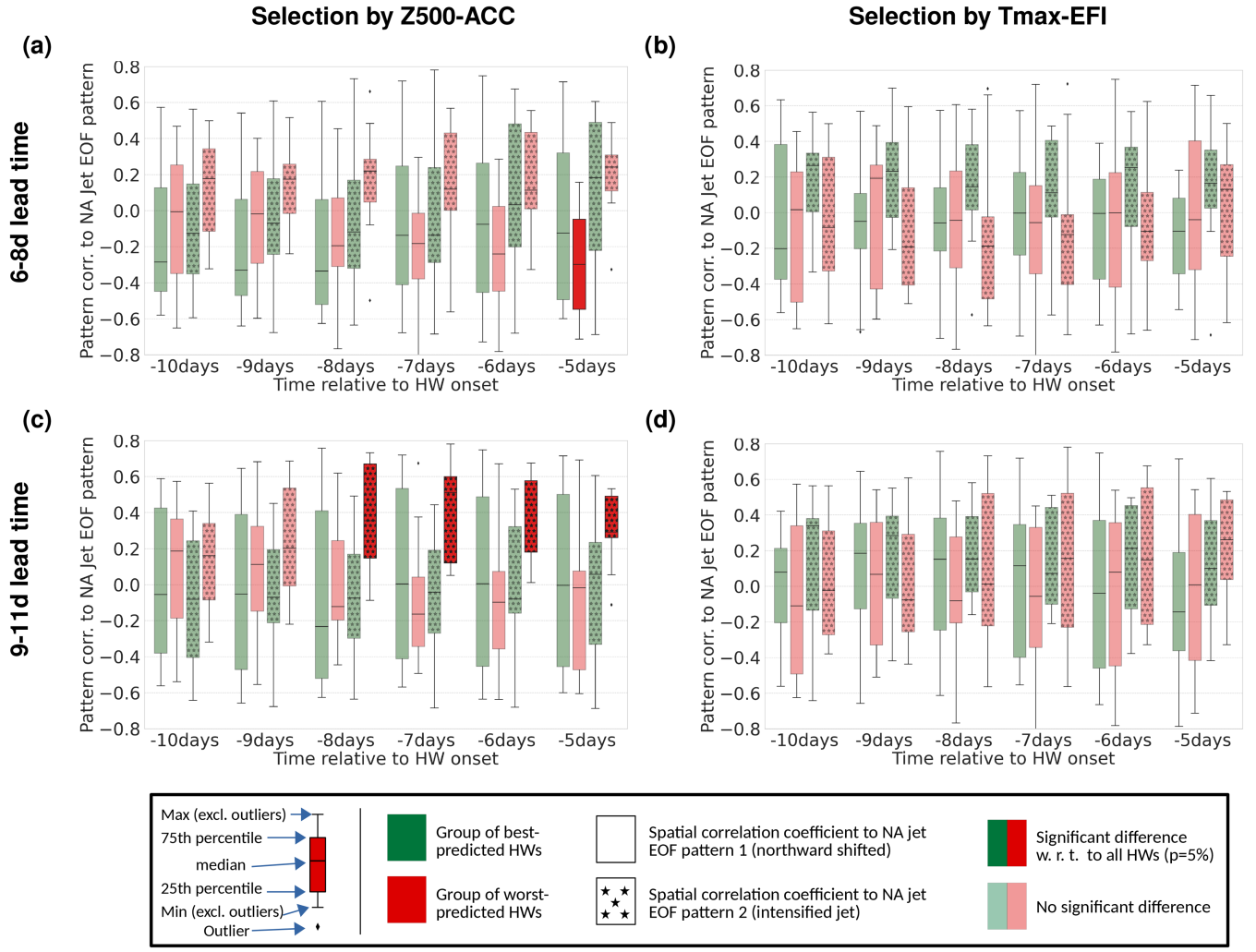

FIGURE 10 Relationship between ECMWF reforecast's predictive skill for Central European HW onsets and the state of the North Atlantic jet stream some days ahead of HW onset. The box plots show the pattern correlations between analysed anomalous zonal wind speed at 250 hPa and the two dominant EOF modes of the North Atlantic Jet for both the best-predicted (green boxes) and worst-predicted HW onsets (red boxes). The pattern correlation is evaluated over the same geographical region to which the EOF analysis was applied (75°W–5°W, 30°N–70°N). The boxes with the hash fill pattern denote the pattern correlation with respect to EOF mode 1 (northward-shifted NA jet) while the boxes without fill pattern show the correlations with mode 2 (intensified NA jet). The upper row presents results for the case of 6–8 day lead-time forecasts, while the bottom row displays results for forecasts with lead times 9–11 days. In the left panels, *Z*500-ACC is used as a selection metric, the right panels feature the relative error in Tmax-EFI as selection metric. Further details about the meaning of the box and whiskers, as well as the way statistical significance is displayed, are provided in the legend at the bottom. [Colour figure can be viewed at wileyonlinelibrary.com]

soil-moisture anomalies over Australia on the upper-level flow both locally and in remote regions. On the other hand, the significant soil moisture differences presented here could instead be a mere effect of anomalously wet or dry weather conditions preceding the heatwaves of the respective sample. A more detailed analysis of a possible causal relationship between soil-moisture anomalies and large-scale flow predictability may be the subject of future studies.

Finally, we assess the role of existing soil-moisture anomalies in the quality of forecasts, with regard to how well the extremeness of near-surface temperatures is captured (as measured by the relative error in Tmax-EFI). ECMWF ensemble reforecasts with a lead time of 6–8 days seem to underestimate near-surface temperatures during Central European HW onsets the most when soils are significantly wetter than normal during the time of forecast initialisation (Figure 11b). This is true for both regional soil-moisture anomalies and anomalies averaged over a larger European scale. An even clearer picture emerges for lead times of 9–11 days (Figure 11d). Again, the worst-predicted HWs in terms of Tmax-EFI are associated

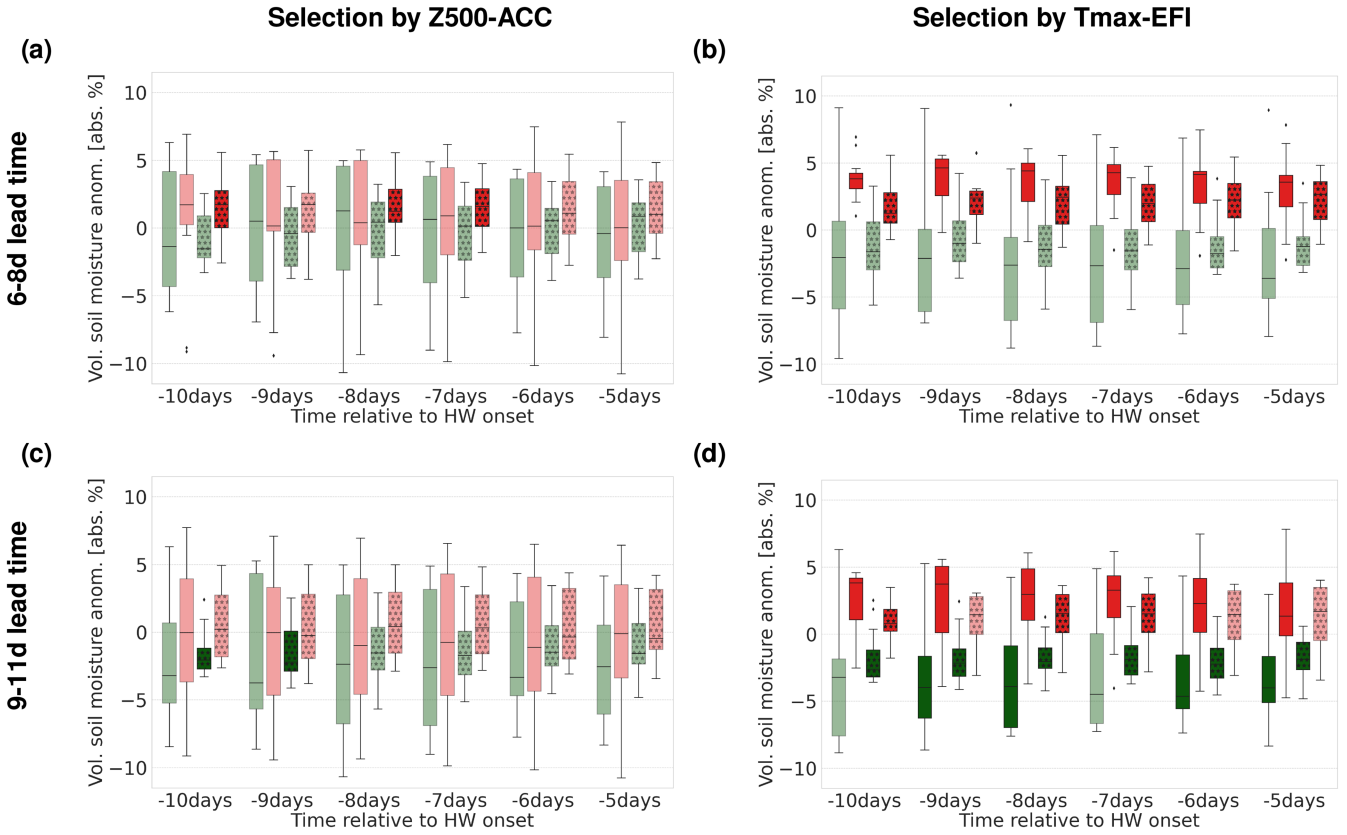

FIGURE 11 Similar to Figure 10, but for analysed anomalous soil moisture instead of jet stream anomalies. The box plots depict analysed 0–20 cm soil-moisture anomalies (from ECMWF reforecast analysis) with respect to summer climatology area-averaged over different spatial scales. The boxes without fill pattern represent anomalies on the regional Central Europe scale (4°E–16°E, 47.5°N–55°N), while boxes with the star fill pattern denote anomalies on a larger supraregional scale (5°W–35°E, 45°N–60°N). Statistically significant differences compared with all HW cases are again denoted by using darker bold colours. [Colour figure can be viewed at wileyonlinelibrary.com]

with anomalously wet soils a week before HW onset. Now, a significant result is also found for well-predicted HWs, suggesting that significantly desiccated soils are linked to particularly good predictability of near-surface temperatures.

Using GEFSv12 instead of ECMWF reforecasts, pre-existing regional and supraregional soil-moisture anomalies are again found to be associated with differences in predictive skill, but only for near-surface temperatures. A direct comparison between both models is complicated both by a slightly different treatment of soil moisture and the fact that TMAX-EFI was not computed because of the low ensemble size in GEFSv12. We have therefore assessed predictive skill alternatively by means of the area-averaged maximum temperature error, and then created a best/worst-predicted HW sample from ERA5 soil-moisture anomalies (first two levels). On longer lead times (9–11 days for ECMWF and again 8–10 days for GEFS), we find good agreement between both models, with better near-surface temperature prediction skill being related to significant pre-existing dry anomalies and vice versa (not shown). On shorter lead times (6–8 and 5–7 days, respectively), we no longer find a clear link between predictive skill for temperature and antecedent soil-moisture anomalies, which is likely due to the choice of a different evaluation metric.

4.2 | What affects the predictability of British Isles HW onsets?

Table 6 lists and compares the characteristics of the best- and worst-predicted HW onsets over the British Isles at 9–11 days lead time, again with respect to the selection metrics *Z*500-ACC and relative error in Tmax-EFI applied to ECMWF reforecasts. In comparison with the Central European HWs, the selected British HWs are in most cases a bit shorter (median of 4 days) and generally less intense. The most poorly predicted HWs are a notable exception, according to the Tmax-EFI underestimation with a median length of 6 days. Similar to the

TABLE 6 As Table 5, but for the best- and worst-predicted HW onsets over the British Isles.

12 best-predicted HWs w.r.t *Z*500-**ACC**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
*	2009-06-29	4	0.62	0.79	0.51	ScBL
*	2018-07-25	3	0.67	0.69	0.48	ScBL
*	2016-06-04	5	0.37	0.63	0.25	AtlRi
*	2006-07-01	5	0.49	0.62	0.18	EuBL
*	2007-06-08	4	0.49	0.59	0.48	ScBL
	2006-07-15	8	0.46	0.57	0.48	EuBL
	2004-08-07	4	0.68	0.54	0.40	ScBL
	2016-09-12	4	0.59	0.53	0.29	ScBL
*	2002-09-11	4	0.45	0.51	0.20	ScBL
	2003-08-04	10	0.73	0.50	0.49	EuBL
*	2014-09-16	4	0.51	0.47	0.24	EuBL
	median	4	0.51	0.57	0.40	

12 best-predicted HWs w.r.t **Tmax-EFI**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
	2018-08-03	5	0.38	0.22	0.50	NoReg
*	2006-07-15	8	0.46	0.57	0.48	EuBL
	2007-06-08	4	0.49	0.59	0.48	ScBL
*	2007-05-01	4	0.44	0.37	0.42	NoReg
*	2001-07-02	4	0.47	0.02	0.40	EuBL
	2009-06-29	4	0.62	0.79	0.51	ScBL
*	2013-07-07	17	0.38	0.38	0.30	EuBL
	2018-05-23	11	0.31	0.44	0.25	EuBL
	2006-05-08	5	0.33	0.47	0.25	ScBL
	2016-08-23	3	0.31	0.23	0.23	ScBL
*	2018-07-25	3	0.67	0.69	0.48	ScBL
	median	4	0.44	0.44	0.42	

12 worst-predicted HWs w.r.t *Z*500-**ACC**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
*	2005-06-17	4	0.57	−0.33	−0.00	ScBL
	2018-06-25	15	0.45	−0.18	−0.01	EuBL
	2005-07-10	4	0.44	−0.13	0.02	EuBL
*	2016-07-18	3	0.63	−0.07	0.22	NoReg
*	2003-07-13	5	0.60	−0.07	−0.02	ScBL
*	2001-05-10	4	0.54	−0.06	0.05	NoReg
	2001-07-02	4	0.47	0.02	0.40	EuBL
*	2016-09-05	3	0.65	0.04	0.23	Zonal
	2016-05-07	6	0.55	0.05	−0.17	ScBL
*	2003-05-29	4	0.52	0.05	0.04	NoReg
	2008-05-04	11	0.58	0.08	0.07	ScBL
	median	4	0.55	−0.06	0.04	

12 worst-predicted HWs w.r.t **Tmax-EFI**

	Onset	Len	EFI_{ana}	ACC	$EFI_{predmax}$	Regime
	2016-05-07	6	0.55	0.05	−0.17	ScBL
*	2006-06-06	7	0.36	0.41	−0.11	EuBL
*	2003-07-13	5	0.60	−0.07	−0.02	ScBL
	2013-09-22	3	0.35	0.10	−0.01	GrBL
	2018-06-25	15	0.45	−0.18	−0.01	EuBL
*	2005-06-17	4	0.57	−0.33	−0.00	ScBL
*	2005-07-10	4	0.44	−0.13	0.02	EuBL
*	2012-05-22	7	0.40	0.37	0.01	EuBL
*	2004-09-03	8	0.51	0.14	0.03	EuBL
	2010-05-20	6	0.47	0.47	0.02	EuBL
	2003-05-29	4	0.52	0.05	0.04	NoReg
	median	6	0.47	0.05	0.0	

Central European case, not even half of the HWs with the best large-scale predictability scores are also found among the best-predicted with respect to near-surface temperatures (intersection fraction of the two selection metric groups is smaller than 50%). However, a higher overlap of slightly more than 50% is found for the worst-predictable HWs. One cross-group intersection case exists, in which Tmax-EFI was predicted well despite a *Z*500-ACC of around zero. As was the case for Central European HWs, a clear majority of well-predicted HWs in terms of *Z*500-ACC fall into the category of either Scandinavian or European blocking. Since "no regime" HWs are generally less common for the British Isles, their share among the poorly predicted is also lower than for the Central European case, but still higher compared with the well-predicted ones, which do not feature a single case of "no regime".

4.2.1 | The dominant role of preceding Euro-Atlantic weather regimes

As in the earlier section about the predictability of Central European HWs, we compare again the extent to which well and poorly predicted HW onsets are characterised by pre-existing anomalies in the atmospheric state. For this purpose, Figure 12 provides the same composite-mean comparison of the best- and

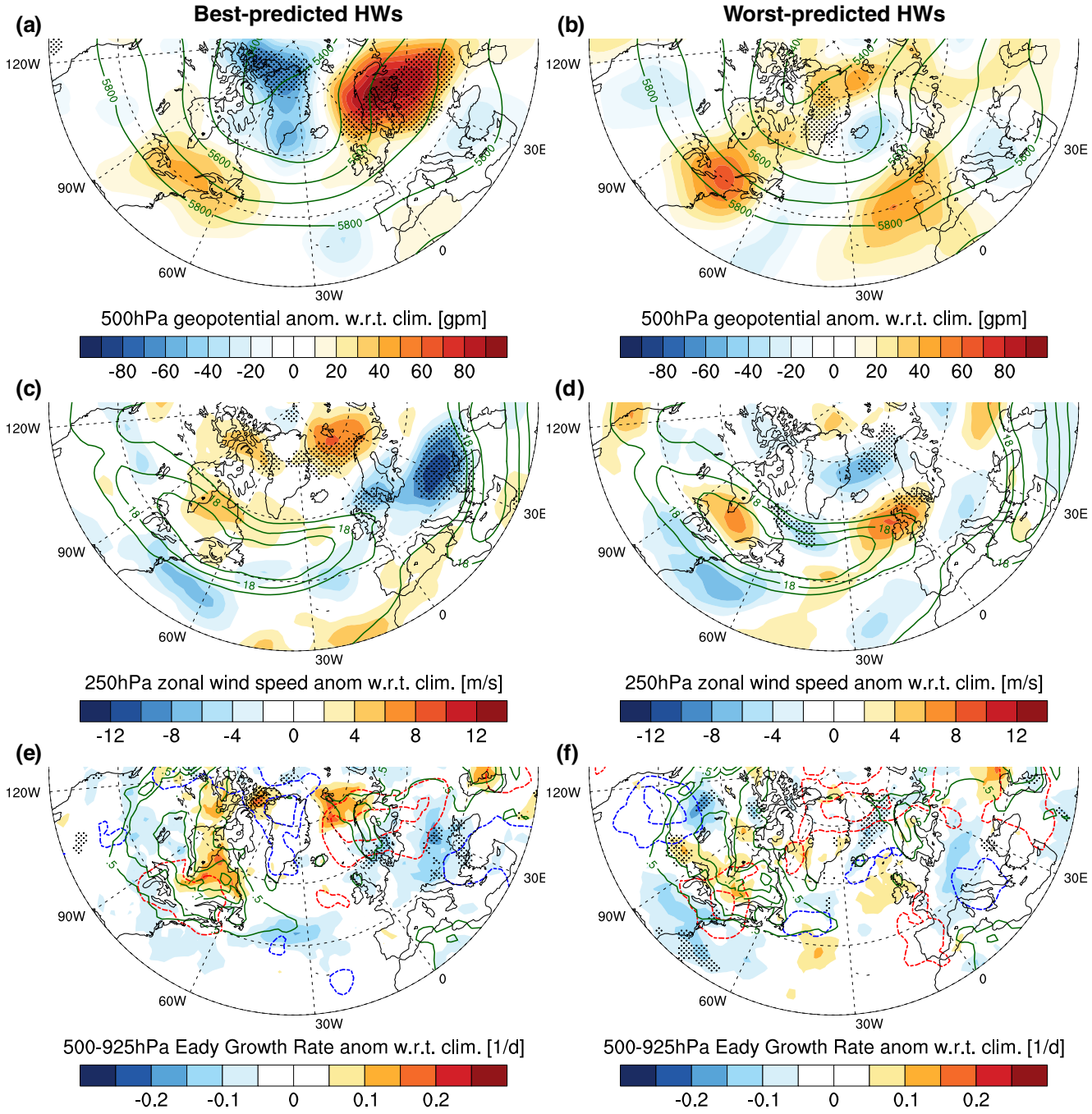

FIGURE 12 As Figure 9, but for the best- and worst-predicted HW onsets over the British Isles. [Colour figure can be viewed at wileyonlinelibrary.com]

worst-predicted HWs according to $Z500$-ACC at lead times of 9–11 days, now for the British Isles HW cases.

In stark contrast to the Central European HW case, we no longer find robust evidence for HW onset predictive skill being impacted by upstream anomalies. For both the best- and the worst-predicted HWs, we do not detect major significant anomalies a week prior to HW onset, either for the baroclinicity over the Western Atlantic or for the state of the NA jet (Figure 12c–f). Instead, the large-scale predictability of British HWs seems to be much more associated with the pre-existing weather regimes over the European continent. In contrast to the Central European case, it is now the best-predicted HWs that already exhibit a substantial anomaly a week before HW onset. During that time, the composite mean depicts a large area of significantly elevated 500-hPa geopotential over large parts of northeastern Europe (Figure 12a). This continental anticyclone then slowly moves/extends westwards, finally

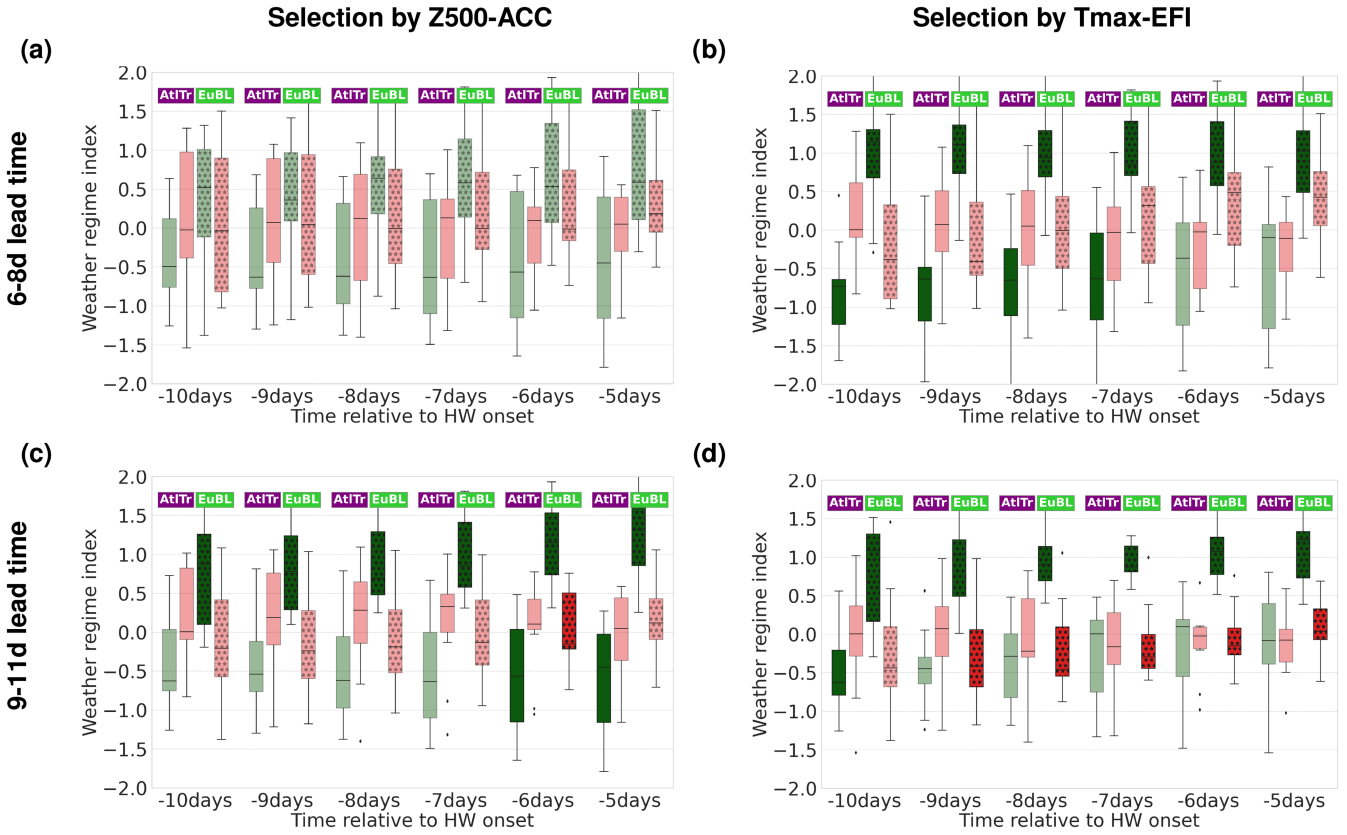

FIGURE 13 Relationship between ECMWF reforecast's predictive skill for British Isles HW onsets and the amplitude of two selected weather regimes some days ahead of HW onset. The box plots show the analysed instantaneous weather regime index for the Atlantic Trough regime (boxes without fill pattern) and the European Blocking regime (boxes with star fill pattern) for the period of 10–5 days before HW onset. The upper row presents results for the case of 6–8 day lead-time forecasts, while the bottom row displays results for forecasts with lead times of 9–11 days. In the left panels, *Z*500-ACC is used as a selection metric, the right panels feature Tmax-EFI as selection metric. Statistically significant differences compared with all HW cases are again denoted by using darker bold colours. [Colour figure can be viewed at wileyonlinelibrary.com]

enabling the development of a HW over the British Isles region. The composite-mean synoptic pattern presented here bears some resemblance to the case of sequential HWs of 2018 described in detail by Spensberger *et al.* (2020). In contrast to the Central European case, it is now the poorly predicted HW onsets that feature an atmospheric state with no significant anomalies to summer climatology during or shortly after forecast initialisation (Figure 12b).

As was the case for the Central European HWs, we extend our statistical investigation to shorter lead times (6–8 days), as well as to the Tmax-EFI-based selection metric. Making use of the analysed Euro-Atlantic weather regime indices, we identified the temporal evolution of the amplitude of the European blocking and the Atlantic Trough weather regimes as being strongly linked to the predictive skill for British HW onsets. For 9–11 days lead time, well-predicted HWs according to *Z*500-ACC first exhibit a significantly elevated EuBL weather regime index already 10 days before HW onset (Figure 13c), consistent with the composite-mean plots shown earlier. At the same time, low-pressure systems are slightly, albeit not yet significantly, less likely to be close to the European continent as represented by the lowered amplitude of the Atlantic Trough regime index. Interestingly, for lower lead times of 6–8 days, the *Z*500-ACC-based HW onset predictive skill is less sensitive to prior anomalies in those weather regimes (Figure 13a). Although seven of the 11 HWs with the highest *Z*500-ACC at 9–11 days lead time are also found among the best-predicted HWs at 6–8 days lead time (see HWs marked with asterisks in Tables 6 or S3 in the Supporting Information), the overall small sample size may lead rather quickly to considerable differences when four HWs within the sample are replaced by others.

In stark contrast to the Central European case, the clear association between antecedent weather regimes and *Z*500-ACC predictability is now also clearly reflected in the forecast quality with respect to Tmax-EFI. For lead times of 6–8 days, a good forecast in near-surface temperature extremeness is already associated with the presence

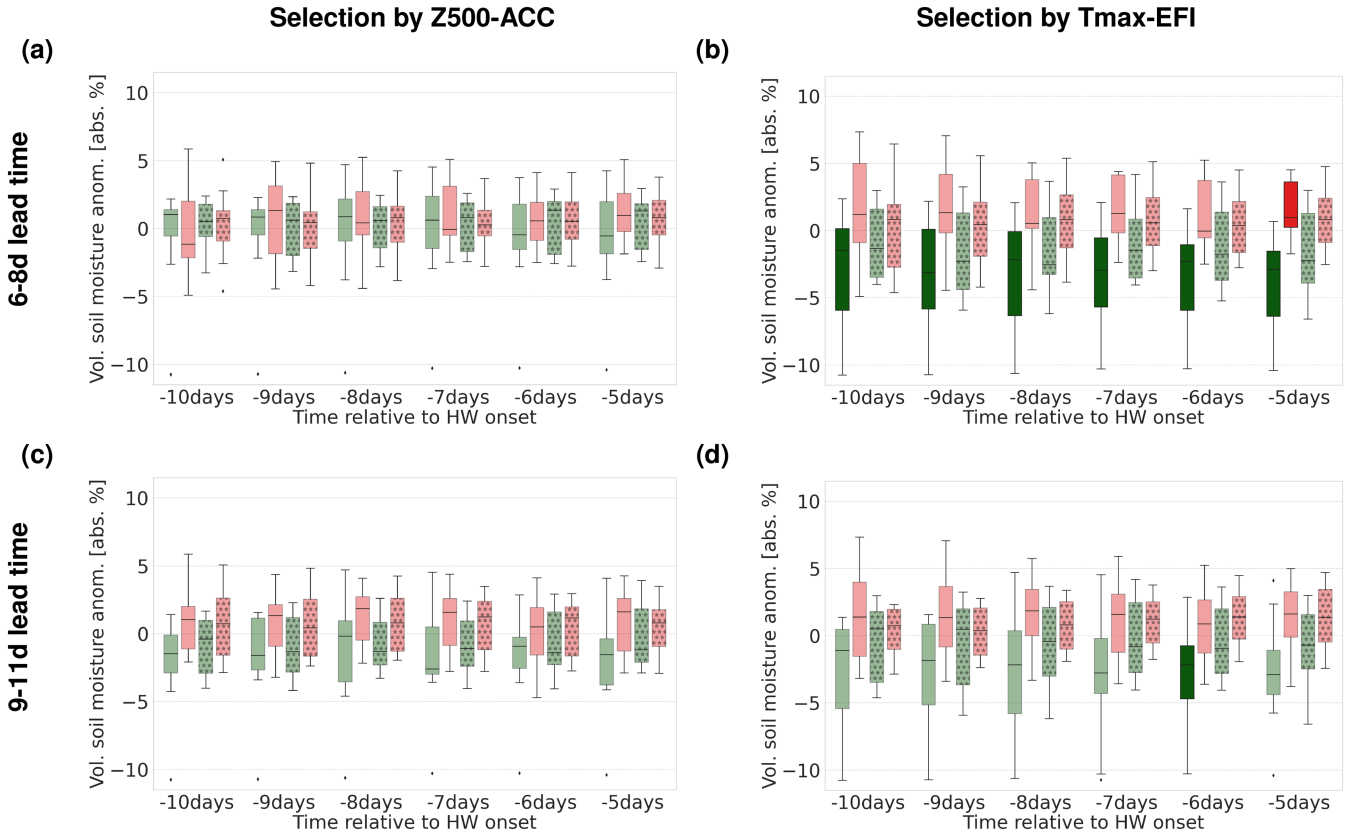

FIGURE 14 As Figure 11, but for the best- and worst-predicted HW onsets over the British Isles. [Colour figure can be viewed at wileyonlinelibrary.com]

of European blocking before the time of forecast initialisation. At the same time, the index of the Atlantic Trough weather regime is significantly lower than in the sample of all HWs (Figure 13b). In other words, when a continental blocking is already established and low-pressure system activity is reduced over the Eastern Atlantic, a good temperature forecast becomes more likely for the British Isles region. A similar picture is found if lead times of 9–11 days are considered. A significantly elevated expression of the EuBL index during the time of forecast initialisation is favourable for a good temperature prediction (Figure 13d). Poor forecasts, in turn, suffer from a significantly reduced amplitude of the EuBL weather regime. In contrast to lower lead times, the amplitude of the Atlantic Trough weather regime has less impact on forecast quality.

Finally, we again test the robustness of our results by performing the same analyses with our secondary forecast dataset (Figure S15). We find that GEFSv12 agrees quite well with ECMWF on which British HWs are the best-predictable according to the median $Z500$-ACC. For 9–11 days lead time, the intersection fraction amounts to 73% and for 6–8 days lead time there is still an overlap of 64%. Consistent with the high agreement between both models leading to the selection of a rather similar subset, we find a very similar and even more robust signal for the best-predicted HWs over the British Isles: significantly increased geopotential is first detected some 8 days before HW onset, and then slowly moves retrogradely towards the British Isles. As Tmax-EFI has not been calculated for GEFSv12 due to the low ensemble number, we instead separate HW onsets by the domain-averaged prediction error in maximum temperature. Again, our secondary reforecast dataset agrees strongly with ECMWF. Hence, the potential for increased predictability of British HWs in times of an already well-established continental blocking appears to be a robust finding.

4.2.2 | The role of pre-existing soil-moisture anomalies

Antecedent anomalies in soil moisture, either locally over the British Isles or over a larger regional area including mainland Europe, are not associated with better or worse large-scale predictability ($Z500$-ACC) in British HWs according to ECMWF reforecasts (Figure 14a,c). Compared with Central European HWs, pre-existing soil-moisture anomalies over the British Isles are also not as strongly linked to the predictive skill in terms of surface temperature extremeness. This is particularly true

for longer lead times of 9–11 days (Figure 14d). Only at lead times of 6–8 days do we find that Tmax-EFI scores are better when the soils over the British Isles are significantly drier than normal (Figure 14b). A likely explanation of the overall less pronounced role of soil moisture is the maritime location and probably also the overall lower likelihood of substantial desiccation of the soils. For anomalously high temperatures to develop, the British Isles probably rely more strongly on the correct prediction of a rather narrow range of synoptic conditions (strong subsidence, no onshore winds), whereas it is easier for anomalous heat to develop over the continental region via land surface–atmosphere interaction (Zschenderlein *et al.*, 2019).

Again we perform a similar kind of analysis, now also including the GEFSv12 reforecasts. As mentioned earlier in the section about Central European HWs, for better comparability between ECMWF and GEFSv12, we have to change the evaluation metric to area-averaged maximum temperature error and then construct the best-/worst-predicted HW samples using ERA5 soil-moisture anomalies with respect to climatology. Doing so, we find for both models a somewhat stronger connection between pre-existing soil-moisture anomalies and temperature prediction skill, now also for longer lead times of 9–11 (ECMWF) and 8–10 days (GEFSv12). Again, better temperature forecasts are being associated with drier soils some seven days before HW onset, whereas poorer prediction skill is related to cases of anomalously wet soils. However, these findings are not as clear as for the Central European case and therefore are not in disagreement with our hypothesis that soil-moisture anomalies may have less impact over the British Isles compared with the European mainland.

5 | SUMMARY AND CONCLUDING DISCUSSION

In this study, we investigated the medium-range predictability of European heatwaves (HWs) that occurred in the months May–September in the period 2001–2018 by using ensemble reforecasts from two state-of-the-art weather models: primarily ECMWF reforecasts, with GEFSv12 reforecasts as support. For four different mid-latitude European regions (Central Europe, British Isles, Scandinavia, Western Russia), we objectively identified around 50 HWs each, which are characterised by a local and regional-scale exceedance of the 90th percentile of maximum temperature for a duration of at least three days. We applied the concept of year-round Euro-Atlantic weather regimes (Grams *et al.*, 2017) to first characterise the large-scale synoptic setup under which HWs form. The findings for the four European regions investigated can be summarised as follows.

- Over the British Isles and Scandinavia, HW onsets are associated with a classic blocking regime, that is, either Scandinavian (ScBL) or European blocking (EuBL), in 66% and 75% of cases, respectively.
- Over Central Europe, ScBL or EuBL cases are less often observed (55% in total) and HW onsets occur more often, in about 30% of cases, in the absence of a pronounced weather regime (NoReg case).
- Over Western Russia, HWs occur most often in conjunction with a Scandinavian trough or in the absence of a regime.

Stratified by observed weather regime during HW onset, forecast quality was then evaluated for a range of lead times from 3–12 days using classic verification metrics such as the 500-hPa geopotential anomaly correlation coefficient (Z500-ACC) and root-mean-square errors (nRMSE) of 850-hPa temperature fields. In addition, we further computed relative errors in the maximum temperature extreme forecast index (Tmax-EFI-RE) as a more forgiving metric to review the forecast ensemble's capacity to adequately predict the likelihood of extreme near-surface temperatures. The following main results have been found to be robust in both the ECMWF and the GEFSv12 reforecasts.

- For Central Europe and the British Isles, HW onsets associated with pronounced blocking regimes (ScBL or EuBL) are linked to overall better and sometimes also significantly elevated skill in the prediction of large-scale synoptic-scale patterns (Z500-ACC), mostly up to lead times of about 10 days.
- For Central Europe and particularly the British Isles and Scandinavia, 850-hPa temperature fields tend to be predicted significantly better at lead times beyond one week for HW onsets associated with Scandinavian blocking (cf. against HW onsets of all regimes, not against non-HW periods).
- In all four regions studied, HW onsets linked to "no regime" mostly display worse Z500-ACC, Z500-nRMSE, and T850-nRMSE scores at medium-range (exception: Z500-ACC for Scandinavian HW onsets).
- The significant differences between weather regimes are not reflected in the Tmax-EFI-RE score, which points to difficulties in the prediction of near-surface temperatures even when large-scale predictability is high.

Using a multi-metric approach, the first part of our study highlights the intricate details of medium-range

forecast behaviour regarding HW onsets in the midlatitude regions of Europe. We have demonstrated that HWs forming under an already well-established continental blocking regime (ScBL and/or EuBL) may offer higher predictability with regard to the correct positioning of ridges and troughs and the introduction of anomalously warm air masses. In contrast, HWs forming in the absence of a pronounced blocking regime (NoReg) may be more prone to prediction errors, which are mostly related to phase errors. Therefore, a somewhat skilful forecast of a regime transition may then also imply either increased confidence (EuBL, ScBL) or higher uncertainty (NoReg) for the regionally accurate prediction of HW onsets. However, "getting the overall large-scale Rossby-wave pattern right" is of course not always sufficient in order to predict a HW onset correctly. This was demonstrated for the case of Scandinavian HWs, where "no regime" type HW onsets featured better $Z500$-ACC scores, but clearly worse predictive skill in terms of more regional temperature-based metrics. Predicting the likelihood of extreme near-surface temperatures proves to be more difficult. The increased predictive skill for $Z500$-ACC and T850n-RMSE—as clearly seen for Central European and and British HWs—is not reflected at all in our used Tmax-EFI error metric. This disagreement may mostly point to the inherent difficulty in predicting local near-surface temperatures correctly. As Lemburg and Fink (2022) demonstrated, even at short lead times of 3 days, maximum temperatures might generally be underestimated during HWs, and errors may often stem from under- or overestimation of more local diabatic processes restricted to the boundary layer or errors in the details of the near-surface flow in coastal regions.

In the second part of our investigations, we aimed to understand the role of the initial state of the atmosphere and soil moisture in affecting the predictability of HW onsets over Central Europe and the British Isles. First we use the $Z500$-ACC metric to select the best- and worst-predicted HWs with respect to the prediction of the large-scale Rossby-wave patterns. Secondly, we assess how well the extremeness of temperatures in our region of interest is captured by considering the relative error in the predicted maximum temperature extreme forecast index (Tmax-EFI-RE). Despite the very limited sample size, we could identify for both European regions statistically robust links between HW predictability and significant anomalies in the initial atmospheric state during or shortly after forecast initialisation. The results for Central Europe can be summarised as follows.

- For Central European HWs, poor $Z500$-ACC predictive skill at lead times of 9–11 days is associated with significant upstream anomalies occurring about one week before HW onset: a stronger than normal North Atlantic jet stream and increased baroclinicity over the western Atlantic, both likely related to synoptic-scale disturbances.
- Well-predicted Central European HWs, however, feature nearly no significant pre-existing anomalies compared with summer climatology.
- Predictive skill for near-surface temperature extremes over Central Europe is significantly connected with pre-existing regional and supraregional soil-moisture anomalies on lead times between 6 and 11 days; worse skill when soils are anomalously wet during forecast initialisation, and vice versa.

These key findings regarding poorly predictable Central European HWs are to some extent supported by the GEFSv12 reforecasts, but only if lead times from 8–10 days are considered and with overall lower statistical confidence. In our opinion, the impact of upstream anomalies on the large-scale predictability of weather further downstream is generally consistent with earlier theoretical or idealised studies stressing that increased meridional temperature gradients will generally reduce the intrinsic predictability of midlatitude weather (Sheshadri *et al.*, 2021; Vallis, 1983). On the other hand, an intensified jet stream acting as a wave guide might also increase downstream predictability (Wirth *et al.*, 2018). However, this applies in our view only to the predictability of Rossby-wave packets, and does not necessarily imply improved predictability of the correct phase, which is vital for a regionally accurate HW forecast.

Interestingly, British heatwaves show entirely different predictability characteristics, which are found robustly in both forecast datasets.

- In contrast to Central Europe, the predictive skill for British HW onsets in terms of both $Z500$-ACC and Tmax-EFI is not linked to upstream flow anomalies.
- Instead, better predictions at lead times of 6–11 days w. r. t. both $Z500$-ACC and Tmax-EFI are strongly linked to significantly more pronounced continental blocking existing already a week before HW onset.
- Compared with Central European heatwaves, pre-existing soil-moisture anomalies are less relevant for predictive skill; only on lead times of 6–8 days are regionally drier than normal soils associated with better predictive skill for surface temperature extremeness.

The second part of our article may provide some understanding of causes of HW-related forecast busts—at least for Central Europe—and may point further to situations with higher inherent predictability, such as for the British HW case. Moreover, the investigation underlines

the complex nature of atmospheric predictability again. Although Central Europe and the British Isles are geographically close to each other, the same inferences about HW onset predictability for one region cannot be transferred to the other. Such regional differences in HW predictability have also been pointed out before (Pyrina & Domeisen, 2023; Wulff & Domeisen, 2019), highlighting the need for region-specific studies. In this context, we want to emphasise again the use of different forecast skill metrics. As we have seen, a useful forecast of the large-scale flow might not be closely linked to a skilful prediction of near-surface maximum temperatures and this relationship might differ between regions as well.

One of the limitations of this study and probably any other statistical study about HWs is undoubtedly the naturally low sample size. We aimed to mitigate this problem by including a rather large number of HWs, some of them being rather short-lived and not extreme by any means. One the one hand, the consideration of about 50 HW cases in combination with the bootstrapping tests conducted allowed us to assess the statistical robustness of our results. On the other hand, the inferences drawn in this study might not apply to the most extreme HWs, for which some increased predictability was shown earlier by Wulff and Domeisen (2019). Finally, we want to stress again the intricacies of selecting the most appropriate metrics in order to evaluate HW predictability objectively. Prior studies have often used probabilistic scores, such as the Brier score, which are based on binary classification with respect to some near-surface temperature threshold. We propose that it may be beneficial also to factor in more classic scores such as $Z500$-ACC, which also account for how well the evolution of the large-scale flow is predicted. This comes, however, with its own drawbacks. As geopotential fields tend to exhibit large and spatially homogeneous deviations from climatology in times of HW onsets, $Z500$-ACC might tend to produce too favourable scores.

As outlined in the Introduction, we aimed to reveal the intricate details of large-scale and regional-scale predictability in relation to the weather regime during HW onsets. For this, we deliberately focused on using weather regimes for stratification of HWs into dynamically distinct subsets, but have not looked more deeply into the predictability of the weather regimes themselves. Although literature on European weather regime predictability for summer already exists (Büeler *et al.*, 2021; Osman *et al.*, 2023), further studies may aim to investigate more specifically how HW-allowing weather regimes develop and whether windows for enhanced predictability may exist. Moreover, the possible role of preceding anomalies in the atmospheric flow further upstream, as demonstrated here for poorly predicted Central European HWs, may also motivate further study. Ultimately, synoptic-scale deviations from summer climatology such as increased baroclinicity and an intensified jet stream may be favoured under a certain dynamical regime. Year-round North American weather regimes, as recently introduced by Lee *et al.* (2023), may therefore offer an opportunity to understand better the possible connection between anomalous weather over North America and a downstream impact on HW predictability over Central Europe.

ACKNOWLEDGEMENTS

This research has been supported by the Deutsche Forschungsgemeinschaft (grant no. SFB/TRR 165, "Waves to Weather") and conducted within the subproject C4: "Predictability of European heatwaves". We thank Christian Grams and his working group "Large-scale dynamics and predictability" for providing weather regime data for both reanalyses and ECMWF reforecasts. Further thanks go out to Georgios Fragkoulidis for providing the algorithm used for the identification of heatwaves, as well as for the preparation of some advanced data analyses. We are also grateful to the two anonymous reviewers, whose comments and suggestions helped to improve this article.

DATA AVAILABILITY STATEMENT

ECMWF reforecasts used in this study are publicly available on the website of the ECMWF at the following link: https://apps.ecmwf.int/datasets/data/s2s-reforecasts-daily-averaged-ecmf/levtype=sfc/type=cf/.

The reforecasts of the GEFSv12 model created at NCEP can be freely downloaded at: https://noaa-gefs-retrospective.s3.amazonaws.com/index.html#GEFSv12/reforecast/.

The ERA5 data used for the detection of heatwave are provided for the public at the following URL: https://cds.climate.copernicus.eu/cdsapp/#!/dataset/reanalysis-era5-single-levels?tab=overview.

The weather regime data are produced and stored by the working group of Christian Grams at the Karlsruhe Institute of Technology and may be provided upon request (grams@kit.edu). Scripts used to generate the plots of this article can be provided by the corresponding author upon request (alexander.lemburg@kit.edu).

ORCID

Alexander Lemburg https://orcid.org/0000-0001-8059-0546

Andreas H. Fink https://orcid.org/0000-0002-5840-2120

REFERENCES

Andersson, E., Persson, A. & Tsonevsky, I. (2015) User guide to ecmwf forecast products. *ECMWF, v2*, (1), 121.

Baumgart, M., Ghinassi, P., Wirth, V., Selz, T., Craig, G.C. & Riemer, M. (2019) Quantitative view on the processes governing the

upscale error growth up to the planetary scale using a stochastic convection scheme. *Monthly Weather Review*, 147, 1713–1731.

Büeler, D., Ferranti, L., Magnusson, L., Quinting, J.F. & Grams, C.M. (2021) Year-round sub-seasonal forecast skill for atlantic–european weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 147, 4283–4309.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P. et al. (2013) Long-term climate change: projections, commitments and irreversibility. In: *Climate change 2013-the physical science basis: contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, UK: Cambridge University Press, pp. 1029–1136.

Dole, R., Hoerling, M., Perlwitz, J., Eischeid, J., Pegion, P., Zhang, T. et al. (2011) Was there a basis for anticipating the 2010 russian heat wave? *Geophysical Research Letters*, 38.

Domeisen, D.I., Eltahir, E.A., Fischer, E.M., Knutti, R., Perkins-Kirkpatrick, S.E., Schär, C. et al. (2023) Prediction and projection of heatwaves. *Nature Reviews Earth & Environment*, 4(1), 36–50.

Ebi, K.L., Teisberg, T.J., Kalkstein, L.S., Robinson, L. & Weiher, R.F. (2004) Heat watch/warning systems save lives: estimated costs and benefits for philadelphia 1995–98. *Bulletin of the American Meteorological Society*, 85, 1067–1074.

Ferranti, L., Corti, S. & Janousek, M. (2015) Flow-dependent verification of the ecmwf ensemble over the euro-atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141, 916–924.

Ferranti, L., Magnusson, L., Vitart, F. & Richardson, D.S. (2018) How far in advance can we predict changes in large-scale flow leading to severe cold conditions over europe? *Quarterly Journal of the Royal Meteorological Society*, 144, 1788–1802.

Fink, A.H., Brücher, T., Krüger, A., Leckebusch, G.C., Pinto, J.G. & Ulbrich, U. (2004) The 2003 european summer heatwaves and drought–synoptic diagnosis and impacts. *Weather*, 59, 209–216.

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.

García-Herrera, R., Díaz, J., Trigo, R.M., Luterbacher, J. & Fischer, E.M. (2010) A review of the european summer heat wave of 2003. *Critical Reviews in Environmental Science and Technology*, 40, 267–306.

Gómez, I., Niclòs, R., Estrela, M.J., Caselles, V. & Barberà, M.J. (2019) Simulation of extreme heat events over the valencia coastal region: sensitivity to initial conditions and boundary layer parameterizations. *Atmospheric Research*, 218, 315–334.

Grams, C.M., Beerli, R., Pfenninger, S., Staffell, I. & Wernli, H. (2017) Balancing europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, 7, 557–562.

Grams, C.M., Magnusson, L. & Madonna, E. (2018) An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: a case study. *Quarterly Journal of the Royal Meteorological Society*, 144, 2577–2591.

Grazzini, F. & Vitart, F. (2015) Atmospheric predictability and rossby wave packets. *Quarterly Journal of the Royal Meteorological Society*, 141, 2793–2802.

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X. et al. (2022) Gefsv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Monthly Weather Review*, 150, 647–665.

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., Ferranti, L. & Prates, F. (2021) Evaluation of ecmwf forecasts, including the 2021 upgrade. https://www.ecmwf.int/node/20142

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.

Hughes, L., Hanna, E. & Fenwick, J. (2016) The silent killer: climate change and the health impacts of extreme heat.

Imran, H.M., Kala, J., Ng, A. & Muthukumaran, S. (2018) An evaluation of the performance of a wrf multi-physics ensemble for heatwave events over the city of melbourne in southeast australia. *Climate Dynamics*, 50, 2553–2586.

Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J.G., Ramos, A.M., Sousa, P.M. et al. (2022) Atmospheric blocking and weather extremes over the euro-atlantic sector–a review. *Weather and Climate Dynamics*, 3, 305–336.

Kueh, M.-T. & Lin, C.-Y. (2020) The 2018 summer heatwaves over northwestern europe and its extended-range prediction. *Scientific Reports*, 10, 19283.

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

Lee, S.H., Tippett, M.K. & Polvani, L.M. (2023) A new year-round weather regime classification for north america. *Journal of Climate*, 36, 7091–7108.

Lemburg, A. & Fink, A.H. (2022) Identifying causes of short-range forecast errors in maximum temperature during recent central european heatwaves using the ecmwf-ifs ensemble. *Weather and Forecasting*, 37, 1885–1902.

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.

Martius, O., Wehrli, K. & Rohrer, M. (2021) Local and remote atmospheric responses to soil moisture anomalies in australia. *Journal of Climate*, 34, 9115–9131.

Matsueda, M. & Palmer, T. (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.

Michel, C. & Rivière, G. (2011) The link between rossby wave breakings and weather regime transitions. *Journal of the Atmospheric Sciences*, 68, 1730–1748.

Miller, D.E., Wang, Z., Li, B., Harnos, D.S. & Ford, T. (2021) Skillful subseasonal prediction of us extreme warm days and standardized precipitation index in boreal summer. *Journal of Climate*, 34, 5887–5898.

Mücke, H.-G. & Litvinovitch, J.M. (2020) Heat extremes, public health impacts, and adaptation policy in germany. *International Journal of Environmental Research and Public Health*, 17, 7862.

Osman, M., Beerli, R., Büeler, D. & Grams, C.M. (2023) Multi-model assessment of sub-seasonal predictive skill for year-round atlantic-european weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 149, 2386–2408.

Pelly, J.L. & Hoskins, B.J. (2003) How well does the ecmwf ensemble prediction system predict blocking? *Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography*, 129, 1683–1702.

Pfahl, S. & Wernli, H. (2012) Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the northern hemisphere on (sub-) daily time scales. *Geophysical Research Letters*, 39.

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

Rodwell, M.J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C. et al. (2013) Characteristics of occasional poor medium-range weather forecasts for europe. *Bulletin of the American Meteorological Society*, 94, 1393–1405.

Rouges, E., Ferranti, L., Kantz, H. & Pappenberger, F. (2023) European heatwaves: link to large-scale circulation patterns and intraseasonal drivers. *International Journal of Climatology*, 43, 3189–3209.

Rousi, E., Fink, A.H., Andersen, L.S., Becker, F.N., Beobide-Arsuaga, G., Breil, M. et al. (2023) The extremely hot and dry 2018 summer in central and northern europe from a multi-faceted weather and climate perspective. *Natural Hazards and Earth System Sciences*, 23, 1699–1718.

Schaller, N., Sillmann, J., Anstey, J., Fischer, E.M., Grams, C.M. & Russo, S. (2018) Influence of blocking on northern european and western russian heatwaves in large climate model ensembles. *Environmental Research Letters*, 13, 054015.

Schulzweida, U. (2022) Cdo user guide. https://doi.org/10.5281/zenodo.7112925

Sheshadri, A., Borrus, M., Yoder, M. & Robinson, T. (2021) Midlatitude error growth in atmospheric gcms: the role of eddy growth rate. *Geophysical Research Letters*, 48, e2021GL096126.

Simmons, A. (1986) Numerical prediction: some results from operational forecasting at ecmwf. In: *Advances in geophysics*, Vol. 29. Elsevier, pp. 305–338.

Sousa, P.M., Barriopedro, D., García-Herrera, R., Ordóñez, C., Soares, P.M. & Trigo, R.M. (2020) Distinct influences of large-scale circulation and regional feedbacks in two exceptional 2019 european heatwaves. *Communications Earth & Environment*, 1, 1–13.

Spensberger, C., Madonna, E., Boettcher, M., Grams, C.M., Papritz, L., Quinting, J.F. et al. (2020) Dynamics of concurrent and sequential central european and scandinavian heatwaves. *Quarterly Journal of the Royal Meteorological Society*, 146, 2998–3013.

Steinfeld, D., Boettcher, M., Forbes, R. & Pfahl, S. (2020) The sensitivity of atmospheric blocking to upstream latent heating–numerical experiments. *Weather and Climate Dynamics*, 1, 405–426.

Teng, H., Branstator, G., Tawfik, A.B. & Callaghan, P. (2019) Circumglobal response to prescribed soil moisture over north america. *Journal of Climate*, 32, 4525–4545.

Trenberth, K.E. & Fasullo, J.T. (2012) Climate extremes and climate change: the russian heat wave and other climate extremes of 2010. *Journal of Geophysical Research: Atmospheres*, 117.

Vallis, G.K. (1983) On the predictability of quasi-geostrophic flow-the effects of beta and baroclinicity. *Journal of the Atmospheric Sciences*, 40, 10–27.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C. et al. (2017) The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, 98, 163–173.

Vitart, F., Balmaseda, M., Ferranti, L., Benedetti, A., Sarojini, B., Tietsche, S. et al. (2019) Extended-range prediction. ECMWF Technical Memorandum No., 854.

Vitart, F. & Robertson, A.W. (2018) The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *Npj Climate and Atmospheric Science*, 1, 3.

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

WMO. (2022) State of the Climate in Europe, 2021.

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

Zschenderlein, P., Fink, A.H., Pfahl, S. & Wernli, H. (2019) Processes determining heat waves across different european climates. *Quarterly Journal of the Royal Meteorological Society*, 145, 2973–2989.

Zsótér, E. (2006) Recent developments in extreme weather forecasting. *ECMWF Newsletter*, 107, 8–17.

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: Lemburg, A. & Fink, A.H. (2024) Investigating the medium-range predictability of European heatwave onsets in relation to weather regimes using ensemble reforecasts. *Quarterly Journal of the Royal Meteorological Society*, 1–32. Available from: https://doi.org/10.1002/qj.4801