Received: 27 January 2023 | Revised: 3 April 2023 | Accepted: 12 June 2023 | Published on: 24 July 2023

DOI: 10.1002/qj.4512

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Multi-model assessment of sub-seasonal predictive skill for year-round Atlantic–European weather regimes

Marisol Osman[1] | Remo Beerli[2] | Dominik Büeler[3] | Christian M. Grams[1]

[1]Institute of Meteorology and Climate Research, Department Troposphere Research (IMK–TRO), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

[2]AXPO Solutions, Baden, Switzerland

[3]Institute for Atmospheric and Climate Science, ETH Zürich, Zürich, Switzerland

Correspondence
Marisol Osman, Institute of Meteorology and Climate Research (IMK–TRO), Karlsruhe Institute of Technology, Postfach 3640, 76021 Karlsruhe, Germany.
Email: marisol.osman@kit.edu

Funding information
Helmholtz Association, Grant/Award Number: VH-NG-1243; AXPO Solutions AGN/A

Abstract

The prediction skill of sub-seasonal forecast models is evaluated for seven year-round weather regimes in the Atlantic–European region. Reforecasts based on models from three prediction centers are considered and verified against weather regimes obtained from ERA-Interim reanalysis. Results show that predicting weather regimes as a proxy for the large-scale circulation outperforms the prediction of raw geopotential height. Greenland blocking tends to have the longest year-round skill horizon for all three models, especially in winter. On the other hand, the skill is lowest for the European blocking regime for all three models, followed by the Scandinavian blocking regime. Furthermore, all models struggle to forecast flow situations that cannot be assigned to a weather regime (so-called no regime), in comparison with weather regimes. Related to this, variability in the occurrence of no regime, which is most frequent in the transition seasons, partly explains the predictability gap between transition seasons and winter and summer. We also show that models have difficulties in discriminating between related regimes. This can lead to misassignments in the predicted regime during flow situations in which related regimes manifest. Finally, we document the changes in skill between model versions, showing important improvements for the ECMWF and NCEP models. This study is the first multi-model assessment of year-round weather regimes in the Atlantic–European domain. It advances our understanding of the predictive skill for weather regimes, reveals strengths and weaknesses of each model, and thus increases our confidence in the forecasts and their usefulness for decision-making.

KEYWORDS

blocking, Europe, North Atlantic oscillation, windows of opportunity

1 | INTRODUCTION

Sub-seasonal prediction (∼10–30 days) has been a topic of great interest during the last decade. This is primarily driven by the need of multiple socio-economic sectors for skillful forecasts beyond the classic medium-range forecast horizon (White *et al.*, 2022), but also by the successful implementation of coordinated databases of sub-seasonal forecasts from leading modelling centers, such as the Subseasonal to Seasonal Prediction (S2S) project (Vitart *et al.*, 2017) or the Subseasonal Experiment (SubX) project (Pegion *et al.*, 2019). Atmospheric predictability on sub-seasonal time scales is related to both the initial conditions and slowly evolving climate variations in the Earth System, such as the ocean or sea-ice. Skillful sub-seasonal forecasts thus arise from the ability of forecast systems to capture persistent planetary-scale patterns that modulate weather conditions lasting longer than a week (Ferranti *et al.*, 2018). At the same time, the forecast question shifts from predicting the local weather at a specific time to the prediction of the larger-scale meteorological conditions aggregated over regions on weekly time scales.

A common topic of research within the S2S community is the quantification of forecast skill for recurring, persistent, and quasistationary large-scale patterns known as "weather regimes" (Matsueda and Palmer, 2018; Vigaud *et al.*, 2018; Büeler *et al.*, 2021; Cortesi *et al.*, 2021). In the Atlantic–European region, several definitions of weather regimes have been proposed (Michelangeli *et al.*, 1995; Cassou, 2008; Ferranti *et al.*, 2018; Falkena *et al.*, 2020), of which one of the most recent definitions captures the year-round large-scale flow variability (Grams *et al.*, 2017). The latter regime definition comprises three cyclonic regimes, in which a negative geopotential height anomaly associated with enhanced cyclonic activity dominates (Atlantic trough AT, Zonal regime ZO, Scandinavian trough ScTr), and four blocked regimes with a dominating positive geopotential height anomaly (Atlantic ridge AR, European blocking EuBL, Scandinavian blocking ScBL, Greenland blocking GL: see Figure 1). Recent works have shown that weather regimes are also closely related to surface weather variability (e.g. Beerli and Grams, 2019; Büeler *et al.*, 2020; Domeisen *et al.*, 2020). In addition, knowledge of the prevailing weather regime provides insight into the relative likelihood for anomalous weather to develop (e.g. Pasquier *et al.*, 2019; Spensberger *et al.*, 2020). Weather regimes can thus be a useful predictor in a range of applications (Zubiate *et al.*, 2017; Charlton-Perez *et al.*, 2019; van der Wiel *et al.*, 2019). A systematic assessment of model performance in forecasting weather regimes can provide better guidance for forecasters in a wide range of societal and economic sectors that are sensitive to weather and climate variability, for example, the energy or health sectors (Charlton-Perez *et al.*, 2019; Bloomfield *et al.*, 2020).

Work by Büeler *et al.* (2021) provides the first assessment of the skill of an S2S model in forecasting the aforementioned seven year-round Atlantic–European weather regimes. Using reforecast data from the European Centre for Medium-Range Weather Forecasts (ECMWF) model, they show that the predictability of weather regimes reaches 14 days on average (when using a stricter level of no skill than zero), and five days more in winter than in summer. In addition, they show that the Zonal and Greenland blocking regimes, which are closely related to the positive and negative phases of the North Atlantic Oscillation (NAO), respectively, have the longest skill horizon, while the skill for the European blocking regime is the poorest. Furthermore, they find that the wintertime skill horizon increases by 5 days after a strong stratospheric polar vortex, and that an active Madden–Julian Oscillation in phase 4 or 7 provides additional skill for some weather regimes. To our knowledge, the only study that has done an assessment of multiple S2S models for wintertime weather regimes is that by Bloomfield *et al.* (2021) for the ECMWF and National Centers for Environmental Prediction (NCEP) CFSv2 model. However, a systematic analysis for the year-round regimes including models by other centers is missing.

For medium-range weather forecasts, a systematic assessment of forecasts over a historical period is very useful, since most of their predictability comes from the initial conditions. However, for longer lead times some flow configurations can be more skillful than others (so-called "windows of opportunity") and therefore different tools that target these situations are needed (Mariotti *et al.*, 2020). In particular, the works of Cortesi *et al.* (2021) and Matsueda and Palmer (2018) have shown strong annual and year-to-year variability in the forecast skill of weather regimes. The flow-dependent verification documented in Büeler *et al.* (2021) provides a starting point to identify such windows of opportunity for sub-seasonal forecasts, but further studies are needed to understand how differences in flow manifest in the daily and year-to-year variability in skill.

Since the original implementation of the S2S database (Vitart, 2014), the models participating in the project have evolved substantially, and new versions are now available to the community. By analyzing the improvements in skill between model versions, it is possible to understand the sources of such improvements. For instance, Vitart (2014) shows that, for the ECMWF model, the sub-seasonal skill of reforecasts has improved mainly due to changes in model physics, whereas changes in model resolution have had very little effect. However, the impact of the initial conditions is not assessed in that study. Many other articles

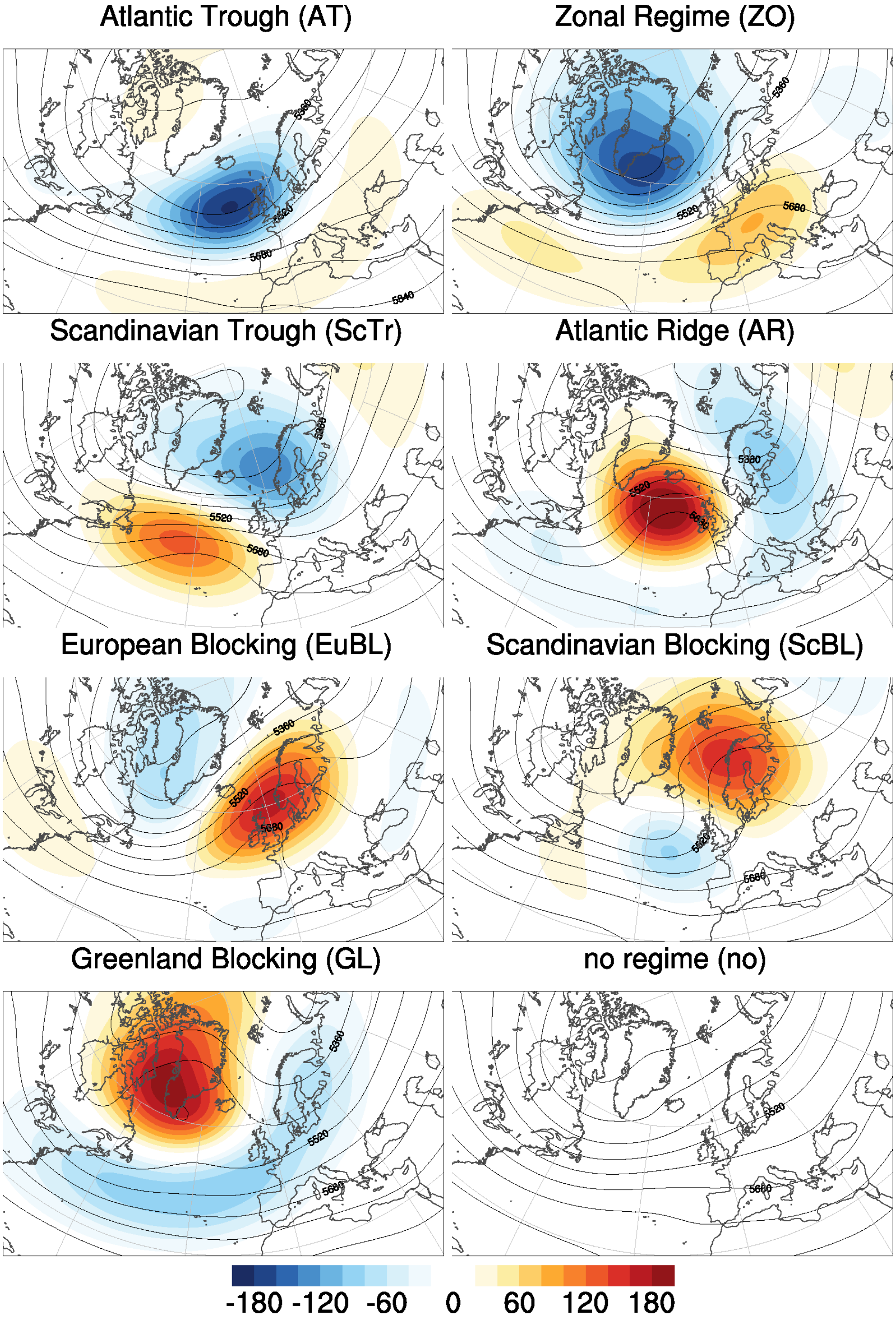

FIGURE 1 Mean 500-hPa geopotential height (contours; geopotential meters (gpm)) and corresponding anomalies (shading; gpm) of the seven year-round Atlantic–European weather regimes, defined based on ERA-Interim data between 1979 and 2015 and the “no regime” category (times at which none of the seven regimes is observed). [Colour figure can be viewed at wileyonlinelibrary.com]

also document the improvements in forecast skill when a new model version is developed (see for instance Maclachlan *et al*., 2015; Zhou *et al*., 2022). However, to our knowledge there is no assessment of the changes in sub-seasonal skill for weather regimes with model versions. Additional gaps in the literature include the absence of a multi-model assessments of year-round weather regimes, and a lack of studies showing how skill varies on daily and interannual time scales and across model versions.

The objective of this study is to address gaps in knowledge regarding the representation and assessment of sub-seasonal forecast skill for year-round weather regimes depicted by different operational sub-seasonal forecasting systems. We put an emphasis on understanding the daily as well as year-to-year variability of forecast skill. In addition, we document the differences in skill between different versions of the S2S models to identify progress and challenges in improving skill. The article is organized as follows: Section 2 describes the models and methods employed while in Section 3 we describe the main findings. The conclusions are presented in Section 4.

TABLE 1 Main characteristics of the S2S models used in this study.

Model	Version	Resolution	Reforecast type	Reforecast period	Reforecast frequency	Reforecast length (days)	Ensemble members
NCEP	GEFS v12	C384L64 ~25 km	Fixed	2000–2019	Every Wednesday	35	11
NCEP	CFS v2	T126 L64 ~100 km	Fixed	1999–2010	Daily	44	4
ECMWF	CY46R1 and CY47R1	Tco639 L91 ~16 km up to day 15 and Tco319 ~32 km after day 15	On the fly	1999–2019	2/week	46	11
ECMWF	CY43R1, CY43R3 and CY45R1	Tco639 L91 ~16 km up to day 15 and Tco319 ~32 km after day 15	On the fly	1997–2017	2/week	46	11
UKMO	GloSea6 (HadGEM3 GC3.2)	N216 L85 ~60 km	On the fly	1993–2015	4/month	60	7
UKMO	GloSea5 (HadGEM3 GC2.0)	N216 L85 ~60 km	On the fly	1993–2015	4/month	60	7

Note: For a definition of reforecasts on the fly and fixed reforecasts, the reader can refer to https://confluence.ecmwf.int/display/CKB/Seasonal+forecasts+and+the+Copernicus+Climate+Change+Service.

2 | DATA AND METHODOLOGY

Models and reference dataset

Models from ECMWF, NCEP, and the UK Met Office (UKMO) are analyzed in this study. These models are selected because they are often used by societal and economic sectors in their forecast-related activities, for instance in the estimation of energy production by renewable sources (wind power and solar) and energy demand. The ECMWF model versions are CY46R1 and CY47R1, implemented on June 11, 2019 and June 30, 2020, respectively. The NCEP model version is the Global Ensemble Forecast System (GEFS) v12, implemented in September 2020. The Met Office Global Seasonal Forecasting System version 6 (GloSea6-HadGEM3 GC3.2), implemented on February 2, 2021, is the UKMO model used in this study. Further details, such as model resolution, reforecast type, and ensemble size, can be found in Table 1. Note that the table also includes older versions of these models, which will be used in Section 3.5. The main assessment is performed with the above-mentioned model versions. While the ECMWF and UKMO model data were obtained from the S2S project database, the NCEP model data were obtained from the NOAA AWS repository.[1] Although each model has a different reforecast period, we select the period 2000–2015, available to all systems, to avoid differences associated with interannual variability in skill. The ECMWF and UKMO data were retrieved with a 1° horizontal resolution, while GEFSv12 was retrieved with a 0.25° resolution for the first 10 days and 0.5° resolution from day 11 onward and then interpolated to a 1° horizontal resolution. The interpolation from native model resolution to final resolutions might have some influence on skill differences. However, this is not quantified in this study. We use ERA-Interim (Dee *et al*., 2011), also retrieved with a 1° horizontal resolution, as a reference dataset to compare the forecasts against. This reanalysis was selected to make the results comparable with those from (Büeler *et al*., 2021). However, we anticipate that our results would remain comparable if we had used the newer ECMWF reanalysis, ERA5. The use of ERA-Interim may penalize the skill seen in the NCEP and UKMO models in comparison with that obtained with ECMWF, as a similar model version is used to produce the latter model and the reanalysis. Nonetheless, the results are not expected to differ significantly if other reanalysis datasets, such as the Climate Forecast System Reanalysis (CFSR) or Modern-Era Retrospective analysis for Research and Applications (MERRA), are used instead of ERA-Interim, since these reanalyses have comparable performance, as demonstrated by Long *et al.* (2017) for the climatology and interannual variability of two dynamical variables.

Computation of weather regime forecasts

The approach adopted in this study builds upon the work performed by Büeler *et al.* (2021). Here we briefly describe the procedure to obtain the weather regime (WR) forecasts, but the reader can refer to Büeler *et al.* (2021) for further details. The climatological mean weather

[1] https://registry.opendata.aws/noaa-gefs-reforecast

regime patterns are defined based on the ERA-Interim period (1979–2015) 6-hr 500-hPa geopotential height anomalies with respect to the corresponding 90-day running mean calendar date climatologies (i.e., +/−45 days centered around each 6-hr time step) within the North Atlantic–European domain from 80°W–40°E and 30°N–90°N. The anomalies are first filtered with a 10-day low-pass filter and then normalized seasonally. This normalization is done by dividing the anomalies by a calendar-day-dependent latitude-weighted average (over the domain of study) of the 31-day running mean temporal standard deviation over all anomalies between 1979 and 2015. In contrast to Büeler *et al.* (2021), we use the 10-day low-pass filter instead of a 5-day low-pass filter to remain close to the original regime definition in Grams *et al.* (2017).[2] An empirical orthogonal function (EOF) analysis is applied to the filtered and normalized anomalies and then *k*-means clustering is applied to the anomalies in the phase space spanned by the first seven EOFs (explaining approximately 70% of the variance), which yields an optimal number of seven cluster means representing the seven weather regimes. These are the aforementioned three cyclonic and four blocked regimes (see Figure 1).

Following Büeler *et al.* (2021), we also identify the weather regimes in the reforecast data. In this study, however, instead of computing the weather regime forecasts for both the raw forecasts and the bias-corrected (i.e., calibrated) forecasts, we only compute the weather regime forecasts for the latter. This is done by computing the geopotential height anomalies with respect to the lead-time dependent 90-day running mean model climatology (based on the entire hindcast period of the model). Likewise, the normalization of the low-pass-filtered anomalies is done with respect to model data in the entire hindcast period. Then, we project the normalized anomalies onto the seven cluster mean geopotential height anomalies obtained from ERA-Interim data for the 1979–2015 period (cf. above), as in Büeler *et al.* (2021):

$$P_{\mathrm{wr}}(t) = \frac{1}{\sum \cos\phi} \sum_{(\phi,\lambda)} \Phi(\lambda,\phi,t)\Phi_{\mathrm{wr}}(\lambda,\phi)\cos\phi.$$

Here $P_{\mathrm{wr}}(t)$ is a scalar measure for the spatial correlation of the instantaneous anomaly field $\Phi(\lambda,\phi,t)$ at lead time t (at each grid point with latitude λ and longitude ϕ within the EOF domain) with the cluster mean anomaly field $\Phi_{\mathrm{wr}}(\lambda,\phi)$ for the regime “wr”. The nondimensional regime index $I_{\mathrm{wr}}(t)$ for each regime and forecast is based on anomalies of the projections $P_{\mathrm{wr}}(t)$ with respect to the climatological mean projection $\overline{P_{\mathrm{wr}}}$ and the climatological standard deviation of the projection based on the calibrated forecasts for the hindcast period as follows:

$$I_{\mathrm{wr}}(t) = \frac{P_{\mathrm{wr}}(t) - \overline{P_{\mathrm{wr}}}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}[P_{\mathrm{wr}}(t) - \overline{P_{\mathrm{wr}}}]^2}}. \qquad (1)$$

To determine the active weather regime at each lead time step t, we apply the same life-cycle criteria as in Grams *et al.* (2017): a regime is active if its $I_{\mathrm{wr}}(t)$ is maximum among all seven $I_{\mathrm{wr}}(t)$ and equal to or above 1.0 for five consecutive days or longer. The time steps in which none of the seven regimes fulfills these criteria are categorized as “no regime”. Note that “no regime” periods might include time steps of shallow pressure distribution as well as episodes of regime transitions, among other cases.

To verify the forecasts, we identify the weather regime evolution in ERA-Interim for the corresponding forecast lead times. This means that we treat ERA-Interim as an additional ensemble member and we apply to it the same steps as above, except that the normalized geopotential height anomalies and the weather regime indices are based on geopotential height and projection data from ERA-Interim over the hindcast period of each model, respectively. The year-round climatological frequencies of the seven regimes and the no regime category in ERA-Interim for the period 2000–2015 are shown in Figure 2. The main difference between this distribution and the one shown in Büeler *et al.* (2021) for the 1997–2017 period is the higher frequency of ScBL in summer, which is nearly 10% larger for 2000–2015 than for 1997–2017. This higher frequency of ScBL in summer is at the expense of the no regime category, the frequency of which is highest in April and May.

In the Supplementary Material, Figure S1 shows an example of the forecast obtained for the NCEP model initialized on January 6, 2010. Figure S1a shows the evolution of $I_{\mathrm{wr}}(t)$ for the seven regimes in the ensemble and the corresponding ERA-Interim value for the forecast days. It gives an overview of the evolution of the dominating and suppressed regimes with lead time, as well as the associated ensemble spread. Figure S1b shows the ensemble forecast probability for a certain regime to be active as a function of lead time and the ERA-Interim weather regime during that period. This categorical forecast will serve as the basis for most of the analysis conducted in this study. For each model, the verification is done until a lead time of $T - 15$ days, with T being the maximum number of lead times of each forecast. This is due to the loss of data at the end of each forecast, which results from both the low-pass filtering (10 days) and a convergence to the “no regime” category due to the life-cycle persistence criterion (5 days).

[2] Note that, in the original work of Grams *et al.* (2017), a latitudinal weighting had not been applied when computing the spatial mean for the normalization weight.

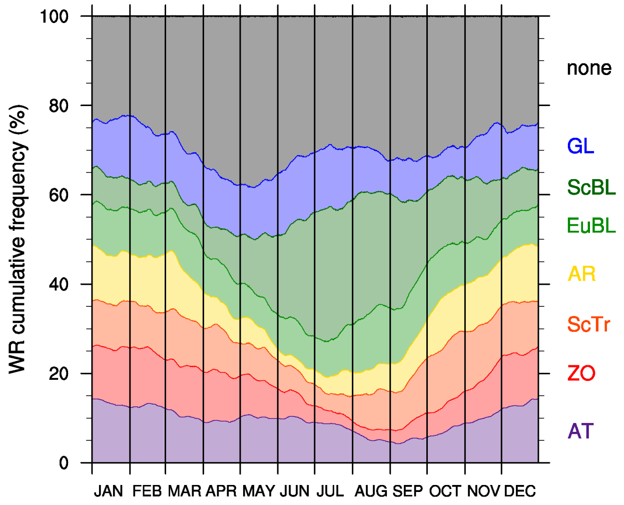

FIGURE 2 91-day running mean calendar day climatological cumulative relative frequency of weather regime life cycles in ERA-Interim defined over the period of study 2000–2015. [Colour figure can be viewed at wileyonlinelibrary.com]

2.1 | Verification

As in Büeler *et al.* (2021), the fair Brier Score (BS, Ferro *et al.*, 2008) is used in this study to verify the categorical ensemble probabilistic forecasts of weather regime life-cycle occurrence at each lead time:

$$BS = \frac{1}{N}\sum_{k=1}^{N}\sum_{wr\in WR}\left((y_k^{\mathrm{wr}} - o_k^{\mathrm{wr}})^2 - \frac{my_k^{\mathrm{wr}}(m - my_k^{\mathrm{wr}})}{m^2(m-1)}\right), \tag{2}$$

where y_k^{wr} is the forecast probability (between 0 and 1) for regime *wr* of forecast *k*, o_k^{wr} is the observed dichotomous counterpart (0 or 1), *m* is the ensemble member size, and *N* is the total number of forecasts. *wr* in the sum can be all regimes together (i.e. *WR* ={AT, ZO, ScTr,AR, EuBL, ScBL, GL, no}) to compute the multicategory BS or an individual regime (i.e., $WR \in$ (AT, ZO, ScTr, AR, EuBL, ScBL, GL, or no) to compute the single-category BS. The fair BS is the classic BS (Brier, 1950; Wilks, 2019), but it also includes a correction term to account for the relatively small ensemble member size. When dealing with multiple categories, the multicategory BS provides a more comprehensive assessment of the overall forecast accuracy, since it evaluates performance across categories.

Finally, we compute the fair Brier skill score (BSS; Wilks, 2019) to relate the fair BS of the numerical model forecast to the BS of a climatological reference forecast (BSref):

$$BSS = 1 - \frac{BS}{BS_{\mathrm{ref}}}. \tag{3}$$

As a reference forecast, we use the 90-day running mean climatological calendar day regime frequency in ERA-Interim at each lead time step. To test the robustness of the BSS, we apply a bootstrapping procedure by randomly resampling (with replacement) 10^4 times a set of forecasts of the same size as the evaluated set of forecasts, and compute the skill score for each of these random samples. We then take the 5th and 95th percentiles derived from these skill score distributions to define the confidence interval for each skill score. A bootstrapping approach is also used to determine whether the biases in the climatological regime occurrence frequencies or transition frequencies (defined as the difference between the regime or transition occurrence frequency in the forecast and in ERA-Interim) are significant. If the bias of the regime or transition occurrence falls outside the 5th or 95th percentiles, the bias is defined to be significant at the 10% level.

In addition to the probabilistic forecasts, we also evaluate the categorical deterministic forecast of the weather regime life cycle using the Heidke Skill Score (HSS; Heidke, 2017):

$$HSS = \frac{PC - E}{1 - E}, \tag{4}$$

where *PC* is the fraction of correct forecasts, defined as

$$PC = \frac{1}{N}\sum_{\mathrm{wr}\in WR} n\left(F_{\mathrm{wr}}, O_{\mathrm{wr}}\right),$$

where $n\left(F_{\mathrm{wr}}, O_{\mathrm{wr}}\right)$ is the number or forecast hits for the weather regime *wr* and *N* is the total number of forecasts. *E* is the fraction of correct forecasts due to random chance, defined as

$$E = \frac{1}{N^2}\sum_{wr\in WR} N\left(F_{\mathrm{wr}}\right) N\left(O_{\mathrm{wr}}\right),$$

where $N\left(F_{\mathrm{wr}}\right)$ and $N\left(O_{\mathrm{wr}}\right)$ are the number of forecasts and observations for the weather regimes *wr*, respectively. The Heidke Score (HS, numerator of Equation 4) measures the accuracy of the forecast in predicting the correct category, in relation to the forecast that would be correct due to random chance. The HSS is the HS normalized by the perfect forecast (denominator in Equation 4) in relation to the forecasts that would be correct due to random chance. To increase the robustness of the score, we use each member of the ensemble as an independent forecast and compare it with the corresponding ERA-Interim value. Both the BSS and HSS account for the reliability and the resolution of forecasts, although the BSS is used for probabilistic forecasts while the HSS is used for deterministic forecasts. While both scores are useful, a probabilistic approach is generally preferred over a deterministic approach. Unlike the BSS, which penalizes the errors in the probabilistic forecasts, the HSS focuses on the correct forecasts. In addition, the BSS uses the climatological probabilistic distribution as a reference forecast, whereas in the HSS the reference forecast is the random occurrence of hits. These differences in the definition of the scores can lead to differences in the relative performance between models when using both skill scores.

The root-mean-square error skill score (RMSESS) was used to verify the reforecast for the geopotential height field over the domain of study at each lead time:

$$RMSESS = 1 - \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(F_k - O_k - \overline{F} + \overline{O}\right)^2}}{\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(O_k - \overline{O}\right)^2}}. \quad (5)$$

Here, N is the total number of forecasts and F and O are the forecast and reference data values, respectively, whereas $\overline{F}$ and $\overline{O}$ are the mean of each field.

3 | RESULTS

3.1 | Assessment of WR frequency biases

We begin with the investigation of the systematic 500-hPa geopotential height model biases, which are removed before computing the weather regime forecasts. From this subsection until Section 3.5 we always refer to the new version of the models, that is, NCEP GEFSv12, ECMWF CY46R1 and CY47R1, and UKMO GloSea6, respectively. Figure 3 shows lead-time dependent 500-hPa geopotential height biases of each model climatology with respect to the ERA-Interim climatology on a calendar day centered in each season. We only show biases for lead times of 10 and 20 days, because biases saturate at sub-seasonal lead times (see Büeler *et al.*, 2021). In general, the NCEP model shows the smallest biases and the smallest seasonal differences in the distribution of those biases. In the Atlantic–European region, the most striking feature is the large positive biases in summer for the ECMWF model, a feature already reported in Büeler *et al.* (2021) for an older version of the model. This positive bias could be partially related to the model's positive bias in the summertime mid-troposphere (850–500hPa) temperature (Magnusson *et al.*, 2022). Both UKMO and NCEP models also present positive biases in summer in the North Atlantic region, which suggests that some physical processes that drive the variability there are not well represented in models yet. Biases in winter are smallest for the three models and are mainly restricted to Greenland and the east coast of North America. In spring, negative biases are observed over Europe for the three systems, while positive biases are evident for the NCEP model over the Atlantic. Finally, in autumn ECMWF and UKMO show positive biases over the eastern North Atlantic, while biases for NCEP are similar to those for spring. Outside the Atlantic–European region, the ECMWF presents large positive biases in summer and autumn over the Pacific, whereas the UKMO model presents large negative biases in autumn over Asia and the Pacific.

To begin the assessment of the weather regime forecasts, we focus on analyzing the main weather regime characteristics, such as their frequency of occurrence, duration, and number. We only select one initialization per week (i.e., Thursdays) from the ECMWF reforecasts for a fair comparison between models. However, results are very similar if all initial conditions are taken into account (not shown). Figure 4 shows the seasonal life-cycle frequency bias with respect to ERA-Interim, as a function of the lead time for each model and for each regime. Removing the mean 500-hPa geopotential height biases in the forecast makes frequency biases very small (the reader can refer to figure 5 in Büeler *et al.* (2021) for a comparison between biased and bias-corrected regime frequency), ranging mostly between −5% and 5%. While the NCEP model only shows a significant bias for EuBL in summer, the ECMWF and UKMO models show a positive bias for the no regime category in winter and a negative bias for ScBL in summer for lead times beyond 10 days (an issue documented for an older version of ECMWF model in Büeler *et al.*, 2021).

The biases in the life-cycle frequency can be related to deficiencies in the forecasts of the number and duration of the life cycle of each weather regime, as well as biases in the transition from one regime into another. To explore these relationships, Figure S2 in the Supplementary Material shows the seasonal duration and total number of regime life-cycle events for each model and for ERA-Interim (throughout all lead times). For the no regime category, we have used a pseudo life-cycle definition (at least 5 days of no regime) for a fair comparison with the regimes. It is important to note that similar frequency biases for two regimes in the same season do not translate directly to a similar bias in the number and/or length of regimes, since the frequency biases account for the per-day frequency of a certain regime, and also because the frequencies are presented in percentages. Figure S3 presents the seasonal frequencies of transitions between these regime life cycles in ERA-Interim (within a time frame of at most 4 days; shading; adding up to 100% along the horizontal axis) and the associated significant biases in the forecasts of each model (numbers). The negative frequency biases for the ScBL in summer in the ECMWF and UKMO models (Figure 4h,i) can be partly explained by the smaller number of life cycles (Figure S2h,i) as well as fewer transitions into this weather

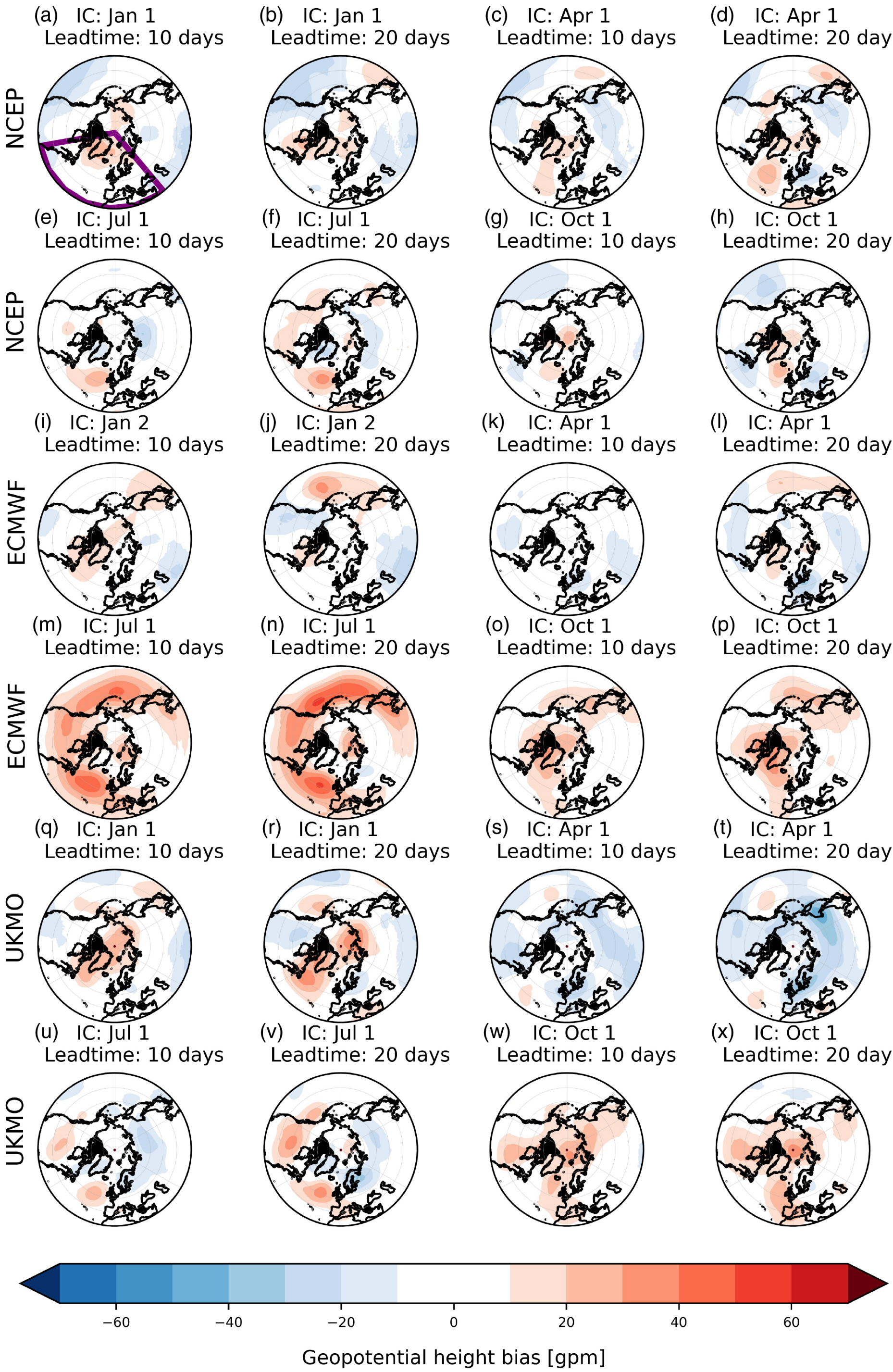
(a) IC: Jan 1 Leadtime: 10 days
(b) IC: Jan 1 Leadtime: 20 days
(c) IC: Apr 1 Leadtime: 10 days
(d) IC: Apr 1 Leadtime: 20 days
NCEP
(e) IC: Jul 1 Leadtime: 10 days
(f) IC: Jul 1 Leadtime: 20 days
(g) IC: Oct 1 Leadtime: 10 days
(h) IC: Oct 1 Leadtime: 20 days
NCEP
(i) IC: Jan 2 Leadtime: 10 days
(j) IC: Jan 2 Leadtime: 20 days
(k) IC: Apr 1 Leadtime: 10 days
(l) IC: Apr 1 Leadtime: 20 days
ECMWF
(m) IC: Jul 1 Leadtime: 10 days
(n) IC: Jul 1 Leadtime: 20 days
(o) IC: Oct 1 Leadtime: 10 days
(p) IC: Oct 1 Leadtime: 20 days
ECMWF
(q) IC: Jan 1 Leadtime: 10 days
(r) IC: Jan 1 Leadtime: 20 days
(s) IC: Apr 1 Leadtime: 10 days
(t) IC: Apr 1 Leadtime: 20 days
UKMO
(u) IC: Jul 1 Leadtime: 10 days
(v) IC: Jul 1 Leadtime: 20 days
(w) IC: Oct 1 Leadtime: 10 days
(x) IC: Oct 1 Leadtime: 20 days
UKMO
−60
−40
−20
0
20
40
60
Geopotential height bias [gpm]

FIGURE 3 500-hPa geopotential height model climatology biases (gpm) in the Northern Hemisphere for (a–h) NCEP, (i–p) ECMWF, and (q–x) UKMO forecasts initialized on (a,b,q,r) January 1, (i,j) January 2, (c,d,k,l,s,t) April 1, (e,f,m,n,u,v) July 1, and (g,h,o,p,w,x) October 1 at (a,c,e,g,i,k,m,o,q,s,u,w) 10 and (b,d,f,h,j,l,n,p,r,t,v,x) 20 days. The purple box indicates the EOF domain in which the weather regimes are defined. [Colour figure can be viewed at wileyonlinelibrary.com]

regime (Figure S3h,i), a finding that was also reported for ECMWF by Büeler *et al.* (2021). In particular, biases in the transition from ScTr to ScBL (a common transition in summer) are more than 10% rarer in both models, while the same is observed for the transition from AT to ScBL in the UKMO model. In addition, both models underestimate the relative number of transitions from no regime to ScBL, another common transition in summer. On the other hand, the positive bias in the number of summer EuBL life cycles in NCEP (Figure S2g) can partially explain the positive bias in the life-cycle frequency for this model and season (Figure 4g).

3.2 | Assessment of WR forecast skill

We now focus on the evaluation of forecast skill by all models. From this section onward, we use all the ECMWF initializations to make our results less sensitive to the sample size, although the conclusions do not change qualitatively

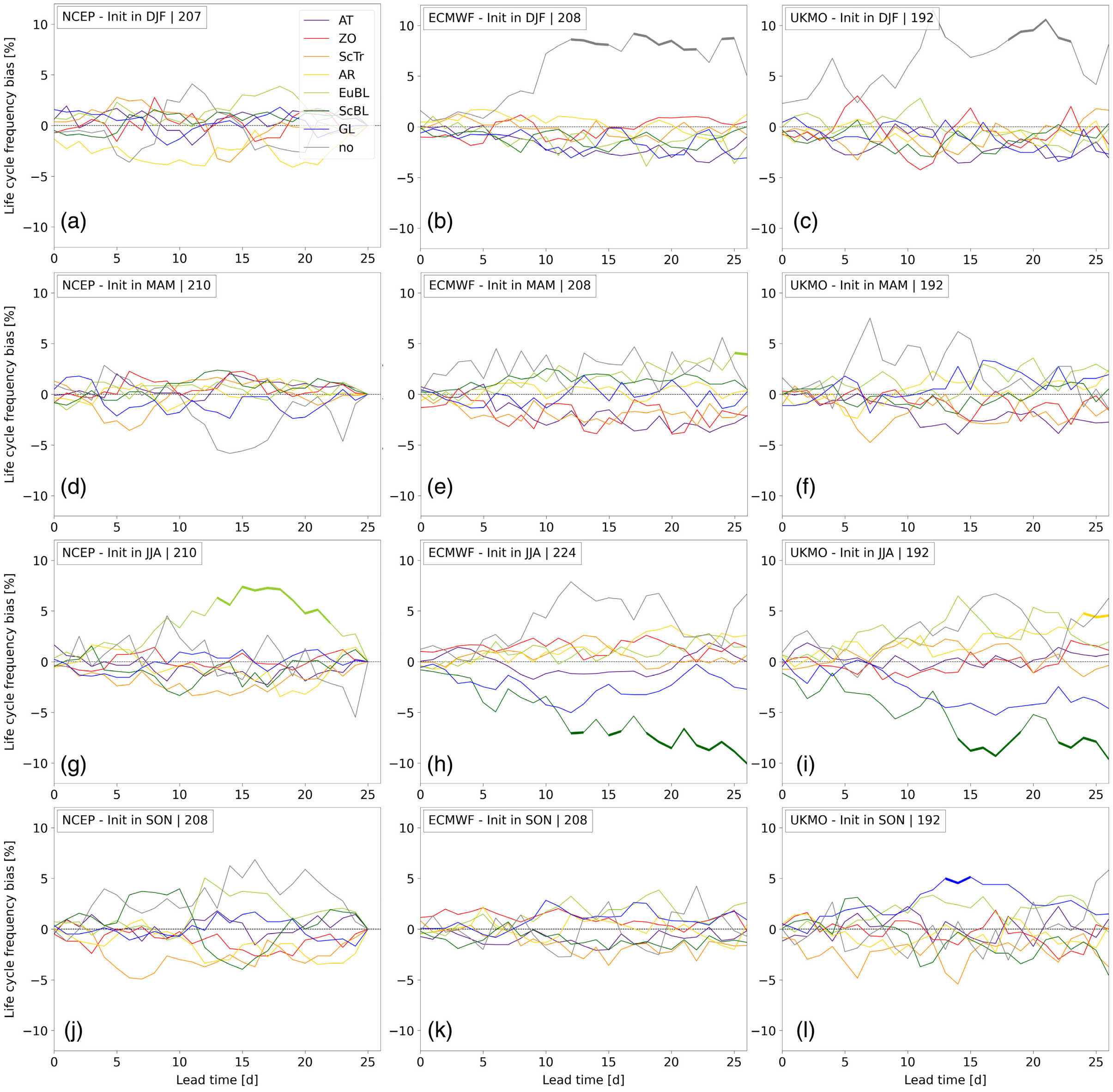

FIGURE 4 Seasonal weather regime life-cycle frequency biases (%; *y*-axis) in (left) NCEP, (middle) ECMWF, and (right) UKMO models with respect to ERA-Interim as a function of lead time (days; *x*-axis). Bold lines indicate significant biases. The seasons and the corresponding available numbers of forecasts are indicated in the boxes. [Colour figure can be viewed at wileyonlinelibrary.com]

compared with those using only one initial condition per week. Figure 5 shows the year-round multicategory (all regimes) life-cycle BSS (Equation 3) for the three models. Due to the definition of the life-cycle persistence criterion (5 days), convergence to “no regime” for lead times beyond 20 days in NCEP model makes BSS values go upward. Although values above zero mean that models are better than climatology, in this study we follow Büeler *et al.* (2021) and define a more rigorous level of BSS equal to 0.1 as a reference to compare skill horizons in the remainder of the study. This 0.1 level is an arbitrary (but reasonable) value. Therefore, the skill horizon for the ECMWF model is 14 days, followed by NCEP at near 13 days and the UKMO model at 10 days. The relatively low BSS for UKMO could be due to deficiencies in the representation of processes by this model, but might also be related to the smaller ensemble size in comparison with the other models. While the BSS accounts for different ensemble sizes, it does so only when a category presents a nonzero forecast probability. Given the number of categories to predict, the small ensemble size in UKMO may not be able to map the uncertainty correctly, resulting in a decrease in skill when the uncertainty is higher (i.e. beyond the first few days). Confirming this hypothesis would require us to recompute the regime forecast using the same ensemble size for the three models, which goes

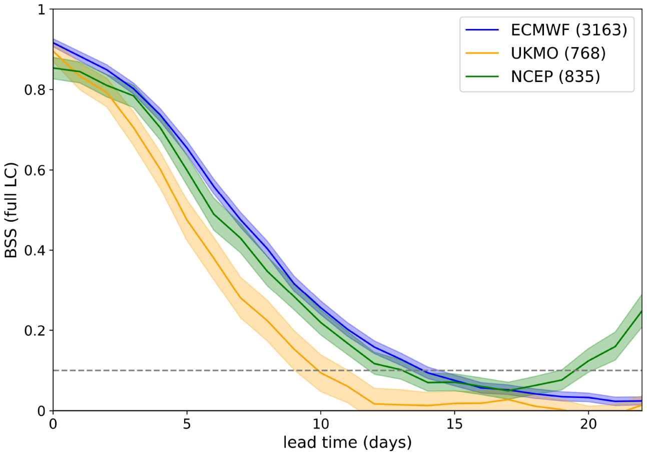

FIGURE 5 Year-round multicategory BSS for all weather regime life cycles as a function of lead time for ECMWF (blue), UKMO (yellow), and NCEP (green) models. The BSS for the life cycle is computed including the "no regime" category. Shading shows the range between the 5th and 95th percentile obtained through a bootstrap procedure. The gray dotted line denotes the 0.1 BSS level. Notice that, due to the convergence to the "no regime" category as a result of the life-cycle persistence criterion (5 days), the BSS values for the NCEP model beyond lead time 20 days go upward. [Colour figure can be viewed at wileyonlinelibrary.com]

beyond the scope of this work. Evaluating the skill for the maximum regime index (i.e. without applying the life-cycle criteria; Figure S4) does not show major differences in the skill horizon nor the relative performance between models, other than for lead times 0–2 days, when the performance of the NCEP model is best using the maximum $I_{\mathrm{wr}}(t)$ (Equation 1) but worst using the full life-cycle definition (i.e., $I_{\mathrm{wr}}(t)$ above 1.0 during at least 5 days). Figure S4 also shows the root-mean-squared error skill score (RMSESS: Equation 5) averaged over the domain in which the weather regimes are computed. We find that the BSS values for both the full LC and the maximum $I_{\mathrm{wr}}(t)$ BSS are higher than the RMSESS for long lead times, although differences are minor. Therefore, predicting weather regimes as a proxy for large-scale circulation outperforms using the full 500-hPa geopotential height field.

Figure 6 shows for each season the multicategory life-cycle BSS for all regimes of the three models in comparison (in order to evaluate the relative performance of models for different times of the year). The smaller sample size in the NCEP and UKMO makes the results less robust than for ECMWF. The skill of the ECMWF model is slightly higher than that of the NCEP model in all seasons but summer, when the skill of both models is virtually identical. The skill of the UKMO model is comparable with the skill of the other models only for short lead times (less than 5 days) in all seasons but summer, when the skill in the UKMO model already decreases sharply 2 days after initialization. For longer lead times, the UKMO skill is significantly lower that for the other two models in all seasons. Figure S5 shows, for each model, the multicategory life-cycle BSS for the four seasons. The skill in winter is highest for all three models, and in the ECMWF and NCEP models at most lead times it is significantly higher than for the remaining seasons. This is in agreement with other studies (Son *et al.*, 2020; Büeler *et al.*, 2021; Cortesi *et al.*, 2021). For these models, the skill horizon for winter of 17 and 15 days in ECMWF and NCEP, respectively, is 3–4 days longer than for the other seasons, which have very little difference between each other.

Figure 7 shows for each weather regime the year-round BSS for the three models. Within the cyclonic regimes, the AT is forecast with similar skill among the three models whereas for the ZO and ScTr the skill of the ECMWF and NCEP models is very similar and higher than that of the UKMO model. On the other hand, for the blocked regimes the skill of the ECMWF and NCEP models is similar for ScBL and AR, while the ECMWF model is slightly better than the NCEP model for EuBL and GL. The skill for the UKMO model is somewhat lower than those for ECMWF and NCEP (although differences are not statistically significant). For the no regime category, the ECMWF skill is the best. The year-round single-category skill for each weather regime in each model is shown in Figure S6. The no regime category has low skill in comparison with the rest of the regimes, which confirms the difficulty that models face in forecasting flow situations that do not fit into one of the distinct patterns. Looking at each weather regime individually, the ZO and GL have the longest skill horizon of almost 18 days (considering the 0.1 level). Conversely, the EuBL, and in the case of ECMWF also the ScBL, have the shortest skill horizon of around 13 days in the ECMWF model, 11 in the NCEP model, and near 8 days in the UKMO model. The better performance for ZO and GL, which are closely related to the positive and negative phases of the NAO, compared with blocking in the European sector has also been documented in previous studies, with other versions of these models either using the set of seven regimes (Büeler *et al.*, 2021) or the more traditional set of four regimes (Ferranti *et al.*, 2018; Matsueda and Palmer, 2018; Cortesi *et al.*, 2021). Finally, Appendix A presents a summary of the skill for individual weather regimes for each season and model. This type of comparison can help to identify strengths and weaknesses in each model.

Overall, the analysis done in this section has shown multiple differences in skill, not only between models but also in the relative performance in the individual seasons. This can be useful for both model developers and forecasters.

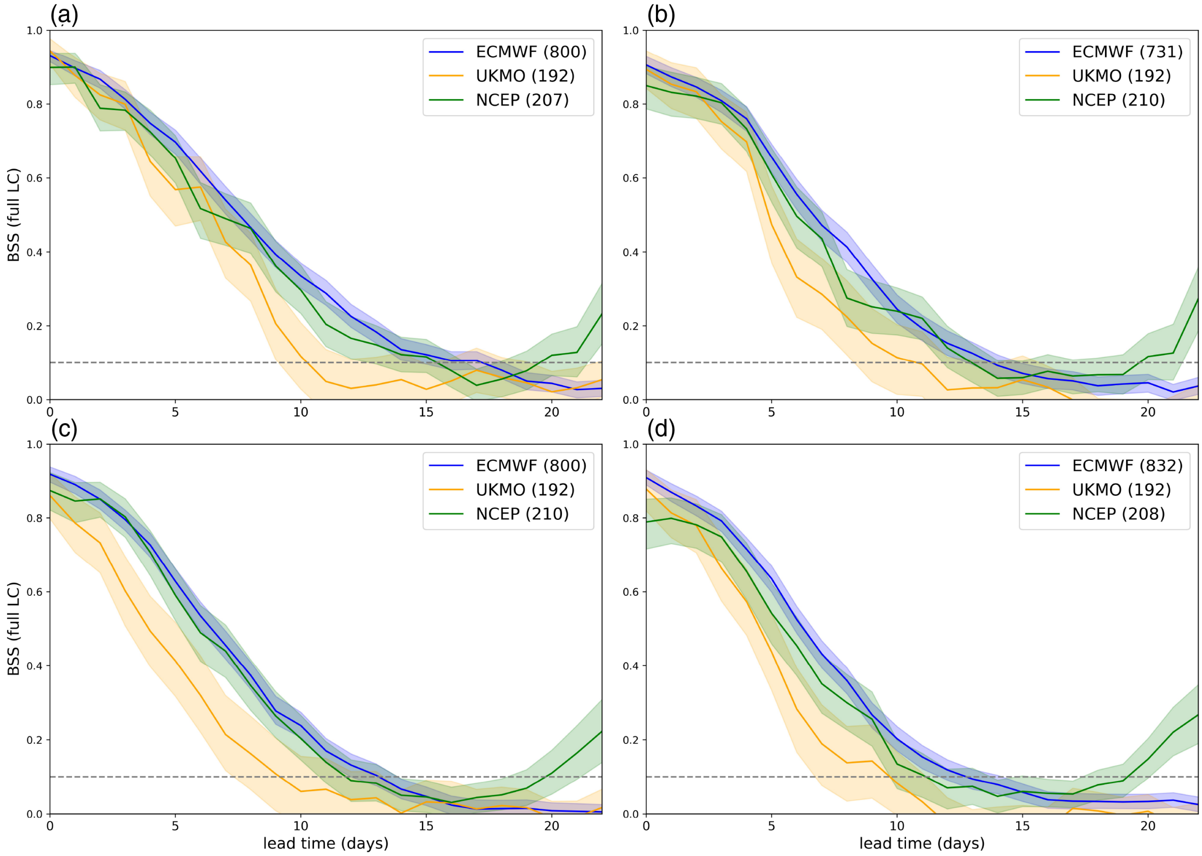

FIGURE 6 Multicategory BSS for life cycle for all weather regimes for ECMWF (blue), UKMO (orange), and NCEP (green) models initialized in (a) DJF, (b) MAM, (c) JJA, and (d)SON. The BSS for the life cycle is computed including the "no regime" category. Shading shows the range between the 5th and 95th percentile obtained through a bootstrap procedure. The gray dotted line denoted the 0.1 BSS level. [Colour figure can be viewed at wileyonlinelibrary.com]

3.3 | Daily to interannual variability of skill

We now turn our attention to investigating how the differences in the forecast skill for each weather regime and season shape the daily to interannual variability in skill. In this section the skill is evaluated through the HSS (Equation 4) and, to increase the robustness of the score, we take each member of the ensemble as an independent forecast and compare it with the corresponding ERA-Interim value. Figure 8 shows the weather regime life cycle HSS for a 91-day moving window centered on each 365-day calendar date, as a function of the lead time. As was mentioned in Section 2, the differences in the formulation of the BSS and HSS can lead to differences in the relative performance between models when comparing the current results with those from the previous section. For short lead times (within the weather time scale, 1–5 days) the HSS values still show some dependence on the day of the year, being slightly smaller between May and June for the three models and between October and November for the NCEP and ECMWF models. For long lead times, January and February present the highest HSS values for all models. In those months, the HSS values above 0.1 remain up to day 20 for NCEP (note that, beyond lead time 20, the life-cycle definition is not valid for this model), 28 for ECMWF, and more than 35 for UKMO. In contrast, the HSS values decrease as calendar dates approach mid May, when they reach a local minimum. At this point, values increase before decreasing again in September and October, where skill has another minimum. This daily variability in skill shows the difficulties forecast systems have in forecasting the transition seasons as skillfully as the extreme seasons. If we compare this figure with the 91-day running mean calendar day climatological relative frequency of weather regime life cycles in ERA-Interim (defined over the period of 2000–2015; Figure 2), we can relate the high values of HSS observed in January and February to the relatively high frequency of the ZO regime, a regime that is well forecast by all models at this time of the year. Likewise, the local maxima in number of days with no regime observed in April–May and

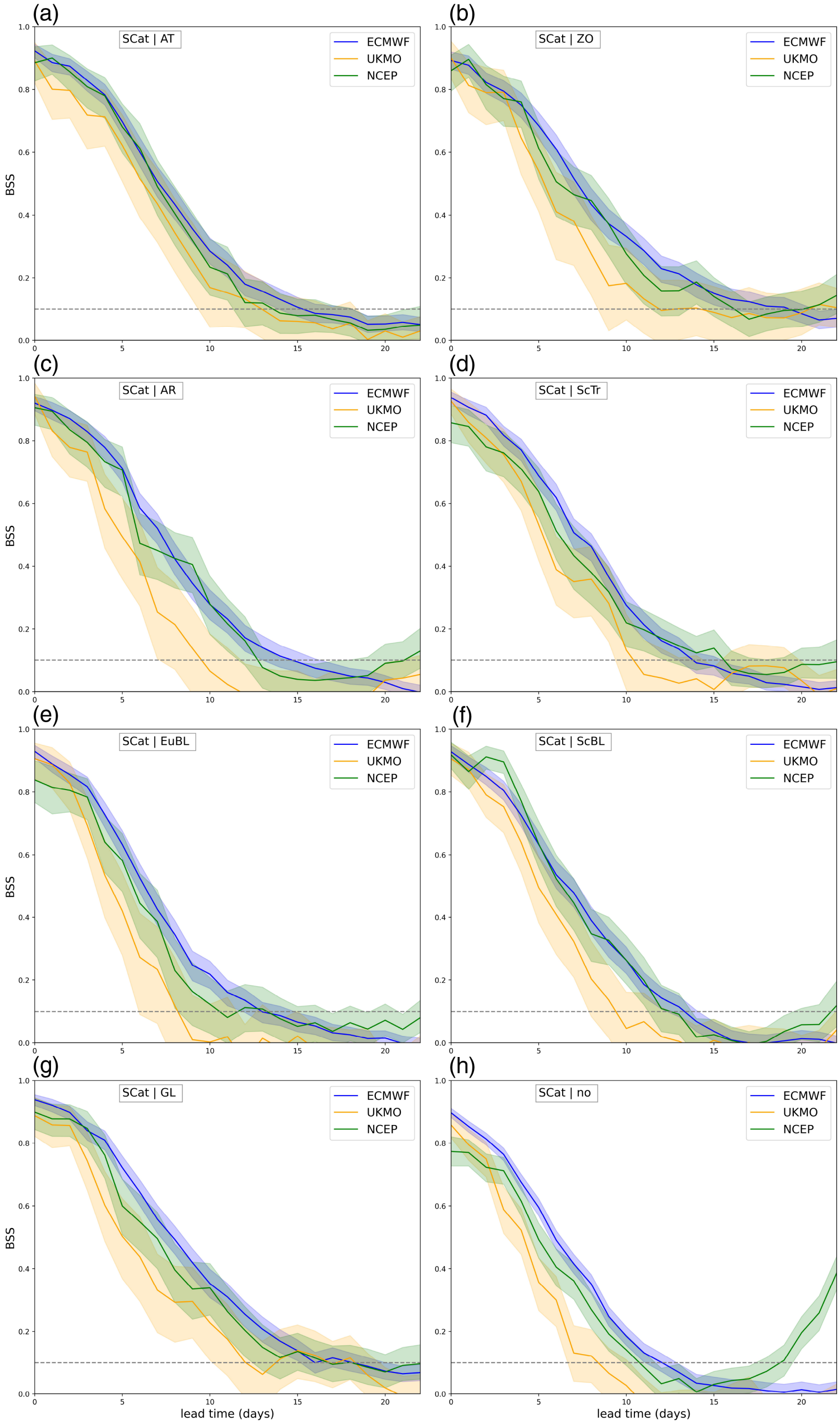

FIGURE 7 Year-round single-category BSS as a function of lead time of the weather regime life cycle for ECMWF (blue), UKMO (orange), and NCEP (green) for (a) AT, (b) ZO, (c) AR, (d) ScTr, (e) EuBL, (f) ScBL, (g) GL, and (h) no regime. Shading shows the range between the 5th and 95th percentiles obtained through a bootstrap procedure. The gray dotted line denotes the 0.1 BSS level. [Colour figure can be viewed at wileyonlinelibrary.com]

September–October can also partially explain the lower values of HSS observed. Therefore, the relative frequency of weather regimes influences the daily variability of skill not only on sub-seasonal time scales but also on shorter time scales.

Figure 9 shows the yearly evolution of the weather regime life-cycle HSS for the three models aggregated over lead times 8–14 days (week 2, left column) and 15–21 days (week 3, right column), together with the relative frequency of occurrence of each weather regime in each year for each of the four seasons. In December–January–February (DJF) for week 2 (Figure 9a), the year-to-year evolution of the skill is similar for the three models. The highest values of HSS for lead

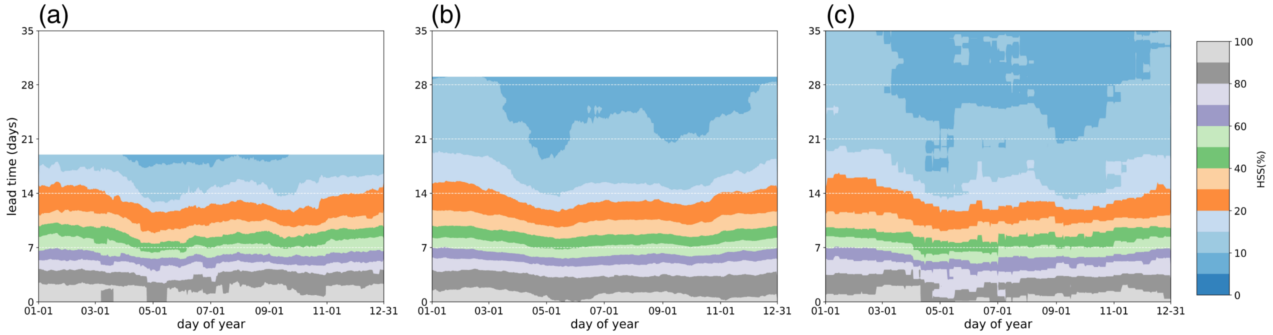

FIGURE 8 Life cycle HSS (%) for all weather regimes as a function of lead time and calendar date. HSS is computed using a 91-day window centered on each calendar day from January 1 to December 31 for (a) NCEP model, (b) ECMWF model, and (c) UKMO model. [Colour figure can be viewed at wileyonlinelibrary.com]

times 8–14 days were observed in 2009/2010, followed by 2013/2014 and 2004/2005. Given that GL is one of the regimes with the best forecast performance in winter, it is not surprising that the 2009/2010 winter has the highest values of HSS, since that winter was characterized by two GL events that lasted for around 30 days (reflected in the positive frequency anomaly for GL) associated with a stratospheric polar vortex split in early December and a sudden stratospheric warming in late January (Dörnbrack *et al.*, 2012). Both 2013/2014 and 2004/2005 experienced more frequent than normal ScTr events, another weather regime forecast well in winter, which could partially explain the relatively good performance of those winters. In addition, the three winters mentioned have very few days with no regime events and experienced an unusually strong stratospheric polar vortex (cf. Fig S1 in Papritz and Grams, 2018). In March–April–May (MAM: Figure 9c), the HSS in week 2 remains around 40% for the three models in the first half of the period, while a small positive trend in skill is observed in the second half, when higher than climatologically expected values of relative frequency for ZO and GL are observed. This trend is significant at the 95% confidence level (based on a *t*-test) for ECMWF and NCEP models. In June–July–August (JJA: Figure 9e), the performance of models for week 2 is very regular and there are no clear summers that stand out. A tendency to more blocking events is evident at the end of the period. However, this does not seem to translate into better or worse skill. Finally, in September–October–November (SON: Figure 9g), the variability of skill for week 2 is less consistent between models and complicates the analysis. Nevertheless, there is a maximum of HSS in 2012 for the UKMO model and a second one in 2015. In 2012 the number of days with ScTr triples the mean values, while autumn 2015 has the second fewest days of no regime. As expected, the values of HSS aggregated over week 3 are lower than week 2. In DJF (Figure 9b), the relatively good performance for week 2 in the 2009/2010 and 2004/2005 winters also extends into week 3. In MAM (Figure 9d), the positive trend in the HSS observed for week 2 is also evident in week 3 (being significant at the 95% confidence level for the three models), although values are lower. In JJA and SON, conclusions similar to those for week 2 can be drawn for week 3. Overall, the largest fluctuations in skill are observed in DJF, which shows potential for exploiting the windows of opportunity that capture the years when the skill is largest, such as the anomalous positive frequency occurrence of GL, ScTr, and ZO or the small no regime frequency.

In some years the performance of models can be attributed to the anomalous frequency of specific weather regimes. As we mentioned previously, the anomalous positive frequency of GL, ScTr, and ZO or the anomalous negative frequency of no regime can partly explain the higher skill observed in DJF and MAM. In line with this, the work by Matsueda and Palmer (2018) shows that the model skill depends on the duration of the negative phase of the NAO, which might be linked to a weak stratospheric polar vortex (Domeisen, 2019). Similarly, Cortesi *et al.* (2021) shows that enhanced periods of skill occur in January and February when regimes that resemble the negative phase of the NAO occur. In this work, we would like to investigate whether this hypothesis on the link between regime frequency anomalies and skill is valid. Figure 10 shows the correlation between the HSS aggregated over lead times 8–14 days (week 2, left column) and 15–21 days (week 3, right column) and the seasonal anomalous occurrence frequency for each weather regime. In DJF, there is no significant relation between the frequency of GL and ScTr and the skill in that season. However, there is a relationship between the skill in week 2 and week 3 and the negative anomalous frequency of no regime and EuBL. These results may indicate the influence of no regime on skill in winter. However, it could also be a by-product of the increased frequency of regime transitions and their lower persistence. In addition, the

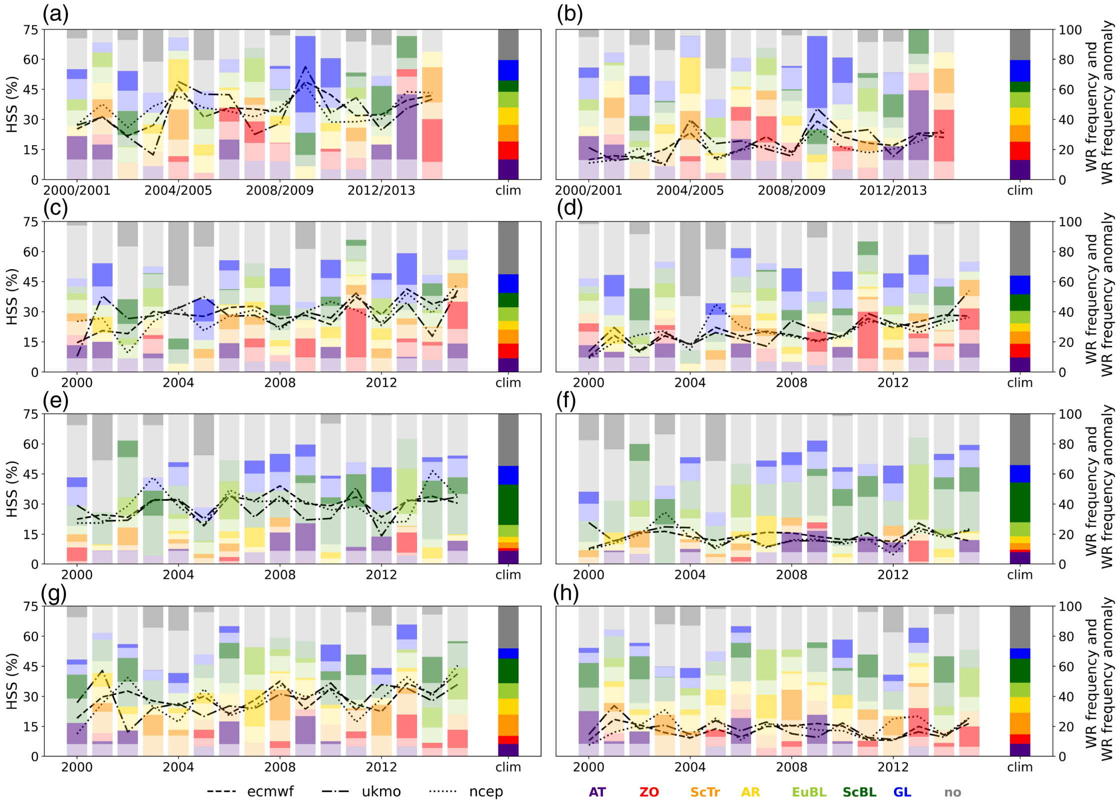

FIGURE 9 Year to year evolution of life cycle HSS (%) for all weather regimes and NCEP (dotted line), ECMWF (dashed line), and UKMO (dot–dashed line) models aggregated over (left column) lead times 8–14 days and (right column) lead times 15–21 days; and year to year evolution of cumulative relative frequency of weather regime life cycles (bars) for (a,b) DJF, (c,d) MAM, (e,f) JJA, and (g,h) SON. Pale bars denote the relative frequency below or equal to the climatological relative frequency (indicated on the end of the plot) of each weather regime (indicated on the end of the plot), while opaque colors denote values exceeding the relative climatological frequency of each weather regime. [Colour figure can be viewed at wileyonlinelibrary.com]

lower skill when there are more EuBL events shows the need for improving the forecast performance for wintertime continental blocking events (already highlighted in Matsueda and Palmer, 2018; Büeler *et al.*, 2021; Ferranti *et al.*, 2018) to enhance the sub-seasonal skill in winter. On the other hand, in MAM the skill of models in weeks 2 and 3 is positively correlated with the anomalous frequency of both ZO and ScTr. This might explain the positive trend in the MAM skill observed in the last years of the period in conjunction with positive anomalies in ZO regime frequency. Finally, the significant relationship between skill in the NCEP model and the frequency of ScBL in JJA may partly explain the relatively good performance of the NCEP model in that season, since more ScBL events are observed in the last years of the period.

3.4 | Representation of regime assignments in the models

When forecasting weather regimes, small deviations in the depicted flow can result in a forecast being assigned to a different regime. To explore whether there are systematic misassignments (i.e., forecasts for a specific weather regime that have a different weather regime as an observed counterpart), Figure 11 shows the contingency table (as percentage of forecasts for each weather regime corresponding to each observed weather regime) for each model at lead times 10–12 days. We selected these lead times because, even though some misassignments are observed at earlier lead times, they stabilize around lead time 10 (see Figures S7, S8, and S9, where the evolution of forecast misassignments as a function of lead time for each model is plotted). In the contingency table, weather regimes are ordered so that related regimes are next to each other. For each weather regime, the contingency tables show large values of correct forecasts (values in the diagonal), but an even larger percentage of false alarms (the forecast for a weather regime verifies as no regime, last row). The latter can be partially explained by cases in which the forecasts marginally fulfill the life-cycle criteria whereas the observations do not. In addition, there are some preferred misassignments common to the three models. Overall, the NCEP model presents the lowest misassignments, while

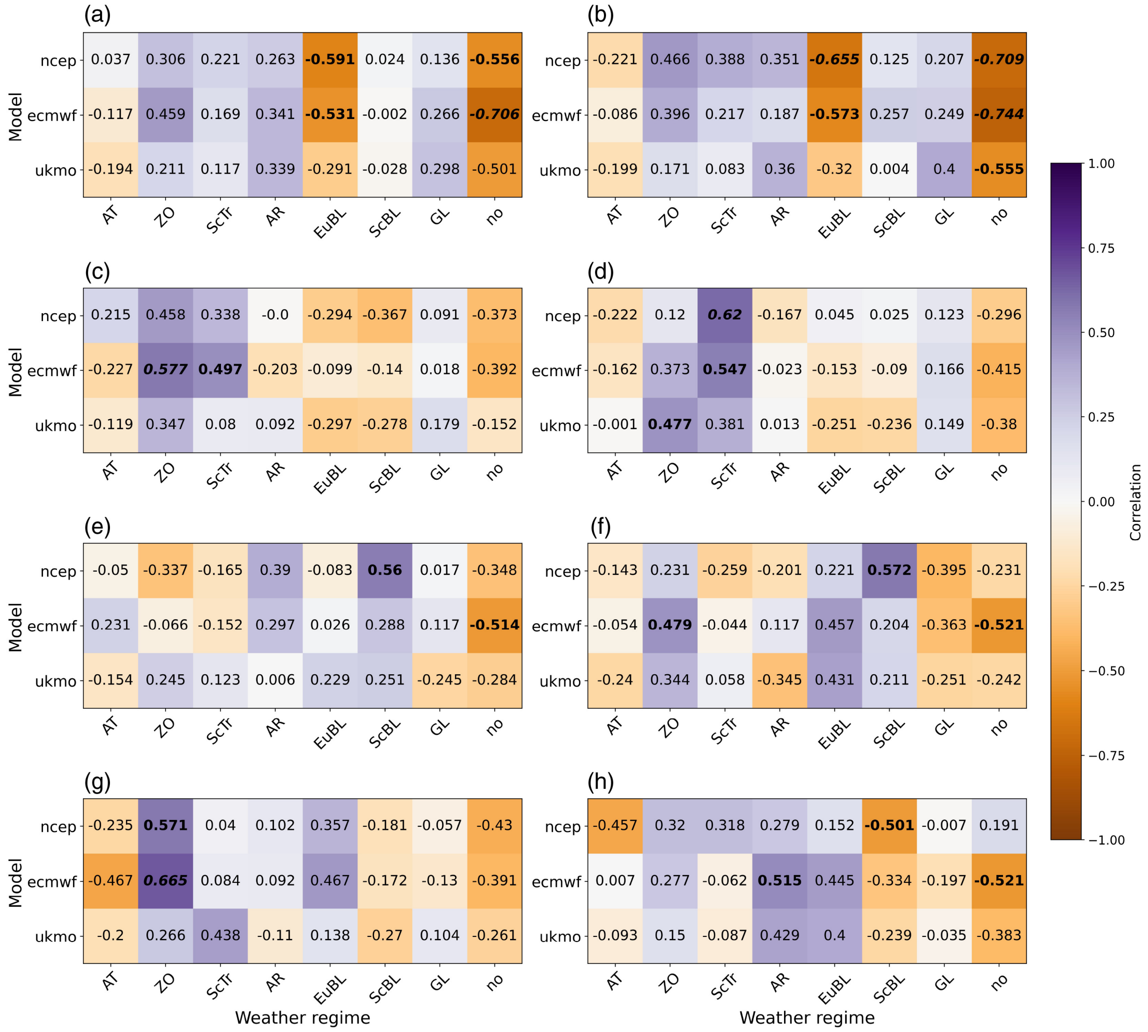

FIGURE 10 Correlation between the HSS aggregated over (left column) lead times 8–14 days (week 2) and (right column) lead times 15–21 days (week 3) and the seasonal anomalous occurrence frequency for each weather regime and the no regime category for (a,b) DJF, (c,d) MAM, (e,f) JJA, and (g,h) SON. Bold italic (bold) values correspond to significant correlations at the 95% (90%) one-tail *t*-test. [Colour figure can be viewed at wileyonlinelibrary.com]

the UKMO shows the largest. With the exception of some missassingments in the UKMO model, most of the documented misassignments occur for the regimes that are related to each other (i.e. the misassignments are close to the diagonal, cf. Figure 1). This opens up the question of the nature of the misassignments: are they due to the existence of two competing signals? Do the models fail in capturing the correct one?

To address these questions, we now analyze whether the misassignments discussed are related to flow situations in which discriminating between two similar regimes is challenging, or instead reflect deficiencies in the models' ability to identify regimes. To do this, we plot composites of $I_{\mathrm{wr}}(t)$ for forecasts and reanalysis for the most common misassignments. As an example, we will show the cases in which the forecasts for AT, ScTr, and EuBL verify as ZO between lead times 10–12 days (Figure 12). The forecast $I_{\mathrm{wr}}(t)$ (dashed lines) shows that the flow situations that are commonly confounded by models project strongly not only in the forecast weather regime but also in the ZO regime (both shown in thicker lines), either at lead time 11 days or some days before. In the models, the ZO weather regime pattern (characterized by a negative $Z500$ anomaly centered over Greenland and positive anomalies in mid-latitudes) appears with three different flavours: either the negative anomaly is shifted southeastwards (characteristic of the AT pattern), or it presents a pattern in which the positive anomaly in the Atlantic is strong and displaced eastwards (and the pattern projects onto the ScTr), or else the positive anomaly over Europe maximizes and is shifted northwards (as when EuBL occurs). This is confirmed when inspecting composites of forecast for geopotential

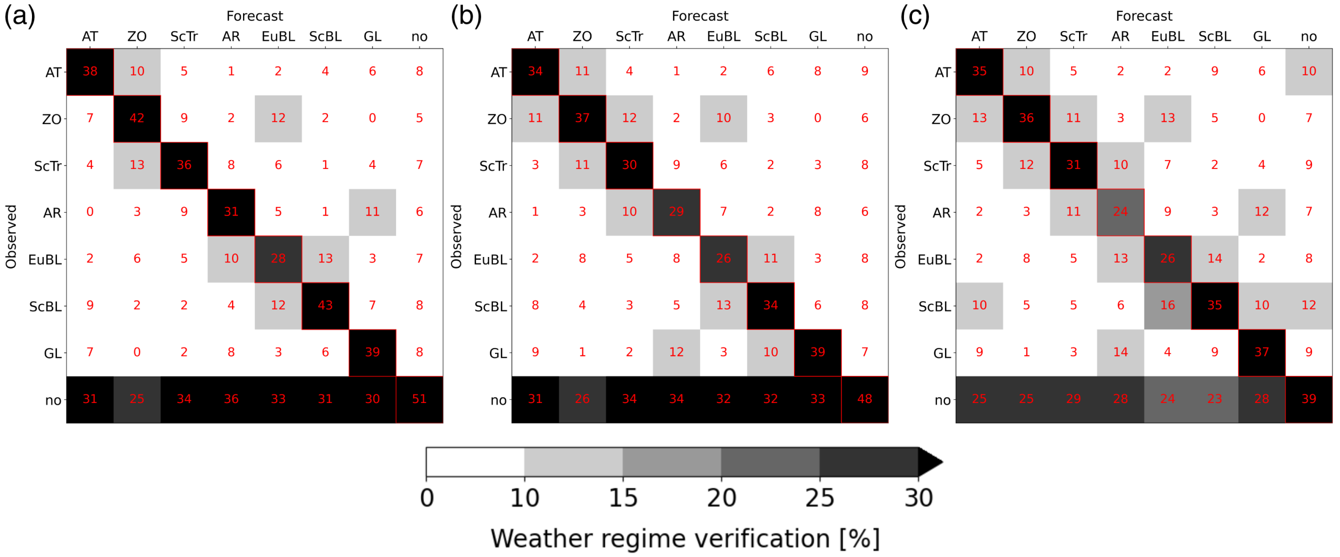

FIGURE 11 Contingency matrix, presented as percentage of forecasts of a weather regime life cycle (columns) at lead time 10 days against the observed weather regime life cycle (rows) for (a) NCEP, (b) ECMWF, and (c) UKMO models. Shading shows values above 10%. [Colour figure can be viewed at wileyonlinelibrary.com]

height anomalies at 500 hPa over the domain of study for the same dates as the misassignments (not shown). If we compare the forecast series with the observed counterpart (solid lines) for the same cases, we can see that, for the misassignment of AT (Figure 12a–c), the maximum of observed $I_{\mathrm{wr}}(t)$ is preceded by a relative maximum of AT that peaks around lead time 7 days and then decays. In the case of the misassignments of ScTr (Figure 12d–f), the maximum of $I_{\mathrm{wr}}(t)$ for ZO is accompanied by a high value of $I_{\mathrm{wr}}(t)$ for ScTr. It seems that the forecast misassignments that occur when models forecast ScTr or AT but ZO is observed occur in flow situations in which two weather regimes have similar signals growing the previous days, and then the ZO signal in models grows and dominates. Models then fail in discriminating between which of these regimes would dominate. Finally, for the misassignment of EuBL (Figure 12g–i), the observed $I_{\mathrm{wr}}(t)$ is maximum for ZO whereas the $I_{\mathrm{wr}}(t)$ for EuBL is half that for ZO, which implies a weaker projection of the flow onto the EuBL. Therefore, the detection of the correct signal in this misassignment is partly related to the existence of two competing signals, but more likely related to the deficiencies in the models in capturing the correct one.

3.5 | Improvements in skill with model versions

Finally, we study the changes in weather regime skill of the models analyzed with respect to previous model versions. This will help us understand whether changes in model forecast system characteristics, such as the model formulation (i.e., dynamical formulation, parametrizations, model components), initialization, or resolution, lead to improvements in skill. To do this, we compare the earlier ECMWF model versions CY43R1, CY45R1, and CY45R3 (those used in Büeler *et al.*, 2021), NCEP Climate Forecast System v2 (CFSv2), and UKMO GloSea5 with the more recent ECMWF, NCEP, and UKMO model versions used in the previous sections, that is, the ECMWF model versions CY46R1 and CY47R1, the NCEP GEFSv12, and the UKMO GloSea6, respectively. Model details can be found in Table 1. These data have been obtained from the S2S project database. In the case of NCEP, instead of selecting the previous version of GEFSv12, the GEFS version implemented in the SubX project, we compared the results with CFSv2, because this model has been used historically in operational activities by the S2S community and also because its reforecasts have been initialized consistently during its entire reforecast period, which facilitates bias correction (Guan *et al.*, 2019, whereas GEFS-SubX has inconsistencies in its re-forecast climatologies;), whereas GEFS-SubX has inconsistencies in its reforecast climatologies. There are several differences between the two NCEP models, in not only atmospheric but also other Earth system components of the model, as well as in the reanalysis used to initialize the reforecasts. The main difference between the ECMWF model versions is the reanalysis used to initialize the reforecasts (ERA-Interim for versions CY43R1-CY45R3 and ERA5 for versions CY46R1 and CY47R1), while the main difference in the UKMO model versions is the atmospheric

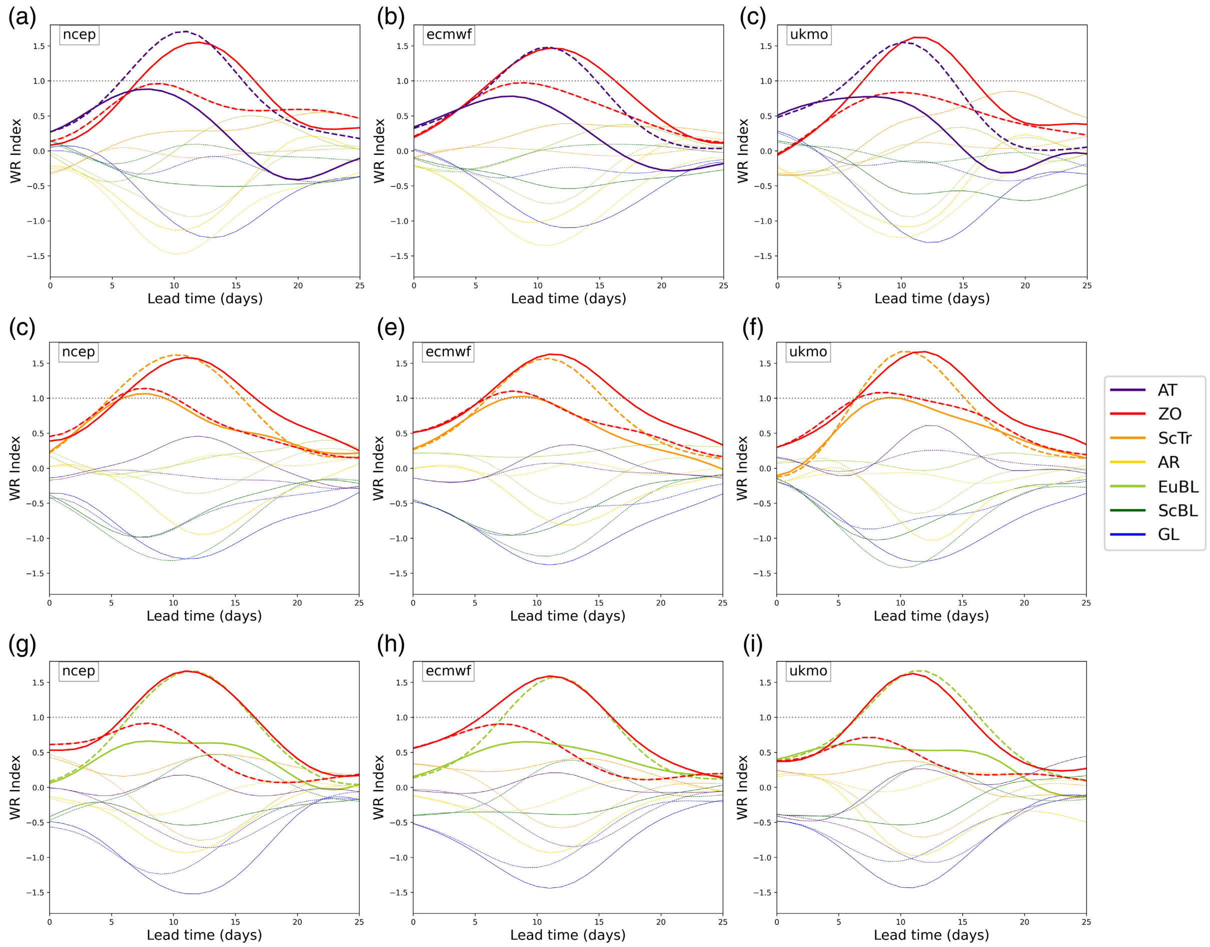

FIGURE 12 Composites of observed (solid) and forecast (dashed) time series of $I_{\mathrm{wr}}(t)$ forecast by (a,d,g) NCEP, (b,e,h) ECMWF, and (c,f,i) UKMO for forecasts that at lead time 10–12 days forecast AT (first row), ScTr (second row), and EuBL (third row) and verify in ZO. Thick lines depict the misassignmments analyzed in detail (see text). The dotted gray line denotes the 1.0 level for I_{wr}, that is, the threshold for an active regime life cycle. [Colour figure can be viewed at wileyonlinelibrary.com]

component of the model (HadGEM3 GC3.2 in GloSea 6 and HadGEM3 GC2.0 in GloSea 5) and the initialization of the land-surface model. For each modelling center, we compare the models in the largest common reforecast period available for the versions compared, that is, the 1999–2010 period for NCEP, the 1999–2017 period for ECMWF, and the 1993–2015 period for UKMO.

Figure 13a shows the difference in BSS between the new (GEFSv12) NCEP version and the old (CFSv2) NCEP version for the year-round multicategory BSS and the multicategory BSS for forecasts initialized in each season (positive values mean that the new NCEP version is better than the old NCEP version and negative values mean the opposite). The performance of the new NCEP version in terms of BSS is better than that for the old NCEP version for almost all lead times, being significant from lead time 1 onward using all initializations. The skill horizon thus extends by about 4 days in the new version with respect to the old version (not shown). Dividing the initial conditions by seasons also shows improvements in all seasons and at almost all lead times, the most important being for JJA between lead times 3 and 11 and for MAM between lead times 10 and 15. If we take a look at the performance for individual weather regimes, as depicted by differences between the single-category BSS for the new version of NCEP and the old version of NCEP for blocked (Figure 13b) and cyclonic (Figure 13c) regimes, it can be seen that both present improvements for all lead times and the improvements in the cyclonic regimes peak a couple of days earlier than those in the blocked regimes. The largest improvements are seen for AR between lead times 8 and 12. A similar analysis is done for ECMWF model versions (Figure 14). It is important to note here that, because models are verified against ERA-Interim, which is the reanalysis used to initialize the old ECMWF version, this may penalize the performance of the new version of ECMWF, initialized with ERA5. The improvements in skill in the new versions are evident until at least lead time

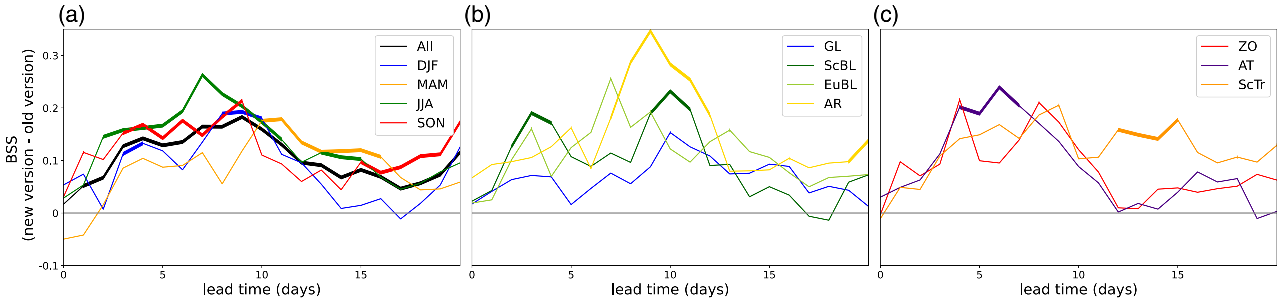

FIGURE 13 The difference in the BSS between NCEP GEFSv12 version (the new NCEP version) and NCEP CFSv2 version (the old NCEP version) for (a) seasonal multicategory life cycle for all weather regimes as a function of lead time, (b) year-round single-category life cycle for blocked regimes, and (c) year-round single-category life cycle for cyclonic regimes. Positive values mean that NCEP GEFSv12 BSS is higher than NCEP CFSv2 BSS. Negative values mean the opposite. Thick lines mean that the difference in the BSS between both versions is significant. [Colour figure can be viewed at wileyonlinelibrary.com]

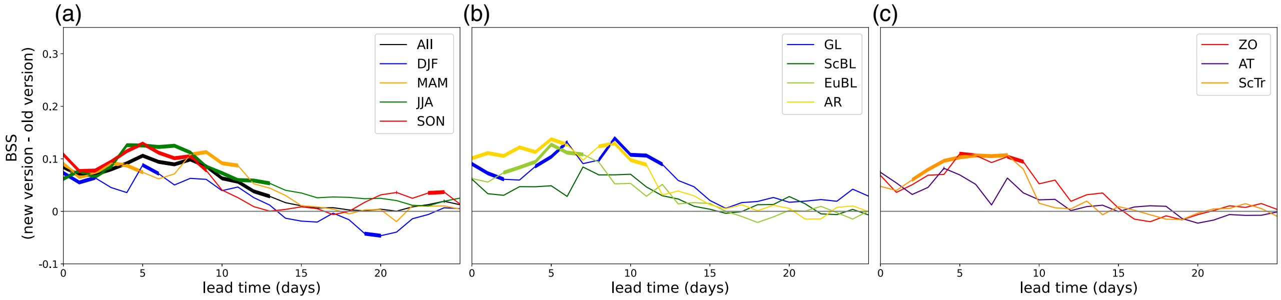

FIGURE 14 The difference in the BSS between ECMWF CY46R1 and CY47R1 versions (the new ECMWF version) and ECMWF CY43R1, CY43R3, and CY45R1 versions (the old ECMWF version) for (a) seasonal multicategory life cycle for all weather regimes as a function of lead time, (b) year-round single-category life cycle for blocked regimes, and (c) year-round single-category life cycle for cyclonic regimes. Positive values mean that ECMWF CY46R1 and CY47R1 BSS is higher than ECMWF CY43R1, CY43R3 and CY45R1 BSS. Negative values mean the opposite. Thick lines mean that the difference in the BSS between both versions is significant. [Colour figure can be viewed at wileyonlinelibrary.com]

15 days for all seasons, although the improvement is significant mainly for forecasts initialized in JJA and SON up to lead times of 13 and 11 days, respectively. In addition, higher skill is obtained with the newer versions for all weather regimes until lead time 15 days. At shorter lead times (1–6 days), those changes are more important for all blocked regimes except ScBL. The smaller improvements in ScBL in the new version of ECMWF can explain the relatively similar performance of this model for EuBL and ScBL. Finally, the comparison of UKMO GloSea versions (Figure 15) shows that the skill has increased mainly at early lead times (1–3 days). Looking at each regime separately, the largest improvements are seen for AT between lead times 5 and 12.

4 | SUMMARY AND CONCLUSIONS

We studied the skill of three sub-seasonal forecasts models, namely NCEP, ECMWF, and UKMO, in forecasting seven year-round Atlantic–European weather regimes. To make fair comparisons between the models, we analyzed the same reforecast period and bias-corrected (i.e., calibrated) all the reforecasts against their own model climate. We showed that the NCEP model represents the main weather regime characteristics, such as the frequency, length, and number, best and has the lowest biases in weather regime transitions. In addition, our analysis revealed that the calibrated forecasts still presented significant positive biases in the frequency of no regime events in winter and negative biases in ScBL in summer for ECMWF and UKMO. The latter can partly be explained by the underestimation of the number of ScBL events and the biases in the transition from other weather regimes into ScBL in that season. We further showed that the performance of the models in forecasting the weather regime life cycle (LC) is better than that for the maximum of the weather regime indices (maxIwr) and the geopotential height anomalies at 500 hPa over the regime domain. This was partly documented by Büeler *et al.* (2021) for the

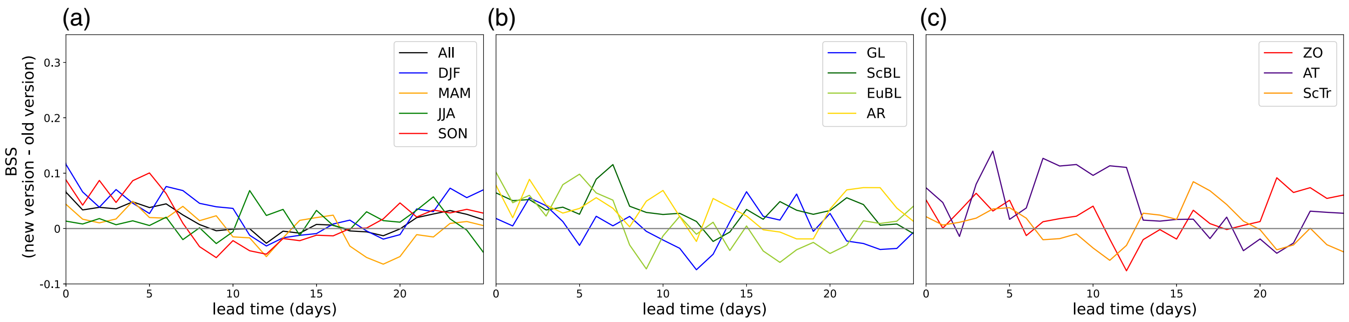

FIGURE 15 The difference in the BSS between EUKMO GloSea6 version (the new UKMO version) and UKMO GloSea5 version (the old UKMO version) for (a) seasonal multicategory life cycle for all weather regimes as a function of lead time, (b) year-round single-category life cycle for blocked regimes, and (c) year-round single-category life cycle for cyclonic regimes. Positive values mean that UKMO GloSea6 BSS is higher than UKMO GloSea5 BSS. Negative values mean the opposite. Thick lines mean that the difference in the BSS between both versions is significant. [Colour figure can be viewed at wileyonlinelibrary.com]

LC and maxIwr for the ECMWF model. We expanded this result to other models and variables. This demonstrates the benefit of introducing the LC definition to identify predictable modes, enhancing the practical S2S predictability of the large-scale flow.

We then analyzed the relative skill of models depending on the season the forecasts were initialized in and the weather regime predicted. We found that ECMWF is the model with the best Brier Skill Score (BSS) for all seasons and weather regimes (the skill horizon is around 14 days), although the performance of NCEP is comparable (skill horizon of around 13 days). Furthermore, we found that in summer and for the AT and ScBL the performance of both models is almost identical. In addition, and in agreement with Büeler *et al.* (2021), there are differences in the skill between seasons and weather regimes: in winter the skill is about 3 days longer than in the remaining seasons for ECMWF and NCEP, whereas the skill has nonsignificant differences between seasons in the UKMO model. The stratification of the skill into the individual weather regimes showed that the EuBL skill horizon is 2–6 days shorter than that for other weather regimes in NCEP and 2–4 days shorter in UKMO. On the other hand, in the ECMWF the skill horizon for EuBL is 1 day shorter than for ScBL and both have a 1–6 days shorter skill horizon than the remaining weather regimes. The skill horizon for ZO and GL tends to be the longest for all three models, which partly explains the high skill in winter, when these two regimes are most persistent. The skill for the no regime category is lowest for all three models, which shows the usefulness of including this category to remove less predictable episodes, such as regime transitions. This result also confirms the difficulties that models have in forecasting situations where no particular flow dominates.

We also assessed the day-to-day and year-to-year variability of skill, depicted by the Heidke Skill Score (HSS), and the role of weather regime occurrence in shaping this variability. The three models present their highest HSS values in January and February, which remain above 0.1 up to 35 days for UKMO, 28 days for ECMWF, and 20 days for NCEP. The three models present a seasonality in the skill, with relative minimum values in May and late September, when the observed no regime frequency peaks. These relative minima are also observed for short lead times (1–5 days), which implies that some flow situations can impact on skill even in lead times where skill is tied to initial conditions. It also demonstrates the predictability gap seen in the transition seasons. On the other hand, the anomalous frequency of no regime days in winter also influences the interannual variability of skill on sub-seasonal time scales in that season: the winters with the fewest no regime days are related to years with enhanced week 2 skill. In addition, higher week 2 and 3 winter-time skill values are observed in the years with the fewest EuBL events, which highlights the challenges in forecasting this regime skillfully. Another significant influence of observed weather regime variability on skill is found in autumn, when skill is enhanced in years with more ZO regimes. Finally, the skill for the NCEP model in summer is higher in years with more ScBL. This model presents the lowest bias in the mean geopotential height in summer and is the best at reproducing the main ScBL characteristics, such as frequency, number, and length. There have been very few studies yet using this NCEP model version (it was implemented in late 2020) and therefore further studies of the representation of dynamical processes during ScBL by this model could help us understand its relatively good performance. For instance, the work by Quinting & Vitart (2019) shows that the NCEP CFSv2 model presents the lowest biases in blocking frequency over the Atlantic–European domain across S2S models, due to a better representation of Rossby-wave packets. Similar studies with NCEP GEFSv12 may shed light on processes that explain this relatively good performance.

We then explored the systematic misassignments of forecasts for long lead times to understand whether they are related to model deficiencies or flow situations in which there are competing regime signals. For the three models, the misassignments grow from lead time 0 and saturate after a lead time of 10 days. We found that most of the misassignments correspond to regimes that have similar configurations. As an example, we showed the flow situations when ZO is observed but models forecast either ScTr, AT, or EuBL. In the cases of ScTr and AT, both observed flow situations are such that the signal is high for ScTr and ZO and AT and ZO, respectively, whereas the misassignments for EuBL are partly related to model deficiencies in distinguishing between ZO and EuBL regimes. The identification of misassignments can help to improve the accuracy of models by placing different weights on ensemble members based on these results. In addition, the study of the misassignments can help to distinguish windows of opportunity of enhanced predictability when multiple models agree on a pattern from situations when models present the same misassignments and all point to the incorrect regime. Future studies on regime misassignments could help us to understand the dynamical processes behind them better.

Finally, we investigated the evolution of forecast skill in each modelling center. We found that the skill has improved in ECMWF and NCEP models with newer versions. For the NCEP model, this improvement leads to an extension of the skill by around 4 days. On the other hand, the skill for UKMO has not changed considerably across the versions considered. Changes in the reanalysis used to initialize the ECMWF model mainly impacted the first 12 days, although the improvements are greater for blocking events, with the exception of ScBL. In the case of NCEP model versions, there may be multiple explanations of the observed improvements, since the models compared have very little in common. However, changes in the atmospheric model, from a spectral model to a finite-volume dynamical core, as well as the perturbation scheme, from a time-lagged ensemble to an ensemble Kalman filter, might be responsible for the substantial improvements seen. Finally, in the UKMO model, changes in the atmospheric model and the initialization of the land-surface model have almost no impact on the skill.

At the moment, the only work that has evaluated the performance of the seven year-round regimes is by Büeler *et al.* (2021). In this study, we expanded that work by including a newer version of the model used in that study and two additional models and by investigating the sources of daily to interannual skill variability. This study is thus the first systematic multi-model assessment of the seven year-round weather regimes with different state-of-the-art sub-seasonal models and represents a contribution towards a more objective assessment of the evolution of weather regimes for days to weeks ahead, allowing better forecast guidance in the decision-making process. For instance, forecasters can weigh the models against each other based on the results of this assessment, or can compare the models and see to what extent they agree or disagree to identify situations of high uncertainty. In addition, it is important to highlight that understanding the impacts of regimes on surface weather variables is as important as (or even more important than) understanding large-scale flow situations (Bloomfield *et al.*, 2021), especially for economic sectors. For example, studies have shown that in winter there is an overproduction of wind power under cyclonic regimes (Grams *et al.*, 2017), whereas EuBL and GL blocking events are associated with periods of low electricity production by renewable sources (wind power and solar) and high electricity demand that lead to stress on the electricity system (Mockert *et al.*, 2022; Otero *et al.*, 2022). In this context, improving the skill for wintertime EuBL is thus crucial for better preparedness for those critical situations. In addition, assessing how models forecast the relationship between weather regimes and surface weather is also very relevant for end users, but it is beyond the scope of this work.

The possibility of analyzing robustly the role of different phases of the Madden–Julian Oscillation or the intensity of the stratospheric polar vortex was limited by the short common reforecast period. This short common reforecast period shows the limitations of current sub-seasonal databases. Newer projects, in collaboration with forecast centers, might consider new forecast strategies, such as aligning the reforecast calendar and extending the common reforecast period, to address the windows of opportunity across models systematically.

AUTHOR CONTRIBUTIONS

Marisol Osman: conceptualization; data curation; formal analysis; investigation; methodology; visualization; writing – original draft. **Remo Beerli:** conceptualization; methodology; writing – original draft. **Dominik Büeler:** data curation; writing – original draft. **Christian M. Grams:** data curation; methodology; writing – original draft.

ACKNOWLEDGEMENTS

This work is based on S2S data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF as an extension of the TIGGE database. The NOAA Global Ensemble Forecast System (GEFS) Reforecast was accessed in

February 2022 from https://registry.opendata.aws/noaa-gefs-reforecast. We are grateful to the members of the Large-Scale Dynamics and Predictability group for valuable comments on this research. We thank the two anonymous reviewers for assessing the manuscript and giving very useful comments that helped to improve this study. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The contribution of Marisol Osman and Remo Beerli was supported by Axpo Solutions AG. The contribution of CMG and DB is funded by the Helmholtz Association as part of the Young Investigator Group "Sub-seasonal Predictability: Understanding the Role of Diabatic Outflow" (SPREADOUT, grant VH-NG-1243).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

ORCID

Marisol Osman https://orcid.org/0000-0002-6275-1454
Remo Beerli https://orcid.org/0000-0002-0897-9650
Dominik Büeler https://orcid.org/0000-0002-9904-6281
Christian M. Grams https://orcid.org/0000-0003-3466-9389

REFERENCES

Beerli, R. and Grams, C.M. (2019) Stratospheric modulation of the large-scale circulation in the Atlantic-european region and its implications for surface weather events. *Quarterly Journal of the Royal Meteorological Society*, 145, 3732–3750 https://onlinelibrary.wiley.com/doi/10.1002/qj.3653.

Bloomfield, H.C., Brayshaw, D.J. and Charlton-Perez, A.J. (2020) Characterizing the winter meteorological drivers of the european electricity system using targeted circulation types. *Meteorological Applications*, 27, e1858 https://onlinelibrary.wiley.com/doi/full/10.1002/met.1858.

Bloomfield, H.C., Brayshaw, D.J., Gonzalez, P.L. and Charlton-Perez, A. (2021) Pattern-based conditioning enhances sub-seasonal prediction skill of european national energy variables. *Meteorological Applications*, 28, e2018 https://onlinelibrary.wiley.com/doi/full/10.1002/met.2018.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Büeler, D., Beerli, R., Wernli, H. and Grams, C.M. (2020) Stratospheric influence on ecmwf sub-seasonal forecast skill for energy-industry-relevant surface weather in european countries. *Quarterly Journal of the Royal Meteorological Society*, 146, 3675–3694 https://onlinelibrary.wiley.com/doi/10.1002/qj.3866.

Büeler, D., Ferranti, L., Magnusson, L., Quinting, J.F. and 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 https://onlinelibrary.wiley.com/doi/10.1002/qj.4178.

Cassou, C. (2008) Intraseasonal interaction between the madden-julian oscillation and the North Atlantic oscillation. *Nature*, 455, 523–527.

Charlton-Perez, A.J., Aldridge, R.W., Grams, C.M. and Lee, R. (2019) Winter pressures on the Uk health system dominated by the Greenland blocking weather regime. *Weather and Climate Extremes*, 25, 100218.

Cortesi, N., Torralba, V., Lledó, L., Manrique-Suñén, A., Gonzalez-Reviriego, N., Soret, A. and Doblas-Reyes, F.J. (2021) Yearly evolution of euro-Atlantic weather regimes and of their sub-seasonal predictability. *Climate Dynamics*, 56, 3933–3964. https://doi.org/10.1007/s00382-021-05679-y.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Källberg, P., Köhler, M., Matricardi, M., Mcnally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N. and Vitart, F. (2011) The era-interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597.

Domeisen, D.I. (2019) Estimating the frequency of sudden stratospheric warming events from surface observations of the North Atlantic oscillation. *Journal of Geophysical Research: Atmospheres*, 124, 3180–3194 https://onlinelibrary.wiley.com/doi/full/10.1029/2018JD030077.

Domeisen, D.I.V., Grams, C.M. and Papritz, L. (2020) The role of North Atlantic-european weather regimes in the surface impact of sudden stratospheric warming events. *Weather and Climate Dynamics*, 1, 373–388.

Dörnbrack, A., Pitts, M.C., Poole, L.R., Orsolini, Y.J., Nishii, K. and Nakamura, H. (2012) The 2009-2010 arctic stratospheric winter-general evolution, mountain waves and predictability of an operational weather forecast model. *Atmospheric Chemistry and Physics*, 12, 3659–3675 https://acp.copernicus.org/articles/12/3659/2012/.

Falkena, S.K., de Wiljes, J., Weisheimer, A. and Shepherd, T.G. (2020) Revisiting the identification of wintertime atmospheric circulation regimes in the euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 146, 2801–2814.

Ferranti, L., Magnusson, L., Vitart, F. and 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 https://onlinelibrary.wiley.com/doi/10.1002/qj.3341.

Ferro, C.A., Richardson, D.S. and Weigel, A.P. (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15, 19–24 https://onlinelibrary.wiley.com/doi/full/10.1002/met.45.

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

Guan, H., Zhu, Y., Sinsky, E., Li, W., Zhou, X., Hou, D., Melhauser, C. and Wobus, R. (2019) Systematic error analysis and calibration of

2-m temperature for the ncep gefs reforecast of the subseasonal experiment (subx) project. *Weather and Forecasting*, 34, 361–376.

Heidke, P. (2017) Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geografiska Annaler*, 8, 301–349 https://www.tandfonline.com/doi/abs/10.1080/20014422.1926.11881138.

Long, C.S., Fujiwara, M., Davis, S., Mitchell, D.M. and Wright, C.J. (2017) Climatology and interannual variability of dynamic variables in multiple reanalyses evaluated by the sparc reanalysis intercomparison project (s-rip). *Atmospheric Chemistry and Physics*, 17, 14593–14629 https://acp.copernicus.org/articles/17/14593/2017/.

Maclachlan, C., Arribas, A., Peterson, K.A., Maidens, A., Fereday, D., Scaife, A.A., Gordon, M., Vellinga, M., Williams, A., Comer, R.E., Camp, J., Xavier, P. and Madec, G. (2015) Global seasonal forecast system version 5 (glosea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141, 1072–1084 https://onlinelibrary.wiley.com/doi/full/10.1002/qj.2396.

Magnusson, L., Alonso-Balmaseda, M., Dahoui, M., Forbes, R., Haiden, T., Lavers, D., Sandu, I. and Tietsche, S. (2022) Summary of the ugrow subproject on tropospheric temperature bias during jja over the northern hemisphere. Tech. Rep. 891, ECMWF. URL: https://www.ecmwf.int/node/20356.

Mariotti, A., Baggett, C., Barnes, E.A., Becker, E., Butler, A., Collins, D.C., Dirmeyer, P.A., Ferranti, L., Johnson, N.C., Jones, J., Kirtman, B.P., Lang, A.L., Molod, A., Newman, M., Robertson, A.W., Schubert, S., Waliser, D.E. and Albers, J. (2020) Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, 101, E608–E625.

Matsueda, M. and Palmer, T.N. (2018) Estimates of flow?Dependent predictability of wintertime euro?Atlantic weather regimes in medium?Range forecasts. *Quarterly Journal of the Royal Meteorological Society*, 144, 1012–1027 https://onlinelibrary.wiley.com/doi/10.1002/qj.3265.

Michelangeli, P.A., Vautard, R. and Legras, B. (1995) Weather regimes: recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, 52, 1237–1256.

Mockert, F., Grams, C.M., Brown, T. and Neumann, F. (2022) Meteorological conditions during dunkelflauten in Germany: characteristics, the role of weather regimes and impacts on demand.

Otero, N., Martius, O., Allen, S., Bloomfield, H. and Schaefli, B. (2022) Characterizing renewable energy compound events across europe using a logistic regression-based approach. *Meteorological Applications*, 29, e2089 https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.2089.

Papritz, L. and Grams, C.M. (2018) Linking low-frequency large-scale circulation patterns to cold air outbreak formation in the northeastern North Atlantic. *Geophysical Research Letters*, 45, 2542–2553 https://onlinelibrary.wiley.com/doi/full/10.1002/2017GL076921.

Pasquier, J.T., Pfahl, S. and Grams, C.M. (2019) Modulation of atmospheric river occurrence and associated precipitation extremes in the North Atlantic region by european weather regimes. *Geophysical Research Letters*, 46, 1014–1023 https://onlinelibrary.wiley.com/doi/full/10.1029/2018GL081194.

Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., Lajoie, E., Burgman, R., Bell, R., Delsole, T., Min, D., Zhu, Y., Li, W., Sinsky, E., Guan, H., Gottschalck, J., Metzger, E.J., Barton, N.P., Achuthavarier, D., Marshak, J., Koster, R.D., Lin, H., Gagnon, N., Bell, M., Tippett, M.K., Robertson, A.W., Sun, S., Benjamin, S.G., Green, B.W., Bleck, R. and Kim, H. (2019) The subseasonal experiment (subx). *Bulletin of the American Meteorological Society*, 100, 2043–2060.

Quinting, J.F. and Vitart, F. (2019) Representation of synoptic-scale rossby wave packets and blocking in the s2s prediction project database. *Geophysical Research Letters*, 46, 1070–1078. https://doi.org/10.1029/2018GL081381.

Son, S.W., Kim, H., Song, K., Kim, S.W., Martineau, P., Hyun, Y.K. and Kim, Y. (2020) Extratropical prediction skill of the subseasonal-to-seasonal (s2s) prediction models. *Journal of Geophysical Research: Atmospheres*, 125, e2019JD031273 https://onlinelibrary.wiley.com/doi/full/10.1029/2019JD031273.

Spensberger, C., Madonna, E., Boettcher, M., Grams, C.M., Papritz, L., Quinting, J.F., Röthlisberger, M., Sprenger, M. and Zschenderlein, P. (2020) Dynamics of concurrent and sequential central european and scandinavian heatwaves. *Quarterly Journal of the Royal Meteorological Society*, 146, 2998–3013 https://onlinelibrary.wiley.com/doi/full/10.1002/qj.3822.

van der Wiel, K., Bloomfield, H.C., Lee, R.W., Stoop, L.P., Blackport, R., Screen, J.A. and Selten, F.M. (2019) The influence of weather regimes on european renewable energy production and demand. *Environmental Research Letters*, 14, 094010 https://iopscience.iop.org/article/10.1088/1748-9326/ab38d3.

Vigaud, N., Robertson, A. and Tippett, M.K. (2018) Predictability of recurrent weather regimes over north america during winter from submonthly reforecasts. *Monthly Weather Review*, 146, 2559–2577 http://journals.ametsoc.org/doi/10.1175/MWR-D-18-0058.1.

Vitart, F. (2014) Evolution of ecmwf sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, 140, 1889–1899 https://onlinelibrary.wiley.com/doi/10.1002/qj.2256.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A.W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.J., Xiao, H., Zaripov, R. and Zhang, L. (2017) The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, 98, 163–173.

White, C.J., Domeisen, D.I., Acharya, N., Adefisan, E.A., Anderson, M.L., Aura, S., Balogun, A.A., Bertram, D., Bluhm, S., Brayshaw, D.J., Browell, J., Büeler, D., Charlton-Perez, A., Chourio, X., Christel, I., Coelho, C.A., DeFlorio, M.J., Monache, L.D., Giuseppe, F.D., Garc, A.M., Gibson, P.B., Goddard, L., Romero, C.G., Graham, R.J., Graham, R.M., Grams, C.M., Halford, A., Huang, W.T., Jensen, K., Kilavi, M., Lawal, K.A., Lee, R.W., MacLeod, D., Manrique-Suñén, A., Martins, E.S., Maxwell, C.J., Merryfield, W.J., Muñoz, Á.G., Olaniyan, E., Otieno, G., Oyedepo, J.A., Palma, L., Pechlivanidis, I.G., Pons, D., Ralph, F.M., Reis, D.S., Remenyi, T.A., Risbey, J.S., Robertson, D.J., Robertson, A.W., Smith, S., Soret, A., Sun, T., Todd, M.C., Tozer, C.R., Vasconcelos, F.C., Vigo, I., Waliser, D.E., Wetterhall, F. and Wilson, R.G. (2022) Advances in the application and utility of subseasonal-to-seasonal predictions. *Bulletin of the American Meteorological Society*, 103, E1448–E1472 https://journals.ametsoc.org/view/journals/bams/103/6/BAMS-D-20-0224.1.xml.

Wilks, D.S. (2019) *Statistical Methods in Atmoshperic Sciences*. Amsterdam: Candice Janco.

Zhou, X., Zhu, Y., Hou, D., Fu, B., Li, W., Guan, H., Sinsky, E., Kolczynski, W., Xue, X., Luo, Y., Peng, J., Yang, B.O., Tallapragada, V. and Pegion, P. (2022) The development of the ncep global ensemble forecast system version 12. *Weather and Forecasting*, 37, 1069–1084 https://journals.ametsoc.org/view/journals/wefo/37/6/WAF-D-21-0112.1.xml.

Zubiate, L., McDermott, F., Sweeney, C. and O'Malley, M. (2017) Spatial variability in winter nao-wind speed relationships in western europe linked to concomitant states of the East Atlantic and scandinavian patterns. *Quarterly Journal of the Royal Meteorological Society*, 143, 552–562 https://onlinelibrary.wiley.com/doi/full/10.1002/qj.2943.

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: 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(755), 2386–2408. Available from: https://doi.org/10.1002/qj.4512

APPENDIX A. WR FORECAST SKILL OF EACH MODEL IN EACH SEASON

Here we briefly describe the main weather regimes forecast skill of each model in each season. Tables S1, S2, and S3 show a summary of the skill for individual regimes year round and for each season and for the NCEP, ECMWF, and UKMO models, respectively. Note that the statistics are less robust for skill in seasons due to the smaller sample size. For the NCEP model (Table S1), the poor skill for short lead times in EuBL is driven by its performance in DJF, whereas the relatively bad performance for ScBL in DJF, MAM, and SON these seasons is compensated by a good performance in JJA. The good performance for GL and ZO for long lead times is related to the good skill observed in all seasons except JJA for GL and also SON for ZO. The relatively low skill for ScBL and EuBL for ECMWF is also observed in DJF and SON, and in JJA, respectively (Table S2). As for the NCEP model, the good performance for year-round long lead times of ZO and GL is mainly observed in DJF and MAM. The AT, which presents together with GL the second largest skill horizon, has good performance in DJF and MAM. For the UKMO model (Table S3), the EuBL, ScBL, AR, and no regime have the lowest skill, while the GL and ZO have the highest skill. In winter, the ZO regime has the largest skill horizon (27 days), whereas the skill horizon for GL in MAM is 21 days. On the other hand, in JJA the AR has the poorest performance, with a skill horizon in 5 days.