Late Spring and Summer Subseasonal Forecasts in the Northern Hemisphere Midlatitudes: Biases and Skill in the ECMWF Model

EMANUEL DUTRA,^{a,b} FREDERICO JOHANNSEN,^a AND LINUS MAGNUSSON^c

^a *Instituto Dom Luiz (IDL), Faculty of Sciences, University of Lisbon, Lisbon, Portugal*

^b *Instituto Português do Mar e da Atmosfera (IPMA), Lisbon, Portugal*

^c *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*

(Manuscript received 20 October 2020, in final form 17 May 2021)

ABSTRACT: Subseasonal forecasts lie between medium-range and seasonal time scales with an emerging attention due to the relevance in society and by the scientific challenges involved. This study aims to (i) evaluate the development of systematic errors with lead time in the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecasts of surface-related variables during late spring and summer, and (ii) investigate potential relationships between the systematic errors and predictive skill. The evaluation is performed over the Northern Hemisphere midlatitudes, focusing on several regions with different climate characteristics. The results indicate five key bias patterns: (i) cold bias of daily maximum temperature (mx2t) in the April–May forecasts at all lead times in most regions; (ii) central North America with a warm bias mostly in the daily minimum temperature (mn2t); (iii) east of the Caspian Sea region with a warm and dry bias; (iv) western and Mediterranean Europe with a cold bias in mn2t mainly in April–May forecasts; and (v) continental Europe with a cold bias in the mx2t and warm bias of mn2t in the June–July forecasts. We also found substantial deviations of soil moisture and terrestrial water storage variation in most regions compared to the fifth-generation ECMWF atmospheric reanalysis (ERA5). Despite the large differences in the systematic error characteristics among the different regions, there is little relation to the skill of the subseasonal forecasts. The systematic temperature biases require further attention from model developers as diurnal cycle improvements could enhance some of the potential predictability coming from the long-memory effect of soil moisture.

SIGNIFICANCE STATEMENT: Weather forecasts up to 1 week in advance have increased quality substantially in the last decades. This drives an increased attention and demand of quality forecasts for longer lead times, with 2 to 3 weeks in advance. This “subseasonal” scale has many challenges due to the chaotic nature of the atmosphere, which limits predictability, but also due to limitations in the way the current models represent Earth’s system. In this work, we assess one of the most significant modeling issues: systematic errors. We found systematic errors on the diurnal range of 2-m temperature that could hamper predictability from long-term memory soil moisture conditions, requiring further attention from model development.

KEYWORDS: Atmosphere-land interaction; Forecast verification/skill; Model errors

1. Introduction

Subseasonal forecasts lie between medium-range and seasonal time scales. They follow the medium-range weather forecasts that depend essentially on the atmospheric initial conditions, and precede the seasonal projections, driven by the slowly evolving boundary conditions (e.g., land and ocean conditions). At the subseasonal scale, much of the impact of the initial atmospheric conditions has been lost while the boundary conditions do not exert a strong influence as in longer time scales. Due to this conjunction of factors, the

subseasonal scale has been previously referred to as a “predictability desert” (Vitart et al. 2012).

Forecasts in the subseasonal time scale have been assessed by the research community and in several operational weather forecasts centers in recent years (Pegion et al. 2019; Vitart et al. 2017). This emerging attention can be explained by the relevance of these forecasts for society and by the scientific challenges involved (White et al. 2017; Merryfield et al. 2020). The challenges of capturing and representing key processes and teleconnections that are prominent at these scales are significant: forecasting temperature extremes associated with weather extremes like heatwaves and droughts (Lavaysse et al. 2019; Wulff and Domeisen 2019; Magnusson et al. 2018) that can have severe consequences in nature and human health. Limitations in forecast skill can arise from the limits of predictability of the chaotic Earth system (Lorenz 1969), as well as from errors in the numerical models (Robertson et al. 2015; Vitart et al. 2019). There are several potential sources of predictability at the subseasonal scale, for example, the Madden–Julian oscillation (Kim et al. 2018) or the land surface (Ardilouze et al. 2017; Orsolini et al. 2013; Prodhomme et al. 2016).

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/MWR-D-20-0342.s1>.

Corresponding author: Emanuel Dutra, emanuel.dutra@ipma.pt

DOI: 10.1175/MWR-D-20-0342.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

In this study we investigate systematic model biases in the ECMWF hindcasts, their evolution with lead time and potential links with forecast skill and other performance metrics, focusing on surface variables (temperature and precipitation). The reference dataset was the fifth-generation ECMWF atmospheric reanalysis (ERA5) and the study was performed over the Northern Hemisphere midlatitudes in late spring and summer. There are several studies focusing on the skill of subseasonal forecasts, in particular taking advantage of the Subseasonal-to-Seasonal (S2S) Prediction Project database (e.g., Albers and Newman 2019; de Andrade et al. 2019; Vigaud et al. 2019; Zhou et al. 2019; Wulff and Domeisen 2019). In this study we take a different view of the forecasts by evaluating in detail the biases and their evolution with forecast lead time. Systematic model errors are a long-standing issue in the numerical weather prediction (NWP) and climate modeling communities (Zadra et al. 2018; Merryfield et al. 2020), which has been further highlighted in a recent survey carried out by the Working Group for Numerical Experimentation (WGNE) of the World Meteorological Organization (WMO) (Reynolds et al. 2019).

In the following section, we briefly describe the data (reanalyses and hindcasts) used and the methods (performance metrics). Section 3 presents the main results followed by the discussion in section 4 with the main conclusions in section 5.

2. Data and methods

a. Data

In this study, an 11-member ensemble using the ECMWF extended-range forecast system was evaluated. The hindcasts (or reforecasts) extend for 6 weeks, starting every 7 days, from 9 April to 30 July, for a 20-yr period (1998–2017). This is a similar setup to the current ECMWF operational ensemble prediction system, mainly differing in the horizontal resolution in the atmosphere (triangular–cubic–octahedral TCo with a spectral truncation of 199 vs 639) and ocean ($1^\circ \times 1^\circ$ vs $0.25^\circ \times 0.25^\circ$). The forecasts are initialized from ERA5 and the evaluation is performed over the Northern Hemisphere midlatitudes considering weekly means. The operational ECMWF hindcasts were not used to guarantee model and initialization consistency, as the operational forecasts had changes in June 2019/20.

ERA5 is the reference dataset used in this study. ERA5 is the latest global atmospheric reanalysis produced by ECMWF (Hersbach et al. 2020). It is a product of a decade of model developments and data assimilation innovations that replaced the previous reanalysis, ERA-Interim, in 2019. ERA5 is based on the 2016 version of the ECMWF model, with a horizontal resolution of about 31 km (triangular–linear with spectral truncation of 639-TL639) and 137 vertical layers (reaching 0.01 hPa at the top of the atmosphere). ERA5 also has a 10-member ensemble of data assimilation at a coarse spatial resolution (about 63 km) which can be used to derive uncertainty estimates. However, in this study, only the high-resolution analysis was used. ERA5 data were also processed for weekly means to compare with the forecasts. Although ERA5 can share some of the model biases, it is still a good reference

dataset due to its land and atmospheric data assimilation. In particular, the 2-m temperature analyses in ERA5 are strongly constrained by in situ observations using a two-dimensional optimal interpolation (de Rosnay et al. 2014; Douville et al. 1998). Moreover, ERA5 precipitation has been shown to also have a good quality (Beck et al. 2019; Nogueira 2020; Tarek et al. 2020).

In addition to ERA5, a land surface–only simulation similar to ERA5-Land (Muñoz Sabater 2019) was also compared to assess the influence of data assimilation in the initialization of the surface fields and their evolution with forecast lead time. ERA5-Land is a global land surface dataset at 9-km resolution, driven by ERA5 near-surface meteorology and fluxes. The land surface simulation was performed at the same resolution as the forecasts (TCo199 ~ 50-km resolution) as it prevents any interpolation-derived problems when compared with ERA5-Land 9-km resolution. Compared with ERA5, which was used to initialize the forecasts, the key difference is that the simulation has a free running land surface that is not constrained by the land data assimilation.

The following atmospheric and land variables were assessed in this work: daily mean 2-m temperature (t2m), daily maximum t2m (mx2t), daily minimum t2m (mn2t), evaporation (e), runoff (ro), total precipitation (tp) and soil moisture index (smi). Special attention was given to the temperature daily extremes and total precipitation. The daily data were averaged to weekly means, therefore mx2t represents the weekly mean of daily maximum temperature (not the maximum temperature over the 7-day period). The soil moisture index is computed by normalizing the top-meter soil moisture (top three soil layers, normally associated to the vegetation root zone) between field capacity and wilting point [$\text{smi} = (\text{soil moisture} - \text{wilting point}) / (\text{field capacity} - \text{wilting point})$]. This calculation is performed at the gridpoint resolution before interpolating to a regular grid of $1^\circ \times 1^\circ$, used in the evaluation, to avoid interpolation errors associated with different soil textures. The smi is normally between 0 (dry, at wilting point) and 1 (wet, at field capacity), but values below 0 or above 1 are possible as soil moisture can fall below wilting point or be above field capacity. The smi can be interpreted as a proxy to soil moisture stress to evaporation, which is independent from spatially varying soil textures. Terrestrial water storage variation (TWSV) was computed from the surface water fluxes (precipitation – evaporation – runoff) providing an integrated measure of the surface water budget.

b. Methods

Various metrics were used to evaluate the hindcasts performance: bias (and relative bias), the root-mean-square error (RMSE), the anomaly correlation coefficient (ACC), the Brier score (BS) and the signal-to-noise ratio (SNR). These metrics address different characteristics of the forecasts. The Bias assesses the systematic differences between the mean forecast and mean observations, while the RMSE measures the average magnitude of the forecast errors. The ACC measures the correspondence between the forecast ensemble mean and observations, independently of the bias, providing a skill measure of the forecasts. The Brier score measures the mean squared

probability error, being a forecast accuracy measure. Finally, the SNR measures the size of the predictable signal relative to the unpredictable chaos in the forecasts ensemble (Eade et al. 2014), providing some guidance on the expected skill of the predictions (Kumar 2009). Below 1 SNR is associated with large ensemble noise, while SNR above 1 indicates consistency between the ensemble members (high signal). SNR is commonly used to investigate potential predictability, where regions with high (low) SNR tend to have higher (lower) predictability (Ehsan et al. 2020b; Saha et al. 2016; Ehsan et al. 2020a). A detailed summary of the metrics' computation is available in the [appendix](#).

In addition to hemispheric maps, the results were also aggregated by regions (see Fig. 1). The European regions were based on the areas defined by Wulff and Domeisen (2019): Scandinavia (SC), western Europe (WEU), eastern Europe (EEU), Russia (RUK), western Mediterranean (WMED), and eastern Mediterranean (EMED). Two additional regions—the United States (covering mostly central North America) and east of the Caspian Sea (CASP)—were selected due to their large systematic temperature biases (see Fig. 1). When computing regional means, a land–sea mask was applied to the data to consider only land points, and the spatial maps only show results over land points. The results were organized in two periods: April–May and June–July start dates, to differentiate late spring from summer forecasts, which have different temperature biases (see Fig. 2).

3. Results

The daily maximum and minimum temperature forecast biases for weeks 1, 3, and 5 (Fig. 2) show several large-scale patterns that tend to be amplified with forecast lead time and differ between the forecasts initialized in April–May and June–July. For mx2t there is a widespread cold bias in the entire Northern Hemisphere midlatitudes for the April–May starting dates, while for June–July starting dates there is a neutral to positive bias over the central United States and east of the Caspian Sea that is amplified with lead time. The bias patterns for mn2t differ from those of mx2t, but there is still some cold bias in the April–May starting dates while for June–July there is a clear warm bias in the continental United States and Europe. The results of the RMSE for mx2t and mn2t with the same organization as Fig. 2 are available in Fig. S1 in the online supplemental material. The RMSE increases with forecast lead time as it accounts for both the systematic errors and the errors variance due to the decrease in predictability. The mx2t errors are larger than those of mn2t. This is particularly evident for longer lead times in northern Eurasia, a region with a strong cold bias (see Fig. 2c). These errors are likely associated with snowmelt and soil freezing differences between ERA5 and the forecasts. Since this region and cold process (snowmelt and soil thawing) are not strongly constrained by the data assimilation in ERA5, these errors are not further explored in this study. However, these errors are worth a closer investigation to understand their sources.

The soil moisture index differences between the forecasts and ERA5 also present large-scale patterns (see supplemental

June–July week 6 forecast mn2t bias

FIG. 1. Spatial extent of the regions used in this study. The underlying map shows the 2-m daily minimum temperature bias (K) of the forecasts initialized in June–July for week 6 lead time.

material Fig. S2), which grow with forecast lead time. For total precipitation, the region to the east of the Caspian Sea presents a clear dry bias (see Fig. S3), while there is an indication of a wet bias in the region affected by the East Asia monsoon in the June–July forecast at longer lead times (Fig. S3f). The bias evolution as a function of lead time for the daily minimum, mean, and maximum temperature, as well as for total precipitation and soil moisture index averaged over the eight regions (Fig. 3), summarizes some of the key temporal–spatial bias patterns: (i) systematic cold mx2t biases in the April–May forecasts at all lead times in all regions except the United States and CASP; (ii) the United States with a warm bias mostly in mn2t; (iii) CASP region with a general warm and dry bias; (iv) WEU, EMED, and WMED with a cold mn2t bias mainly in April–May forecasts and (v) EEU, SC and RUK with a cold mx2t and warm mn2t biases in the June–July forecasts. A detailed evaluation of the biases and their evolution with lead time is presented in the discussion section supported by time series of the mean forecast evolution and ERA5.

The SNR for mx2t, mn2t, tp, and smi for the forecasts initialized in April–May and June–July is shown in Fig. 4. SNR is higher in week 1 in all variables and regions, which is expected due to the memory of the atmospheric initial conditions and reduced ensemble spread, when compared with long lead times. In week 1, the SNR is, in general, higher in April–May than in June–July for the daily temperature extremes and precipitation. There is a clear annual cycle in the forecast skill of both the deterministic 500-hPa geopotential height and in the 850-hPa temperature ensemble forecasts, showing higher skill in winter when compared with summer in the Northern Hemisphere midlatitudes (Haiden et al. 2019). This could explain the higher SNR in April–May forecasts in week 1 compared with June–July. The higher predictability in late spring,

FIG. 2. Temperature biases (K) on (left) week 1, (center) week 3, and (right) week 5 for daily maximum temperature (mx2t) forecasts initialized in (a)–(c) April–May and (d)–(f) June–July and for daily minimum temperature (mn2t) forecasts initialized in (g)–(i) April–May and (j)–(l) June–July.

FIG. 3. Forecast ensemble-mean bias for each week lead time (horizontal axis) and region (colored lines) organized by the forecasts initialized in April–May and June–July for (a),(b) t2m; (c),(d) mx2t; (e),(f) mn2t; (g),(h) tp; and (i),(j) smi.

when compared with early summer, in the midlatitudes is likely associated with prevailing synoptic activity that is well captured by the initial conditions that dominate the skill and SNR of the forecasts in week 1.

By week 2, the SNR decreases to values near 1 for both mn2t and mx2t. Precipitation shows SNR much lower than temperature in week 1, falling to values below 1 in week 2. By contrast, the soil moisture index presents higher values overall, with some regions showing SNR above 1 up to week 6. This shows the memory effect of the soil moisture initial conditions providing a predictable signal of soil moisture beyond the first 2 weeks (Dirmeyer et al. 2018). The forecasts initialized in June–July tend to have a higher soil moisture index SNR when compared with the April–May forecasts. This can be attributed to mean drier climate during those months allowing for the initial conditions signal to persist with forecast lead time.

Forecast skill was assessed via the ACC shown in Fig. 5. Both mn2t and mx2t have ACCs above 0.4 up to week 2, with a drop of skill from week 3 onward, consistently in all regions. Similar results are found for precipitation, but with the ACC in week 2 comparable with the ACC of temperature in week 3. The maximum temperature presents higher values (around 0.9) and more consistency between regions, for both April–May and June–July starting dates. The minimum temperature shows values between 0.8 and 0.9 for both periods, while the precipitation ACC is between 0.7 and 0.8 (0.6 and 0.7) in April–May (June–July) start dates. In week 2, there is a larger dispersion between regions in all variables. The maximum (minimum) temperature has values between 0.5 and 0.7 (0.4 and 0.6) for both periods. The precipitation ACC is smaller, with values around 0.2 and 0.3. From week 3 onward, the ACC is, in

general, lower than 0.2 for both mn2t and mx2t and is close to 0 for precipitation, with several regions not showing statistical significance. The difference in ACC between mn2t and mx2t can be primarily attributed to local effects associated with the development of stable boundary layers during nighttime, which are challenging to parameterize (Holtlag et al. 2013). Therefore, it is expected that mn2t is more challenging to be represented in the model than mx2t, with implications on forecast skill. The drop in skill from week 1 to weeks 2 and 3 in all regions and variables analyzed are mostly driven by a drop in the atmospheric predictability. This is evidenced in the SNR results (see Fig. 4) with a clear drop of the SNR from week 1 to week 2. The ACC of the soil moisture index shows a much smaller reduction with lead time, with ACCs mostly above 0.5 in all regions in week 6 with the exception of Scandinavia in the April–May forecasts. These results are consistent with the SNR, which also showed a lower SNR for smi during the April–May forecasts. This can be primarily attributed to snowmelt that still affects some regions in Scandinavia, impacting soil moisture evolution.

The Brier score for both temperature extremes, percentiles, and periods in week 1 is very similar (between 0.06 and 0.09) in all regions (see Fig. S4). For precipitation, the values are slightly higher (between 0.09 and 0.14). In week 2, the BS values are once again very similar, between 0.13 and 0.17 for the temperature extremes and between 0.15 and 0.20 for the precipitation. From week 3 onward, the BS is above its reference value (Brier score of a climatological forecast ~ 0.2 for below the 25th and above the 75th percentile) in most regions and for all variables. There is no clear difference between high and low extremes in the different regions, despite the distinct biases

FIG. 4. Signal-to-noise ratio of the ensemble forecasts as function of lead time (weeks) for each region (colored lines) for the forecasts initialized in (left) April–May and (right) June–July for (a),(b) mx2t; (c),(d) mn2t; (e),(f) tp; and (g),(h) smi. Please note the different ranges of the vertical axes for the different variables. The SNR equal to 1 is highlighted in dashed black.

with lead time. Similar results can be seen for precipitation, with the exception being the CASP region, having a constant BS value for the 25th percentile in June–July of ~ 0.15 , considerably lower than the BS in the other regions in weeks 2 and 3. This is associated with the dry bias of the model in the region, resulting in BS of forecasts below the 25th percentile with higher skill, when compared with other regions.

The forecast metrics used above are stratified by the forecast's biases, to assess possible relations between forecast biases and skill, accuracy, and predictability. Systematic biases affect the model mean state that could drive nonlinear effects on the land–atmosphere coupling (Williams et al. 2016). Therefore, forecast biases could affect skill. However, such effects might be restricted to particular events or dominated by

FIG. 5. Anomaly correlation coefficient (ACC) of the ensemble forecasts, using ERA5 as reference, as function of lead time for each region (colored lines) for the forecasts initialized in (left) April–May and (right) June–July of (a),(b) mx2t; (c),(d) mn2t; (e),(f) tp; and (g),(h) smi. Crosses indicate that the ACC is different from 0 with 99% confidence following a two-sided t test for the null hypothesis that the expected ACC is zero.

atmospheric predictability. Figure 6 displays the bias versus ACC scatterplots for mx2t, mn2t, and tp in weeks 1 and 2 lead time, distinguishing the April–May and June–July starting dates. For each variable and forecast lead time, the relation between bias and ACC in all regions and April–May and June–July starting data ($8 \times 2 = 16$ points) is assessed via the Spearman rank-order correlation, displayed in the top tile of

each panel in Fig. 6. The results indicate that the ACC of mx2t in week 2 tends to be higher in regions with lower biases (Fig. 6d) with the only significant rank correlation of 0.65. The bias versus SNR only indicates some relation for the mx2t in week 1 (Fig. 7a), but with a negative rank correlation of -0.69 . This suggests that regions with higher cold biases have a higher potential predictability. However, all values of SNR for mx2t

FIG. 6. Forecast bias (x axis) vs ACC (y axis) of the forecasts initialized in April–May (circles) and June–July (crosses), for the (a)–(c) week 1 and (d)–(f) week 2 forecasts lead times of (a),(d) mx2t; (b),(e) mn2d; and (c),(f) tp. In each panel title, SR indicates the Spearman rank-order correlation coefficient with an asterisk (*) denoting significance at 99% confidence.

in week 1 are above 3, which is already a very high indication of potential predictability in the ensemble. There is no strong relation between the biases and Brier score for forecasts below the 25th percentile (Fig. S5) and above the 75th percentile (Fig. S6). For mx2t the biases are always negative in weeks 1 and 2, but for mn2t and tp there are positive and negative biases (below and above 100 in the case of tp). We also computed the rank correlations using the absolute biases to test if there could be some relation between the magnitude of the biases and forecast metrics. However, that did not change the results; i.e., no relation for mn2t and tp was found. The relationship between bias and the different forecast performance metrics are based on a very small sample (only 16 points). Different spatial aggregations were tested, even considering each grid point independently, without presenting any further robust relationships. Therefore, further investigations are required to understand if the relationships found for mx2t and ACC in week 2 and SNR in week 1 are robust. These results do not identify a general and consistent relationship between biases and skill (ACC), accuracy (BS) or potential predictability (SNR) of the forecasts in the different regions. However, the relationship between mx2t and ACC in week 2

and SNR in week 1 provide some guidance for further investigations, in particular, to explore the physical processes responsible for the systematic cold biases.

4. Discussion

This study identified the CASP region with larger temperature biases that are associated with a dry bias (see Fig. 8). The CASP region encompasses the southwestern part of central Asia (covering Turkmenistan, Uzbekistan, and Kazakhstan). It is an arid area of central Asia, with most of it having a cold desert climate (BWk in the Köppen climate classification), characterized by very warm and dry summer months (Jiang et al. 2019). central Asia is one of the regions that stand out in this study, due to its very distinct results across the various metrics. There is coherence between the bias, and SNR in this region for precipitation (negative bias and low SNR). Also, the positive bias in both temperature extremes seems to relate to low SNRs in June–July. The dry bias might trigger the warm bias, as less precipitation means less soil moisture to evaporate, which in turn might cause the warming of the surface and the air above it. This is visible in the time series in Fig. 8 with a drift

FIG. 7. As in Fig. 6, but for the signal-to-noise ratio.

of the smi forecasts with lead time from ERA5 to drier conditions. The ERA5 smi initial conditions are drier than ERA5-Land, with a notable difference in the terrestrial water storage variation which is much higher (negative) in ERA5 than in ERA5-Land. It is also important to mention that the CASP region has a low density of weather stations, due to its low population density, mostly because of the arid climate in the region. Additionally, the number of stations suffered a reduction in the early 1990s (Borovikova 1997) and it has yet to recover, yet ERA5 is one of the reanalyses with better results in terms of precipitable water vapor in central Asia (Jiang et al. 2019). However, the lack of observations may result in less accurate representations of the climate in central Asia, limiting the use of ERA5 as a reference dataset for the evaluation of bias and skill of the ensemble forecasts.

The U.S. domain is another region depicted in this study that stands out for its results. This region is mostly comprised by the Great Plains, which is a well-known area with weather and climate models showing warm and dry biases (Lin et al. 2017; Morcrette et al. 2018; Ardilouze et al. 2019). In this work, it is shown that the ECMWF hindcasts have similar warm biases to previous studies, mainly on the daily minimum temperature. The forecasts drift from ERA5 smi to wetter conditions in May–June and drier conditions in July–August

approaching ERA5-Land smi (see Fig. 9). This is due to negative soil moisture increments in May–June in ERA5 and positive during July–August. These results suggest that the cold mx2t bias in May–June and warm bias in July–August visible in the forecasts is present right at the start of the short-range forecasts of ERA5 explaining the soil moisture increments and differences in respect to ERA5-Land. Although the representation of precipitation in reanalyses has uncertainties, it has been shown that, over the United States, ERA5 can reproduce the interannual variability of precipitation reasonably well (Beck et al. 2019; Tarek et al. 2020) and similarly for surface soil moisture (Beck et al. 2021). Thus, it is likely that the warm bias reported in this study has a negligible relation with precipitation or surface soil moisture errors. The U.S. region, just like CASP, has lower SNR values for the maximum temperature and precipitation, displaying less coherence between the ensembles. The relatively lower SNR starting in week 1 is an indication of noise in the ensemble that will hamper predictability in the following lead times.

The European regions have, in general, smaller biases than the aforementioned regions. All regions have systematic biases in the temperature extremes that do not seem to have a direct impact on forecast skill, and the variability is similar to ERA5.

FIG. 8. Mean weekly climate over CASP region of the forecasts for each starting date (blue), ERA5 (red), and ERA5-Land (orange). The variables displayed are as follows: (a) total precipitation (mm day^{-1}), (b) evaporation (mm day^{-1}), (c) runoff (mm day^{-1}), (d) soil moisture index (0–1), (e) 2-m daily mean temperature (K), (f) 2-m daily maximum temperature (K), (g) 2-m daily minimum temperature (K), and (h) terrestrial water storage variation (mm day^{-1}).

ACC and BS are very similar in all European regions as well, which is consistent with the previous results of Wulff and Domeisen (2019) using the S2S database.

The continental European regions (SC, EEU, and RUK) show a consistent cold mx2t bias in May–June with drier smi in ERA5 than in ERA5-Land with the forecasts drifting from the initial conditions to ERA5-Land state (see Figs. S7–S9). The TWSV in this May–June period is also smaller in ERA5 and the forecasts than in ERA5-Land. This is related with the water removal in the root-zone soil moisture by the land data assimilation in ERA5 to compensate for the cold temperature bias. Scandinavia region also shows large differences in terms of runoff and TWSV between ERA5 and the forecasts and ERA5-Land that are likely associated with snow mass removal by ERA5 data assimilation, which is known to affect northern

basins river discharge (Zsoter et al. 2019). In July–August, the forecasts show a warm minimum temperature bias, which combined with the cold maximum temperature bias, results in an underestimation of the diurnal cycle amplitude.

Contrasting with the continental European regions, the Mediterranean and western Europe regions (WEU, WMED, and EMED) do not present such large smi drifts in the forecast when compared with ERA5 (Figs. S10–S12). However, smi in ERA5 is lower than ERA5-Land, in all regions also showing a persistent daily maximum temperature bias in May–June (from the April–May forecasts), which is mostly negligible in July–August. Johannsen et al. (2019) found a large cold bias of maximum land surface temperature (LST) in ERA5 over the Iberian Peninsula during summer when compared with satellite LST estimates. These biases were

FIG. 9. As in Fig. 8, but for the U.S. region.

linked to vegetation cover in ERA5 (Nogueira et al. 2020), and are likely to also affect ERA5 2-m temperature. Therefore, some caution must be taken when considering ERA5 as a reference dataset, in particular over small areas as some of the signals might be linked with errors in ERA5 as well as different in situ observations density used in the data assimilation. Despite the limitations of using ERA5 as the reference dataset for the forecasts evaluation, the evaluation of near-surface parameters with in situ observation is affected by several issues associated with representativeness errors (ben Bouallegue et al. 2020).

5. Conclusions

This study is focused on surface-related variables during late spring and summer in the Northern Hemisphere midlatitudes, aiming to (i) document the development of systematic errors with lead time in the ECMWF ensemble forecasts and (ii) investigate potential relations between the systematic errors and predictive skill. The biases evolution as a function of lead time for the daily temperature extremes, precipitation and soil moisture index revealed five key temporal–spatial bias patterns:

- Systematic cold bias of daily maximum temperature in the April–May forecasts at all lead times in all regions except the United States and CASP;
- The United States with a warm bias mostly in the daily minimum temperature;
- CASP region with a general warm and dry bias;
- Western and Mediterranean Europe with a cold bias in daily minimum temperature mainly in April–May forecasts;
- Continental Europe with a cold bias in the daily maximum temperature and warm bias of daily minimum temperature in the June–July forecasts, resulting in an underestimation of the diurnal cycle amplitude;

We also found substantial deviations of the soil moisture evolution with forecast lead time from ERA5 state to conditions closer to ERA5-Land. Further diagnostics including soil moisture increments in ERA5 are required to disentangle the effects of land data assimilation from forecasts biases.

Despite the seasonal dependence of the systematic biases, we did not find such a dependence on forecast skill, nor a robust relationship between biases and forecast skill. The main conclusion is, while there exist large differences in the systematic error characteristics, there is little relation to the skill for the subseasonal forecasts. However, these results do not reject the hypothesis that systematic biases affect forecast skill. Such a test would require a controlled set of forecasts with reduced systematic biases (Ardilouze et al. 2019). Despite this, the general and systematic cold maximum daily temperature and warm minimum daily temperature biases require further attention from model developers (Beljaars 2020; Nogueira et al. 2020). Reducing this underestimation of the diurnal temperature range through model development is likely to enhance some of the potential predictability coming from the long-memory effect of root-zone soil moisture conditions (Dirmeyer 2005; Koster et al. 2011).

Acknowledgments. This research was funded by Fundação para a Ciência e a Tecnologia (FCT) Grant PTDC/CTA-MET/28946/2017 (CONTROL). The authors would also like to acknowledge the financial support of FCT through project UIDB/50019/2020–IDL. The authors thank David Richardson, Gianpaolo Balsamo, and three anonymous reviewers for their comments and suggestions that helped to improve the manuscript. Acknowledgement is made for the use of ECMWF's computing and archive facilities in this research.

Data availability statement. ERA5 data can be obtained freely from the Copernicus Climate Change Service Information website (<https://climate.copernicus.eu/>). The simulations carried out in this study are available in the ECMWF data archive (required login at: <https://apps.ecmwf.int/mars-catalogue/>); hindcasts expver=h80p, class=rd; land surface simulation: expver=a04i, class=pt.

APPENDIX

Forecast Skill Metrics

This section presents the calculation details of the different metrics used in this study with the following description:

- Bias represents the difference between the average forecast ensemble mean and the average of the reference product, and the relative bias (used for precipitation) normalizes the bias by the mean reference with 100 representing zero bias. The root-mean-square error (RMSE) measures the mean magnitude of the forecast errors [Eq. (1)];
- Anomaly correlation coefficient (ACC) evaluates the skill of the ensemble mean at reproducing the reference product spatial pattern anomaly [Eq. (2)];
- Brier score (BS) evaluates the accuracy of the ensemble in forecasting a specific event, in this study, below the 25th percentile or above the 75th percentile, by computing the mean square difference between the probability of forecasting the event Y_k (between 0 and 1) and the actual outcome of the event O_k (0 or 1) [Eq. (3)]. The percentiles defining the event are computed independently for the forecasts (P_y) and for the reference product (P_o) [Eq. (5)];
- Signal-to-noise ratio (SNR) measures the coherence between the different ensemble members [Eq. (4)].

In Eqs. (A1)–(A5) y and o represent the forecast and the reference product (ERA5 in this study), respectively; N , J , and M represent the number of years (20), the number of ensemble members (11), and the number of grid points in a particular region, respectively; and $\bar{y}$ is the ensemble temporal mean, $\bar{y}_k$ is the ensemble mean, and $\bar{o}$ the reference temporal mean. In the ACC calculation, y'_m and o'_m are anomalies relative to the climatological averages, and $\bar{y}'$ and $\bar{o}'$ are these anomalies averaged over a given region with M grid points.

The different metrics are four-dimensional (start date, lead time, latitude, longitude), except for the ACC which is only two dimensional (start date, lead time), and all are averaged over the total number of start dates (split between April–May and June–July). A regional mean is applied for all metrics (except

ACC) for each forecast lead time. For the ACC, the Fisher Z transform is applied before the temporal averages and then inverted to the correlation, which will represent the average ACC over a given region. A two-sided t test for the null hypothesis that the expected ACC is zero was performed on the Fisher Z transform:

$$\text{Bias} = \frac{1}{N} \sum_{k=1}^N \bar{y}_k - o_k,$$

$$\text{Relative bias} = 100 + 100 \times \frac{\text{Bias}}{\frac{1}{N} \sum_{k=1}^N o_k},$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^N (\bar{y}_k - o_k)^2}; \quad (\text{A1})$$

$$\text{ACC} = \frac{\sum_{m=1}^M (y'_m - \bar{y}') (o'_m - \bar{o}')}{\left[\sum_{m=1}^M (y'_m - \bar{y}')^2 \sum_{m=1}^M (o'_m - \bar{o}')^2 \right]^{1/2}}; \quad (\text{A2})$$

$$\text{BS} = \frac{1}{N} \sum_{k=1}^N (Y_k - O_k)^2; \quad (\text{A3})$$

$$\text{SNR} = \sqrt{\frac{\frac{1}{N-1} \sum_{k=1}^N (\bar{y}_k - \bar{y})^2}{\frac{1}{N \times J - N} \sum_{k=1}^N \sum_{j=1}^J (y_{kj} - \bar{y}_k)^2}}; \quad (\text{A4})$$

$$\bar{y} = \frac{1}{N \times J} \sum_{k=1}^N \sum_{j=1}^J (y_{kj}); \quad \bar{y}_k = \frac{1}{J} \sum_{j=1}^J (y_{kj}),$$

$$Y_k = \frac{1}{J} \sum_{j=1}^J f(y_{kj}, P_y); \quad O_k = f(o_k, P_o),$$

$$f(x, P) = \begin{cases} 1 & x > P \\ 0 & x < P \end{cases}, \quad P \text{ is the 75th percentile,}$$

$$f(x, P) = \begin{cases} 1 & x < P \\ 0 & x > P \end{cases}, \quad P \text{ is the 25th percentile.} \quad (\text{A5})$$

REFERENCES

- Albers, J. R., and M. Newman, 2019: A priori identification of skillful extratropical subseasonal forecasts. *Geophys. Res. Lett.*, **46**, 12 527–12 536, <https://doi.org/10.1029/2019GL085270>.
- Ardilouze, C., and Coauthors, 2017: Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. *Climate Dyn.*, **49**, 3959–3974, <https://doi.org/10.1007/s00382-017-3555-7>.
- , L. Batté, B. Decharme, and M. Déqué, 2019: On the link between summer dry bias over the U.S. Great Plains and seasonal temperature prediction skill in a dynamical forecast system. *Wea. Forecasting*, **34**, 1161–1172, <https://doi.org/10.1175/WAF-D-19-0023.1>.
- Beck, H. E., and Coauthors, 2019: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol. Earth Syst. Sci.*, **23**, 207–224, <https://doi.org/10.5194/hess-23-207-2019>.
- , and Coauthors, 2021: Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors. *Hydrol. Earth Syst. Sci.*, **25**, 17–40, <https://doi.org/10.5194/hess-25-17-2021>.
- Beljaars, A., 2020: Towards optimal parameters for the prediction of near surface temperature and dewpoint. ECMWF Tech. Memo. 868, 44 pp., <https://doi.org/10.21957/yt64x7rth>.
- ben Bouallegue, Z., T. Haiden, N. J. Weber, T. M. Hamill, and D. S. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Wea. Rev.*, **148**, 2049–2062, <https://doi.org/10.1175/MWR-D-19-0323.1>.
- Borovikova, L. N., 1997: Description of the state of the National Hydrometeorological Surveys and concept for their future development. Rep. 4, World Bank Program, Improvement of Hydrometeorological Surveys in Central Asia, 57 pp.
- de Andrade, F. M., C. A. S. Coelho, and I. F. A. Cavalcanti, 2019: Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dyn.*, **52**, 5451–5475, <https://doi.org/10.1007/s00382-018-4457-z>.
- de Rosnay, P., G. Balsamo, C. Albergel, J. Muñoz-Sabater, and L. Isaksen, 2014: Initialisation of land surface variables for numerical weather prediction. *Surv. Geophys.*, **35**, 607–621, <https://doi.org/10.1007/s10712-012-9207-x>.
- Dirmeyer, P. A., 2005: The land surface contribution to the potential predictability of boreal summer season climate. *J. Hydrometeorol.*, **6**, 618–632, <https://doi.org/10.1175/JHM444.1>.
- , S. Halder, and R. Bombardi, 2018: On the harvest of predictability from land states in a global forecast model. *J. Geophys. Res. Atmos.*, **123**, 13 111–13 127, <https://doi.org/10.1029/2018JD029103>.
- Douville, H., J.-F. Mahfouf, S. Saarinen, and P. Viterbo, 1998: The ECMWF surface analysis: Diagnostics and prospects. ECMWF Tech. Memo., 258 pp., <https://doi.org/10.21957/8pikk2m8>.
- Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, <https://doi.org/10.1002/2014GL061146>.
- Ehsan, M. A., F. Kucharski, and M. Almazroui, 2020a: Potential predictability of boreal winter precipitation over central-southwest Asia in the North American multi-model ensemble. *Climate Dyn.*, **54**, 473–490, <https://doi.org/10.1007/s00382-019-05009-3>.
- , M. K. Tippett, F. Kucharski, M. Almazroui, and M. Ismail, 2020b: Predicting peak summer monsoon precipitation over Pakistan in ECMWF SEAS5 and North American Multimodel Ensemble. *Int. J. Climatol.*, **40**, 5556–5573, <https://doi.org/10.1002/joc.6535>.
- Haiden, T., M. Janousek, F. Vitart, L. Ferranti, and F. Prates, 2019: Evaluation of ECMWF forecasts, including the 2019 upgrade. ECMWF Tech. Memo. 853, 54 pp.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Holtslag, A. A. M., and Coauthors, 2013: Stable atmospheric boundary layers and diurnal cycles: Challenges for weather and climate models. *Bull. Amer. Meteor. Soc.*, **94**, 1691–1706, <https://doi.org/10.1175/BAMS-D-11-00187.1>.
- Jiang, J., T. Zhou, and W. Zhang, 2019: Evaluation of satellite and reanalysis precipitable water vapor data sets against radiosonde observations in Central Asia. *Earth Space Sci.*, **6**, 1129–1148, <https://doi.org/10.1029/2019EA000654>.

- Johannsen, F., S. Ermida, J. P. A. Martins, I. F. Trigo, M. Nogueira, and E. Dutra, 2019: Cold bias of ERA5 summertime daily maximum land surface temperature over Iberian Peninsula. *Remote Sens.*, **11**, 2570, <https://doi.org/10.3390/rs11212570>.
- Kim, H., F. Vitart, and D. E. Waliser, 2018: Prediction of the Madden–Julian oscillation: A review. *J. Climate*, **31**, 9425–9443, <https://doi.org/10.1175/JCLI-D-18-0210.1>.
- Koster, R. D., and Coauthors, 2011: The second phase of the Global Land–Atmosphere Coupling Experiment: Soil moisture contributions to subseasonal forecast skill. *J. Hydrometeorol.*, **12**, 805–822, <https://doi.org/10.1175/2011JHM1365.1>.
- Kumar, A., 2009: Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Mon. Wea. Rev.*, **137**, 2622–2631, <https://doi.org/10.1175/2009MWR2814.1>.
- Lavaysse, C., G. Naumann, L. Alfieri, P. Salamon, and J. Vogt, 2019: Predictability of the European heat and cold waves. *Climate Dyn.*, **52**, 2481–2495, <https://doi.org/10.1007/s00382-018-4273-5>.
- Lin, Y., W. Dong, M. Zhang, Y. Xie, W. Xue, J. Huang, and Y. Luo, 2017: Causes of model dry and warm bias over central U.S. and impact on climate projections. *Nat. Commun.*, **8**, 881, <https://doi.org/10.1038/s41467-017-01040-2>.
- Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, <https://doi.org/10.3402/tellusa.v21i3.10086>.
- Magnusson, L., L. Ferranti, and F. Vamborg, 2018: Forecasting the 2018 European heatwave. ECMWF Newsletter, No. 157, ECMWF, Reading, United Kingdom, 2–3.
- Merryfield, W. J., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. *Bull. Amer. Meteor. Soc.*, **101**, E869–E896, <https://doi.org/10.1175/BAMS-D-19-0037.1>.
- Morcrette, C. J., and Coauthors, 2018: Introduction to CAUSES: Description of weather and climate models and their near-surface temperature errors in 5 day hindcasts near the Southern Great Plains. *J. Geophys. Res. Atmos.*, **123**, 2655–2683, <https://doi.org/10.1002/2017JD027199>.
- Muñoz Sabater, J., 2019: First ERA5-Land dataset to be released this spring. ECMWF Newsletter, No. 159, ECMWF, Reading, United Kingdom, 8–9.
- Nogueira, M., 2020: Inter-comparison of ERA-5, ERA-interim and GPCP rainfall over the last 40 years: Process-based analysis of systematic and random differences. *J. Hydrol.*, **583**, 124632, <https://doi.org/10.1016/j.jhydrol.2020.124632>.
- , C. Albergel, S. Boussetta, F. Johannsen, I. F. Trigo, S. L. Ermida, J. P. A. Martins, and E. Dutra, 2020: Role of vegetation in representing land surface temperature in the CHTESSEL (CY45R1) and SURFEX-ISBA (v8.1) land surface models: A case study over Iberia. *Geosci. Model Dev.*, **13**, 3975–3993, <https://doi.org/10.5194/gmd-13-3975-2020>.
- Orsolini, Y. J., R. Senan, G. Balsamo, F. J. Doblas-Reyes, F. Vitart, A. Weisheimer, A. Carrasco, and R. E. Benestad, 2013: Impact of snow initialization on sub-seasonal forecasts. *Climate Dyn.*, **41**, 1969–1982, <https://doi.org/10.1007/s00382-013-1782-0>.
- Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SUBX). *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, <https://doi.org/10.1175/BAMS-D-18-0270.1>.
- Prodhomme, C., F. Doblas-Reyes, O. Bellprat, and E. Dutra, 2016: Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate Dyn.*, **47**, 919–935, <https://doi.org/10.1007/s00382-015-2879-4>.
- Reynolds, C., K. Williams, and A. Zadra, 2019: WGNE systematic error survey results summary. Accessed 15 October 2020, https://www.wcrp-climate.org/JSC40/12.7b.WGNE_Systematic_Error_Survey_Results_20190211.pdf.
- Robertson, A. W., A. Kumar, M. Peña, and F. Vitart, 2015: Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.*, **96**, ES49–ES53, <https://doi.org/10.1175/BAMS-D-14-00139.1>.
- Saha, S. K., and Coauthors, 2016: Potential predictability of Indian summer monsoon rainfall in NCEP CFSv2. *J. Adv. Model. Earth Syst.*, **8**, 96–120, <https://doi.org/10.1002/2015MS000542>.
- Tarek, M., F. P. Brissette, and R. Arsenault, 2020: Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrol. Earth Syst. Sci.*, **24**, 2527–2544, <https://doi.org/10.5194/hess-24-2527-2020>.
- Vigaud, N., M. K. Tippett, J. Yuan, A. W. Robertson, and N. Acharya, 2019: Probabilistic skill of subseasonal surface temperature forecasts over North America. *Wea. Forecasting*, **34**, 1789–1806, <https://doi.org/10.1175/WAF-D-19-0117.1>.
- Vitart, F., A. W. Robertson, and D. L. T. Anderson, 2012: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *Bull. WMO*, **61**, 23 pp.
- , and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>.
- , and Coauthors, 2019: Extended-range prediction. ECMWF Tech. Memo. 854, 60 pp., <https://doi.org/10.21957/pdivp3t9m>.
- White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, <https://doi.org/10.1002/met.1654>.
- Williams, I. N., Y. Lu, L. M. Kueppers, W. J. Riley, S. C. Biraud, J. E. Bagley, and M. S. Torn, 2016: Land-atmosphere coupling and climate prediction over the U.S. Southern Great Plains. *J. Geophys. Res. Atmos.*, **121**, 12 125–12 144, <https://doi.org/10.1002/2016JD025223>.
- Wulff, C. O., and D. I. V. Domeisen, 2019: Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers. *Geophys. Res. Lett.*, **46**, 11 520–11 529, <https://doi.org/10.1029/2019GL084314>.
- Zadra, A., and Coauthors, 2018: Systematic errors in weather and climate models: Nature, origins, and ways forward. *Bull. Amer. Meteor. Soc.*, **99**, ES67–ES70, <https://doi.org/10.1175/BAMS-D-17-0287.1>.
- Zhou, Y., B. Yang, H. Chen, Y. Zhang, A. Huang, and M. La, 2019: Effects of the Madden–Julian Oscillation on 2-m air temperature prediction over China during boreal winter in the S2S database. *Climate Dyn.*, **52**, 6671–6689, <https://doi.org/10.1007/s00382-018-4538-z>.
- Zsoter, E., H. Cloke, E. Stephens, P. de Rosnay, J. Muñoz-Sabater, C. Prudhomme, and F. Pappenberger, 2019: How well do operational numerical weather prediction configurations represent hydrology? *J. Hydrometeorol.*, **20**, 1533–1552, <https://doi.org/10.1175/JHM-D-18-0086.1>.