Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

873

Evaluation of biases and skill of ECMWF Summer sub-seasonal forecasts in the Northern Hemisphere

Frederico Johannsen[1], Linus Magnusson[2], Emanuel Dutra[1,3]

[1] Instituto Dom Luiz, IDL, Faculty of Sciences, University of Lisbon, Portugal
[2] ECMWF Forecast Department
[3] Instituto Português do Mar e da Atmosfera, IPMA, Portugal

November 2020

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our website under:
http://www.ecmwf.int/en/publications

Contact: library@ecmwf.int

Abstract

Sub-seasonal forecasts lie between medium-range and seasonal time scales with an emerging attention due to its relevance in society and by the scientific challenges involved. This report aims to (i) document the development of systematic errors with lead-time in ECMWF ensemble forecasts of surface-related variables during spring and summer, and (ii) investigate the relationship between the systematic errors and predictive skill. The evaluation has been performed over the northern hemisphere, focusing on several regions with different characteristics. The results indicate five key temporal/spatial bias patterns: (i) systematic cold bias of daily maximum temperature in the April-May forecasts at all lead times in most regions; (ii) USA with a warm bias mostly in the daily minimum temperature; (iii) East of Caspian Sea region with a general warm and dry bias; (iv) Western and Mediterranean Europe with a cold bias in daily minimum temperature mainly in April-May forecasts and (v) 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. The main conclusion is that while there exist large differences in the systematic error characteristics, there is little relation to the skill for the sub-seasonal forecasts. However, these results do not reject the hypothesis that systematic biases affect forecast skill. Despite this, the general and systematic cold maximum daily temperature and warm minimum daily temperature biases require further attention from model development as diurnal cycle improvements are likely to enhance some of the potential predictability coming from the long-memory effect of soil moisture conditions.

1 Introduction

Sub-seasonal 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 sub-seasonal 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 sub-seasonal scale has been previously referred to as a "predictability desert" (Vitart et al., 2012).

Forecasts in the sub-seasonal time scale have been assessed by the research community and in several operational weather forecasts centres 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). The challenges of capturing and representing key processes and teleconnections which are prominent at these scales are significant: forecasting temperature extremes associated with weather extremes like heatwaves and droughts (Lavaysse et al., 2019; Magnusson et al., 2018; Wulff & Domeisen, 2019) 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, Alonso-Balmaseda, et al., 2019). There are several potential sources of predictability at the sub-seasonal 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).

ECMWF also has a long experience in developing global ERA reanalyses. The reanalyses are created by the Integrated Forecasting System (IFS), a global data assimilation and forecasting system developed by ECMWF for weather forecasting. The most recent atmospheric reanalysis, ERA5, is available to the

public since 2019 (Hersbach et al., 2020). Global products combining observations and state-of-the-art model simulations, reanalyses are a key tool in scientific research as they can serve as a benchmark in climate studies. Furthermore, these reanalyses are also essential to provide consistent initial conditions to the reforecasts (Vitart, Balsamo, et al., 2019).

In this study we investigate systematic model biases in the ECMWF reforecasts, 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 ERA5 reanalysis and the study was performed over the Northern Hemisphere in late Spring and Summer. There are several studies focusing on the skill of sub-seasonal forecasts, in particular taking advantage of the Subseasonal-to-Seasonal (S2S) Prediction Project database (e.g. Albers & Newman, 2019; de Andrade et al., 2019; Vigaud et al., 2019; Zhou et al., 2019). However, it is not so common to investigate in detail model biases and their evolution with forecast lead time as accessed in this study.

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

2 Data and Methods

2.1 Data

In this study, an 11-member ensemble from an experimental setup of ECMWF extended-range forecast system (experiment h80p, class=rd) was evaluated. The hindcasts (or reforecasts) ran for 6 weeks, starting every 7 days, from April 9th to July 30th, for a 20-year period 1998-2017. This is a similar setup to the current operational ensemble prediction system, mainly differing in the horizontal resolution in the atmosphere (TCo199 vs TCo639) and ocean (1x1 vs 0.25x0.25). The forecasts are initialized from ERA5 and the evaluation is performed over the Northern Hemisphere considering weekly means.

The ERA5 reanalysis 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. Based on a 2016 version of the IFS (cycle 41r2), its horizontal resolution is about 31 km (TL639) and its vertical resolution has 137 layers (reaching 0.01 hPa at the top of the atmosphere). ERA5 data was also processed for weekly means to compare with the forecasts. Although ERA5 shares some of the model biases, it is still a good reference dataset due its land and atmospheric data assimilation. In particular, 2-meter temperature analysis are strongly constrained by in-situ observations.

In addition to ERA5, a surface experiment 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 surface experiment was carried out at the same resolution as the forecasts (TCo199 ~50 km resolution, expver=a04i, class=pt) as it prevents any interpolation-derived problems when compared with ERA5-Land 9 km resolution.

The following atmospheric and land variables were assessed in this work: daily mean 2-metre temperature (**t2m**), daily maximum t2m (**mx2t**), daily minimum t2m (**mn2t**), evaporation (**e**), runoff

(**ro**), total precipitation (**tp**) and soil moisture index (**smi**). Special focus was given to the temperature extremes and total precipitation. The daily data was averaged to weekly means, therefore mx2t represents the weekly mean of daily maximum temperature (not the maximum temperature over the 7 days period). The soil moisture index is computed by normalizing the top first metre soil moisture (top 3 soil layers, normally associated within the vegetation root-zone) between field capacity and wilting point (smi = (soil moisture - wilting point)/(field capacity - wilting point). This calculation is performed at the grid-point resolution before interpolating to a regular grid of 1x1 to avoid interpolation errors associated with different soil textures. 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.

2.2 Methods

Various metrics were used to evaluate the hindcasts performance: Bias (and relative bias), the Standard Deviation Ratio (SDR), the Anomaly Correlation Coefficient (ACC), the Brier Score and the Signal-to-Noise Ratio. These metrics access different characteristics of the forecasts. The Bias and Relative Bias represent the systematic errors, while the variability is accounted by the SDR. The ACC and Brier Score give information on the skill of the forecasts and the Signal-to-Noise Ratio shows the coherence between the different members of the ensemble. A detailed summary of the metrics' computation is available in Appendix I.

In addition to hemispheric maps, the results were also aggregated by regions (Figure A 1). The European regions were based on the areas defined by Wulff and Domeisen (2019) (SC: Scandinavia; WEU: Western Europe; EEU: Eastern Europe; RUK: Russia; WMED: Western Mediterranean; EMED: Eastern Mediterranean), while the USA and Caspian (CASP) regions were selected due to their large systematic temperature biases (see Figure 1). When computing regional means, a land-sea mask was applied to the data to consider only land points. The results were organized into two periods: April-May and June-July start dates, due to the different spatial patterns of the bias maps (see Figure 1).

3 Results

The daily maximum and minimum temperature forecast biases for weeks 1,3 and 5 (Figure 1) 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. The soil moisture index differences between the forecasts and ERA5 also present large-scale patterns (see Auxiliary Figure A 2 in Appendix II) while for total precipitation only the region to the East of the Caspian Sea presents a clear dry bias (see Figure A 3). The bias evolution as 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 8 regions (Figure 2), 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 USA and CASP; (ii) USA 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. In addition to the mean bias, the signal strength of the forecasts was accessed by computing the standard deviation ratio between the ensemble mean and ERA5 (Figure A 10). For both temperature and precipitation this ratio is mostly between 0.9 and 1.1 in all regions and lead times, indicating that the model's variability is similar to ERA5. The only exception is the CASP region with a clear decrease in precipitation variability with lead time for the forecasts initialized in June-July (Figure A 10 f). A detailed evaluation of the biases and their evolution with lead is presented in the discussion section supported by time series of the mean forecast evolution and ERA5 (e.g. Figure 6).

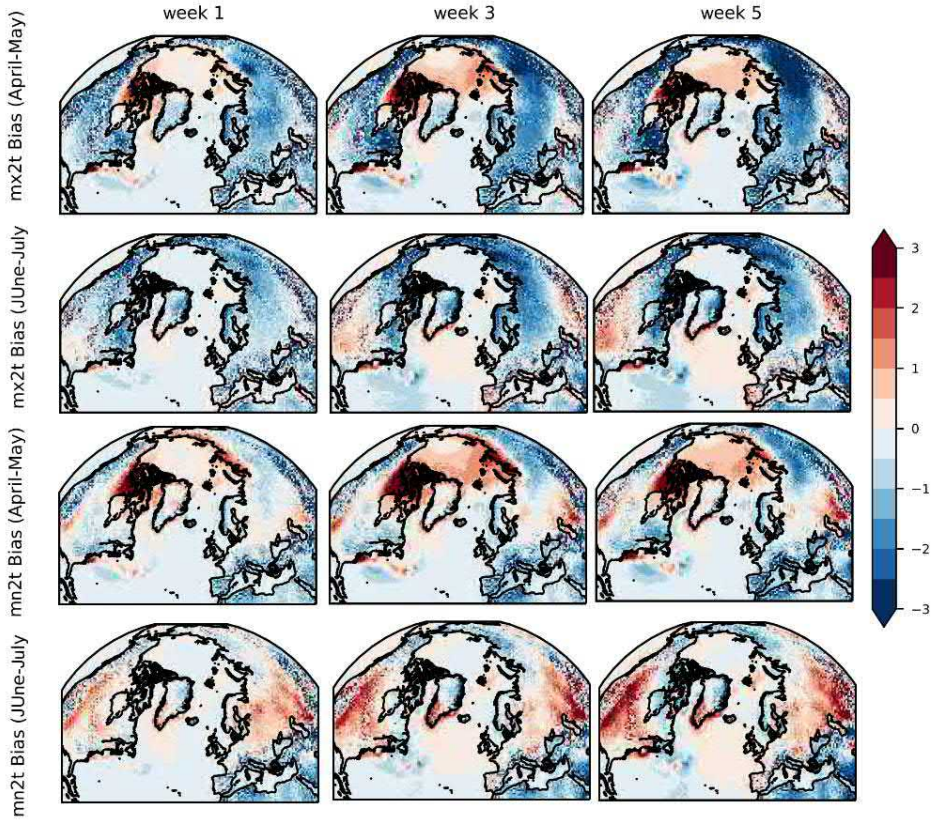

Figure 1- Temperature biases in week 1 (left) week 3 (centre) and week 5 (right) for daily maximum temperature forecasts initialized in April-May (first row) and June-July (second row) and for daily minimum temperature forecasts initialized in April-May (third row) and June-July (forth row).

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). SNR above 1 indicates consistency between the ensemble members (high signal) but does not imply higher skill. Below 1 SNR is associated with large ensemble noise. The SNR for mx2t, mn2t, tp and SMI for the forecasts initialized in April-May and June-July is shown in Figure 3. SNR is higher on week 1 in all variables and regions, which is expected due to the memory of the atmosphere initial conditions and reduced ensemble spread, when compared with long lead times. On week 1, the SNR is, in general, higher in the April-May than in June-July for the daily temperature extremes and precipitation. By week 2 the SNR decreases to values near 1 for both mn2t and mx2t. Precipitation

shows SNR much lower than temperature on week 1, falling to values below 1 on 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.

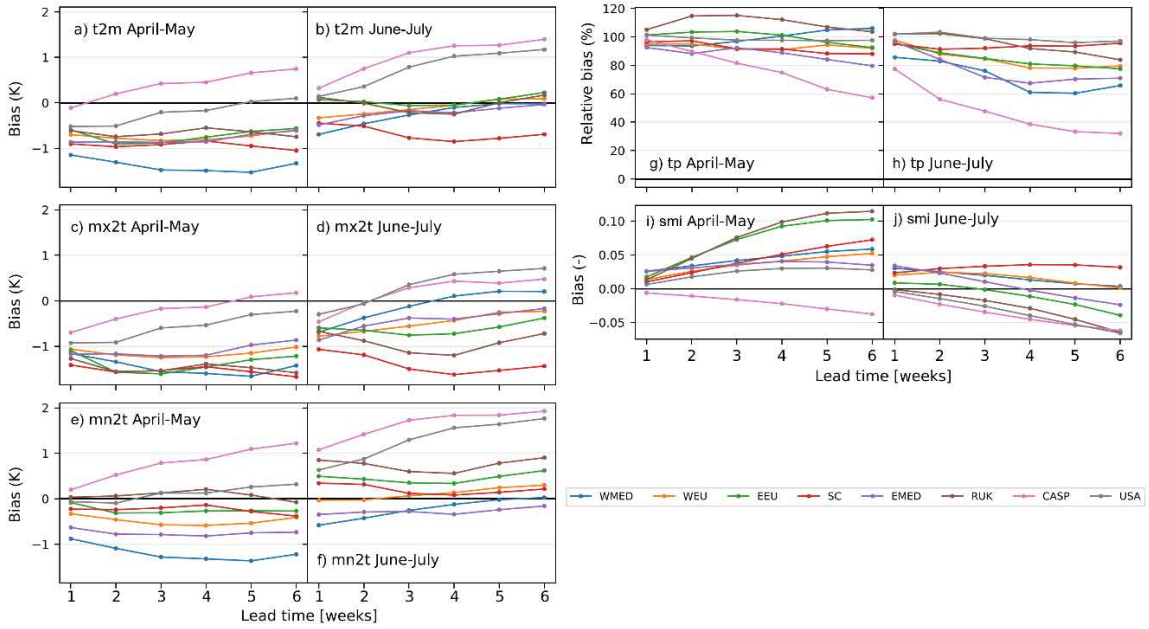

Figure 2- Forecast ensemble mean bias for each week lead time (horizontal axis) and region (color lines) organized by the forecasts initialized in April-May and June-July for t2m (a,b), mx2t (c,d), mn2t (e,f) tp (g,h) and smi (i,j).

Forecast skill was assessed via the anomaly correlation (ACC) shown in Figure 4. The ACC shows that the temperature forecasts are skillful up to week 2, with a drop of skill from week 3 onwards, consistently in all regions. Similar results are found for precipitation, but with the ACC on week 2 comparable with the ACC of temperature on week 3. On week 1, the ACC is high in all regions for the temperature extremes and precipitation. The maximum temperature presents higher values (around 0.9) and more consistency between regions, for both periods. 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. On 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 onwards, the ACC is, in general, lower than 0.2 for the temperature extremes and is close to 0 for precipitation. The difference in skill between mn2t and mx2t can be primarily attributed to local effects which are challenging to represent in the model.

The brier score for both temperature extremes, percentiles and periods on week 1 is very similar (between 0.06 and 0.09) in all regions (see Figure A 11). For precipitation, the values are slightly higher

(between 0.09 and 0.14). On 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 onwards, the BS is above its reference value (Brier Score of a climatological forecast ~0.2) in most regions and for all variables. There is no clear difference between high/low extremes in the different regions, despite the distinct biases 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 on 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 with higher skill, when compared with other regions.

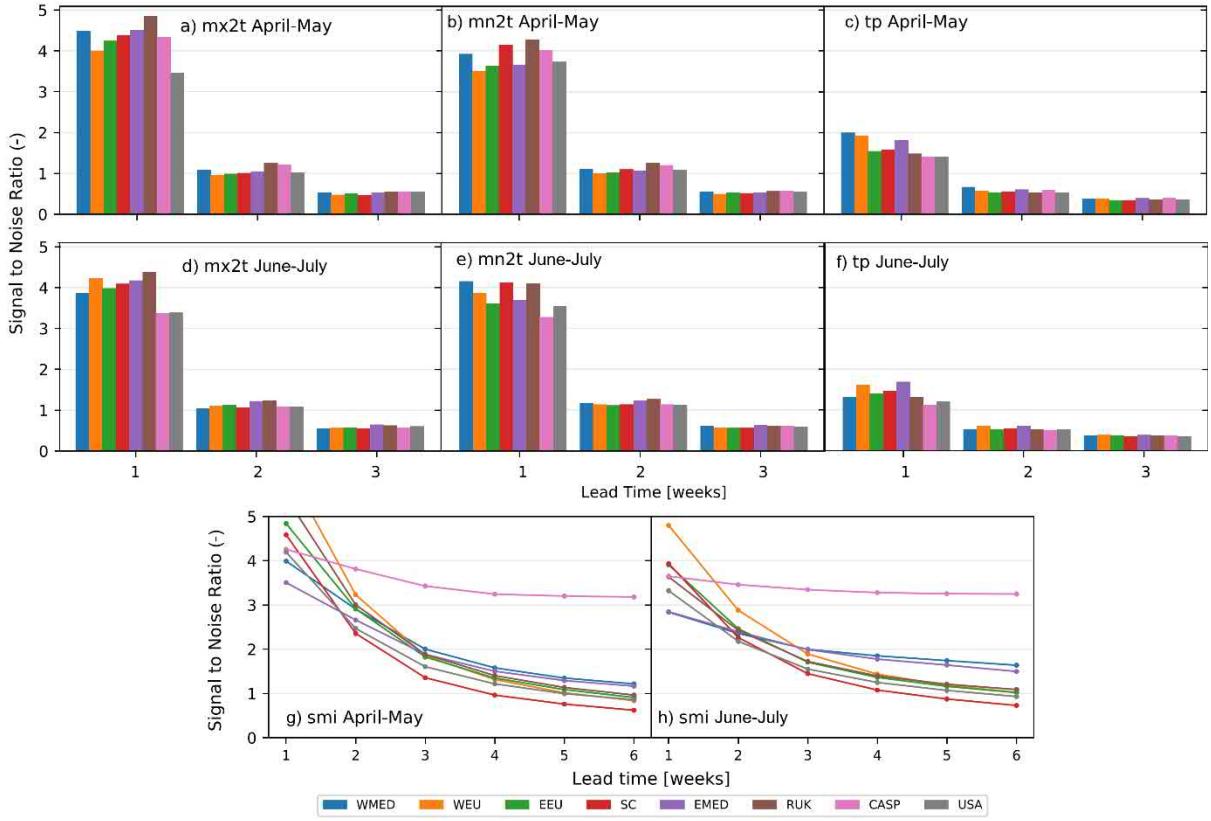

Figure 3- Signal-to-noise ratio of the ensemble forecasts as function of lead time for each region (bars and color lines) for the forecasts initialized in April-May (a,b,c,g) and June-July (d,e,f,h) of mx2t (a,d), mn2t (b,e) tp (c,f) and smi (g,h).

4 Discussion

The previous results of forecast skill are stratified by the forecasts biases, to assess possible relations between forecast biases and skill. The representation in scatterplots of the various metrics is a concise way of visualizing the data. Figure 5 displays the bias vs ACC scatterplots. No clear relation is present for any of the variables in all regions. The same can be said about the bias vs 25th percentile BS, represented in Figure A 12. The scatter plots of the BS of mx2t above the 75th percentile as function of mx2t bias on week 2 for the June-July forecast (Figure A 13) suggest that regions with higher cold biases have a lower BS. The mx2t bias and SNR also appear to have some relation on week 1, with smaller SNRs matching smaller cold biases (Figure A 14). These two relations are counter-intuitive, suggesting

that regions with larger cold mx2t biases have higher predictability and higher skill in predicting mx2t above the 75th percentile. Further investigation is required to understand if these relationships are robust (e.g. considering confidence intervals). Overall, the current analysis does not show any consistent relation between the bias and the skill (ACC, BS) or the predictability (SNR) of the forecasts in the different regions.

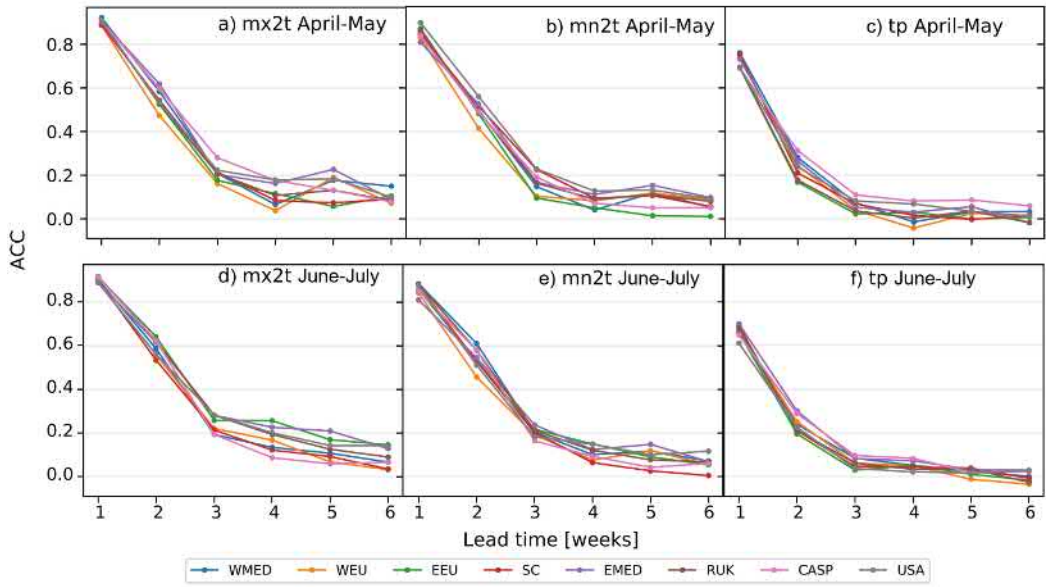

Figure 4- Anomaly correlation coefficient of the ensemble forecasts, using ERA5 as reference, as function of lead time for each region (color lines) for the forecasts initialized in April-May (a,b,c) and June-July (d,e,f) of mx2t (a,d), mn2t (b,e) and tp (c,f).

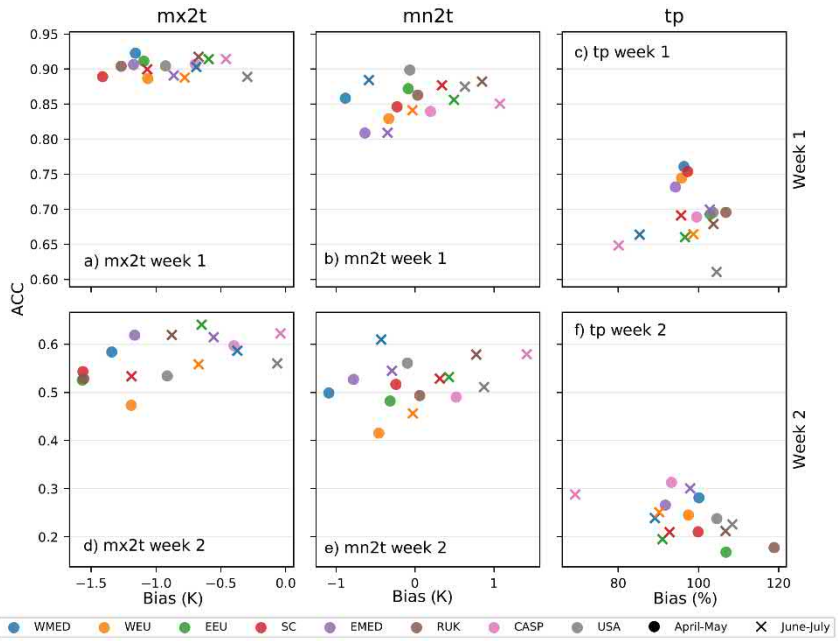

Figure 5 – Forecast bias (x-axis) versus ACC (y-axis) of the forecasts initialized in April-May (circles) and June-July (crosses), for the week 1 (a-c) and week 2 (d-f) forecasts lead times of mx2t (a,d), mn2d (b,e) and tp (c,f).

This study identified the CASP region with larger temperature biases that are associated with a dry bias (see Figure 6). The CASP region encompasses the southwestern part of Central Asia (composed by the countries of Turkmenistan, Uzbekistan and Kazakhstan). It is a very 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 (Lioubimtseva & Cole, 2006; Schiemann et al., 2008). When looking at the results of this study, Central Asia is one of the regions that stand out, due to its very distinct results across the various metrics. There is coherence between the bias, SD and SNR in this region for precipitation (negative bias, small variability 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 Figure 6 with a drift 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 (TWSV) 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 after the fall of the USSR and it has yet to recover, yet ERA5 is one of the reanalysis 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.

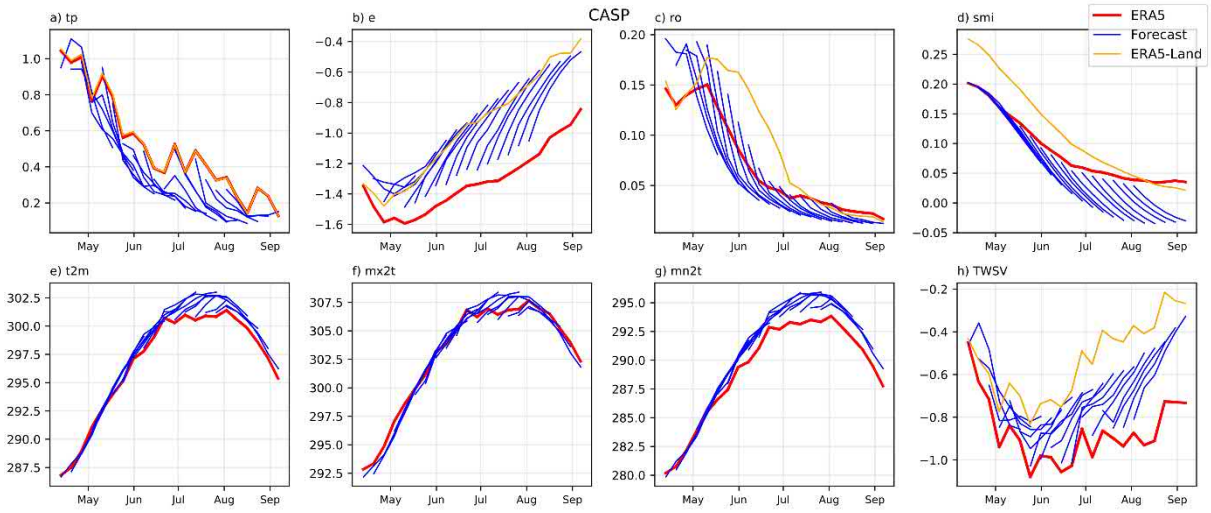

Figure 6- Mean weekly climate over CASP region of the forecasts for each starting date (blue) ERA5 (red) and ERA5-Land (orange). The variables displayed are total precipitation (a), evaporation (b), runoff (c) soil moisture index (d), 2-meter daily mean temperature (e), 2-meter daily maximum temperature (f), 2-meter daily minimum temperature (g) and terrestrial water storage variation (h).

The USA 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 warm and dry area in weather and climate models (Ardilouze et al., 2017, 2019; Lin et al., 2017). In this work, it is shown that the ECMWF reforecasts have similar warm biases to previous studies, mainly on the daily minimum temperature, while a dry bias is not present. The forecasts drift from ERA5 SMI to wetter conditions in May-June and drier conditions in July-August approaching ERA5-Land SMI (see Figure 7). 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 US, ERA5 can reproduce the interannual variability of precipitation reasonably well (H. E. Beck et al., 2019; Tarek et al., 2020) and similarly for surface soil moisture (H. Beck et al., 2020). Thus, it is likely that the warm bias reported in this study has a negligible relation with precipitation or surface soil moisture errors. The USA region, just like CASP, has lower SNR values for the maximum temperature and precipitation, displaying less coherence between the ensemble. 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.

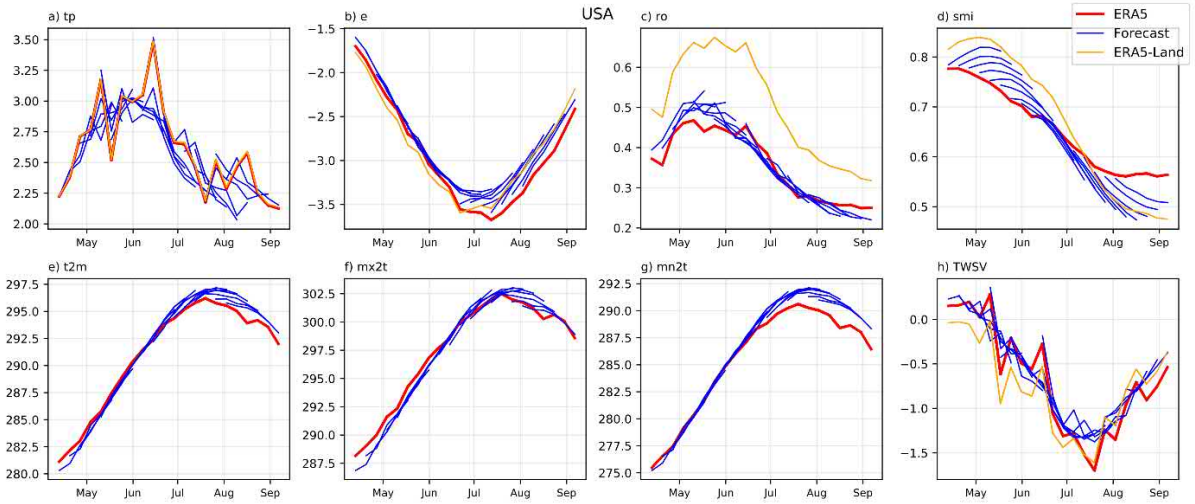

Figure 7- As Figure 6 but for the USA region.

The European regions have, in general, milder results 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. 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 Figure A 4, Figure A 5, Figure A 6). 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 (SC) 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 under-estimation of the diurnal cycle amplitude.

Contrasting with the continental European regions, the Mediterranean and Western Europe regions (WEU, WMED, EMED) do not present such large SMI drifts from ERA5 (Figure A 7 Figure A 8 Figure A 9). 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 Iberian Peninsula during summer when compared with satellite LST estimates. These biases were linked to vegetation cover in ERA5 (Nogueira et al., 2020), and are likely to also affect ERA5 2-metres 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.

5 Conclusions

This report aims to (i) document the ECMWF ensemble forecasts lead-time development of systematic errors in surface-related variables during spring and summer, and (ii) investigate the relation between the systematic errors and predictive skill. The evaluation has been done for regions on the northern hemisphere with different characteristics. The biases evolution as function of lead time for the daily temperature extremes, precipitation and soil moisture index revealed five key temporal/spatial bias patterns: (i) systematic cold bias of daily maximum temperature in the April-May forecasts at all lead times in all regions except USA and CASP; (ii) USA with a warm bias mostly in the daily minimum temperature; (iii) CASP region with a general warm and dry bias; (iv) Western and Mediterranean Europe with a cold bias in daily minimum temperature mainly in April-May forecasts and (v) 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.

The current analysis did not identify any clear dependence between the forecast biases and skill. The main conclusion is while there exist large differences in the systematic error characteristics, there is little relation to the skill for the sub-seasonal forecasts. However, these results do not reject the hypothesis that systematic biases affect forecast skill. Despite this, the general and systematic cold maximum daily temperature and warm minimum daily temperature biases require further attention from model development (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).

Acknowledgements

This work and F. Johannsen were funded by FCT under project CONTROL: PTDC/CTA-MET/28946/2017. The authors would like to thank David Richardson and Gianpaolo Balsamo for comments and suggestions that helped to improve the manuscript.

References

Albers, J. R., & Newman, M. (2019). A Priori Identification of Skillful Extratropical Subseasonal Forecasts. *Geophysical Research Letters*, *46*(21), 12527–12536. https://doi.org/10.1029/2019GL085270

Ardilouze, C., Batté, L., Bunzel, F., Decremer, D., Déqué, M., Doblas-Reyes, F. J., Douville, H., Fereday, D., Guemas, V., MacLachlan, C., Müller, W., & Prodhomme, C. (2017). Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. *Climate Dynamics*, *49*(11–12), 3959–3974. https://doi.org/10.1007/s00382-017-3555-7

Ardilouze, C., Batté, L., Decharme, B., & Déqué, M. (2019). On the link between summer dry bias over

the U.S. great plains and seasonal temperature prediction skill in a dynamical forecast system. *Weather and Forecasting*, *34*(4), 1161–1172. https://doi.org/10.1175/WAF-D-19-0023.1

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I. J. M., McVicar, T. R., & Adler, R. F. (2019). MSWep v2 Global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1

Beck, H., Pan, M., Miralles, D., Reichle, R., Dorigo, W., Hahn, S., Sheffield, J., Karthikeyan, L., Balsamo, G., Parinussa, R., van Dijk, A., Du, J., Kimball, J., Vergopolan, N., & Wood, E. (2020). Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors. *Hydrology and Earth System Sciences Discussions*, 1–35. https://doi.org/10.5194/hess-2020-184

Beljaars, A. (2020). Towards optimal parameters for the prediction of near surface temperature and dewpoint. *ECMWF Tech. Memo.*, *868*. https://doi.org/10.21957/yt64x7rth

de Andrade, F. M., Coelho, C. A. S., & Cavalcanti, I. F. A. (2019). Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dynamics*, *52*(9–10), 5451–5475. https://doi.org/10.1007/s00382-018-4457-z

Dirmeyer, P. A. (2005). The land surface contribution to the potential predictability of boreal summer season climate. *Journal of Hydrometeorology*, *6*(5), 618–632. https://doi.org/10.1175/JHM444.1

Dirmeyer, P. A., Halder, S., & Bombardi, R. (2018). On the Harvest of Predictability From Land States in a Global Forecast Model. *Journal of Geophysical Research: Atmospheres*, *123*(23), 13,111-13,127. https://doi.org/10.1029/2018JD029103

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*(15), 5620–5628. https://doi.org/10.1002/2014GL061146

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Jiang, J., Zhou, T., & Zhang, W. (2019). Evaluation of Satellite and Reanalysis Precipitable Water Vapor Data Sets Against Radiosonde Observations in Central Asia. *Earth and Space Science*, *6*(7), 1129–1148. https://doi.org/10.1029/2019EA000654

Johannsen, F., Ermida, S., Martins, J. P. A., Trigo, I. F., Nogueira, M., & Dutra, E. (2019). Cold bias of ERA5 summertime daily maximum land surface temperature over Iberian Peninsula. *Remote Sensing*, *11*(21), 2570. https://doi.org/10.3390/rs11212570

Kim, H., Vitart, F., & Waliser, D. E. (2018). Prediction of the Madden-Julian oscillation: A review. *Journal of Climate*, *31*(23), 9425–9443. https://doi.org/10.1175/JCLI-D-18-0210.1

Koster, R. D., Mahanama, S. P. P., Yamada, T. J., Balsamo, G., Berg, A. A., Boisserie, M., Dirmeyer, P. A., Doblas-Reyes, F. J., Drewitt, G., Gordon, C. T., Guo, Z., Jeong, J. H., Lee, W. S., Li, Z., Luo, L., Malyshev, S., Merryfield, W. J., Seneviratne, S. I., Stanelle, T., … Wood, E. F. (2011). The second phase of the global land-atmosphere coupling experiment: Soil moisture contributions to subseasonal forecast skill. *Journal of Hydrometeorology*, *12*(5), 805–822. https://doi.org/10.1175/2011JHM1365.1

Kumar, A. (2009). Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review*, *137*(8), 2622–2631. https://doi.org/10.1175/2009MWR2814.1

Lavaysse, C., Naumann, G., Alfieri, L., Salamon, P., & Vogt, J. (2019). Predictability of the European heat and cold waves. *Climate Dynamics*, *52*(3–4), 2481–2495. https://doi.org/10.1007/s00382-018-4273-5

Lin, Y., Dong, W., Zhang, M., Xie, Y., Xue, W., Huang, J., & Luo, Y. (2017). Causes of model dry and warm bias over central U.S. and impact on climate projections. *Nature Communications*, *8*(1), 1–8. https://doi.org/10.1038/s41467-017-01040-2

Lioubimtseva, E., & Cole, R. (2006). Uncertainties of Climate Change in Arid Environments of Central Asia. *Reviews in Fisheries Science*, *14*(1–2), 29–49. https://doi.org/10.1080/10641260500340603

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*(3), 289–307. https://doi.org/10.3402/tellusa.v21i3.10086

Magnusson, L., Ferranti, L., & Vamborg, F. (2018). Forecasting the 2018 European heatwave. *ECMWF Newsletter*, *157*(157). https://www.ecmwf.int/en/newsletter/157/news/forecasting-2018-european-heatwave

Muñoz Sabater, J. (2019). First ERA5-Land dataset to be released this spring. *ECMWF Newsletter*, *159*, 8–9. https://www.ecmwf.int/sites/default/files/elibrary/2019/19001-newsletter-no-159-spring-2019.pdf

Nogueira, M., Albergel, C., Boussetta, S., Johannsen, F., Trigo, I. F., Ermida, S. L., Martins, J. P. A., & Dutra, E. (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. *Geoscientific Model Development Discussions*, *2020*, 1–29. https://doi.org/10.5194/gmd-2020-49

Orsolini, Y. J., Senan, R., Balsamo, G., Doblas-Reyes, F. J., Vitart, F., Weisheimer, A., Carrasco, A., & Benestad, R. E. (2013). Impact of snow initialization on sub-seasonal forecasts. *Climate Dynamics*, *41*(7–8), 1969–1982. https://doi.org/10.1007/s00382-013-1782-0

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., Joseph Metzger, E., Barton, N. P., Achuthavarier, D., Marshak, J., Koster, R. D., … Kim, H. (2019). The subseasonal experiment (SUBX). *Bulletin of the American Meteorological Society*, *100*(10), 2043–2060. https://doi.org/10.1175/BAMS-D-18-0270.1

Prodhomme, C., Doblas-Reyes, F., Bellprat, O., & Dutra, E. (2016). Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate Dynamics*, *47*(3–4), 919–935. https://doi.org/10.1007/s00382-015-2879-4

Robertson, A. W., Kumar, A., Peña, M., & Vitart, F. (2015). Improving and promoting subseasonal to seasonal prediction. *Bulletin of the American Meteorological Society*, *96*(3), ES49–ES53. https://doi.org/10.1175/BAMS-D-14-00139.1

Schiemann, R., Lüthi, D., Vidale, P. L., & Schär, C. (2008). The precipitation climate of Central Asia—intercomparison of observational and numerical data sources in a remote semiarid region. *International Journal of Climatology*, *28*(3), 295–314. https://doi.org/10.1002/joc.1532

Tarek, M., Brissette, F. P., & Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrology and Earth System Sciences*, *24*(5), 2527–2544. https://doi.org/10.5194/hess-24-2527-2020

Vigaud, N., Tippett, M. K., Yuan, J., Robertson, A. W., & Acharya, N. (2019). Probabilistic skill of subseasonal surface temperature forecasts over North America. *Weather and Forecasting*, *34*(6), 1789–1806. https://doi.org/10.1175/WAF-D-19-0117.1

Vitart, F., Alonso-Balmaseda, M., Ferranti, L., Benedetti, A., Balan-Sarojini, B., Tietsche, S., Yao, J.,

Janousek, M., Balsamo, G., Leutbecher, M., Bechtold, P., Polichtchouk, I., Richardson, D., Stockdale, T., & Roberts, C. D. (2019). Extended-range prediction. *ECMWF Tech. Memo.*, *854*. https://doi.org/10.21957/pdivp3t9m

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., … Zhang, L. (2017). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1

Vitart, F., Balsamo, G., Bidlot, J.-R., Lang, S., Tsonevsky, I., Richardson, D., & Alonso-Balmaseda, M. (2019). Use of ERA5 reanalysis to initialise re-forecasts proves beneficial. *ECMWF Newsletter*, *161*, 26–31. https://doi.org/10.21957/g71fv083lm

Vitart, F., Robertson, A. W., & Anderson, D. L. T. (2012). Subseasonal to Seasonal Prediction Project: Bridging the Gap between Weather and Climate. *Bulletin of the World Meteorological Organization*, *61*(2), 23.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A. P., Eggen, B., Graham, R., Kjellström, E., Becker, E., … Zebiak, S. E. (2017). Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, *24*(3), 315–325. https://doi.org/10.1002/met.1654

Wulff, C. O., & Domeisen, D. I. V. V. (2019). Higher Subseasonal Predictability of Extreme Hot European Summer Temperatures as Compared to Average Summers. *Geophysical Research Letters*, *46*(20), 11520–11529. https://doi.org/10.1029/2019GL084314

Zhou, Y., Yang, B., Chen, H., Zhang, Y., Huang, A., & La, M. (2019). Effects of the Madden–Julian Oscillation on 2-m air temperature prediction over China during boreal winter in the S2S database. *Climate Dynamics*, *52*(11), 6671–6689. https://doi.org/10.1007/s00382-018-4538-z

Zsoter, E., Cloke, H., Stephens, E., DE ROSNAY, P., Muñoz-Sabater, J., Prudhomme, C., & Pappenberger, F. (2019). How well do operational numerical weather prediction configurations represent hydrology? *Journal of Hydrometeorology*, *20*(8), 1533–1552. https://doi.org/10.1175/JHM-D-18-0086.1

Appendix I Forecast skill metrics

This section presents the calculation details of the different metrics used in this study. The equations are given in Table A 1 with the following description:

- Bias represents the difference between the average forecast ensemble mean and the average of the reference product (Eq. 1);
- Standard Deviation Ratio (SDR) is the ratio between the temporal standard deviation of the ensemble mean and the temporal standard deviation of the reference product, which can be also computed as the square root of the ratio between the ensemble mean and reference product temporal variance (Eq. 2);
- Anomaly Correlation Coefficient (ACC) evaluates the skill of the ensemble mean at reproducing the reference product spatial pattern anomaly (Eq. 3);
- Brier score (BS) evaluates the accuracy of the ensemble in forecasting a specific event, in this study below the 25th or above the 75th percentile, by computing the mean square difference

between the probability of forecasting the event (between 0 and 1) and the actual outcome of the event (0 or 1) (Eq. 4);

- Signal-to-Noise Ratio (SNR) measures the coherence between the different ensemble members (Eq. 5).

In Table A 1 y and o represent the forecast and the reference product (ERA5 in this study), respectively. N, J, M represent the number of year (20), the number of ensemble members (11) and the number of grid points in a particular region, respectively. $\bar{y}$ 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 data, lead time), and are averaged over the total number of start dates (split between April-May and June-July). For the ACC, the Fisher Z transform is applied to the ACC before the temporal averages and then inverted to the correlation, that will represent the average ACC over a given region. Lastly, a regional mean is applied for all metrics (except ACC) for each forecast lead time.

Table A 1 – Metrics equations

$$Bias = \frac{1}{N}\sum_{k=1}^{N} \overline{y_k} - o_k \qquad \text{Eq. 1}$$

$$SDR = \sqrt{\frac{\sum_{k=1}^{N}(\overline{y_k} - \bar{y})^2}{\sum_{k=1}^{N}(o_k - \bar{o})^2}} \qquad \text{Eq. 2}$$

$$ACC = \frac{\sum_{m=1}^{M}(y'_m - \bar{y'})(o'_m - \bar{o'})}{[\sum_{m=1}^{M}(y'_m - \bar{y'})^2 \sum_{m=1}^{M}(o'_m - \bar{o'})^2]^{1/2}} \qquad \text{Eq. 3}$$

$$BS = \frac{1}{N}\sum_{k=1}^{N}(Y_k - O_k)^2 \qquad \text{Eq. 4}$$

$$SNR = \sqrt{\frac{\frac{1}{N-1}\sum_{k=1}^{N}(\overline{y_k} - \bar{y})^2}{\frac{1}{N \times J - N}\sum_{k=1}^{N}\sum_{j=1}^{J}\left(y_{k,j} - \overline{y_k}\right)^2}} \qquad \text{Eq. 5}$$

$$\bar{y} = \frac{1}{N \times J}\sum_{k=1}^{N}\sum_{j=1}^{J}\left(y_{k,j}\right) \;\; ; \overline{y_k} = \frac{1}{J}\sum_{j=1}^{J}\left(y_{k,j}\right) \qquad \text{Eq. 6}$$

Appendix II Auxiliary figures

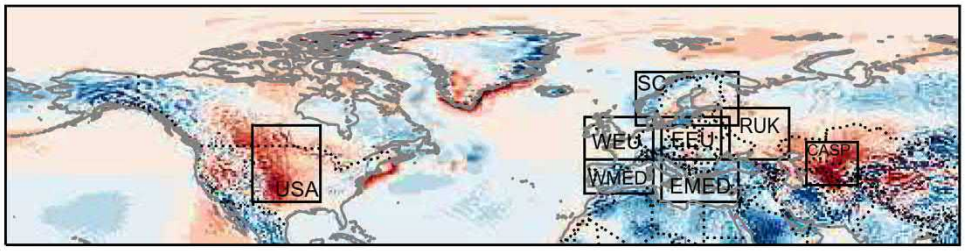

Figure A 1 – Spatial extent of the regions used in this study. The underlying map shows the minimum temperature bias of the forecasts initialized in June-July for week 5 lead time, as in Figure 1.

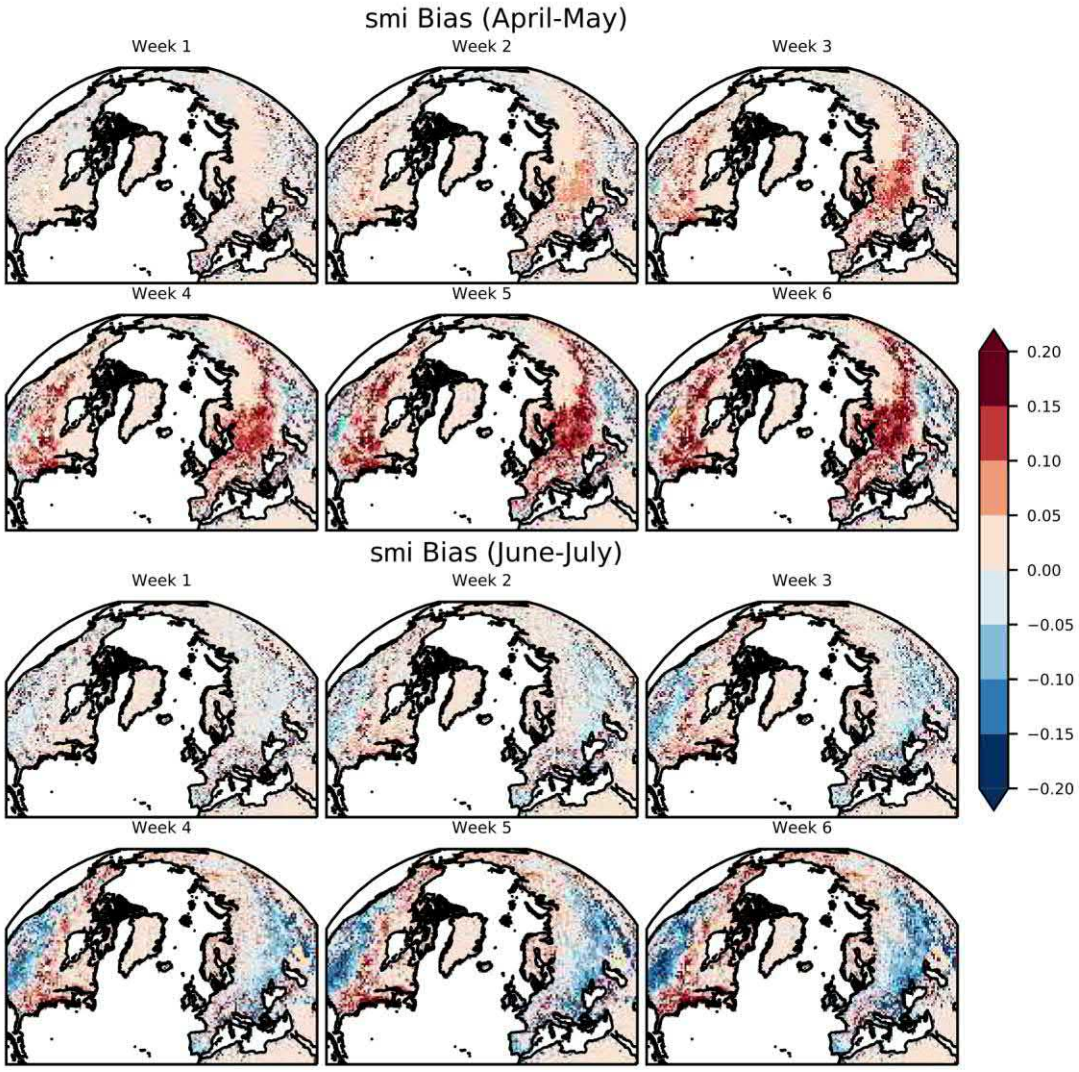

Figure A 2 – Soil moisture index biases in week 1 to 6 of the forecasts initialized in April-May (top) and June-July (bottom).

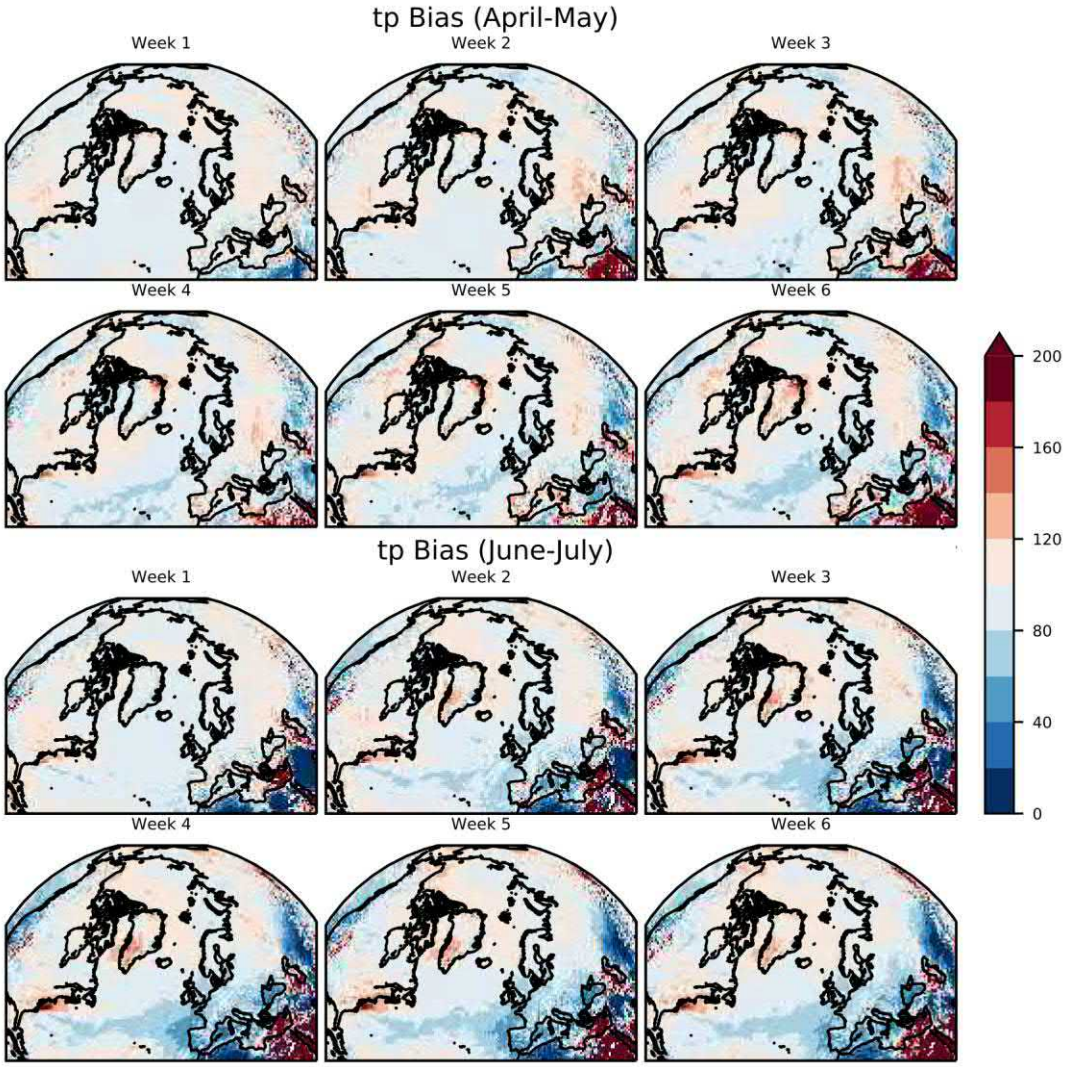

Figure A 3 – Total precipitation percent biases in week 1 to 6 of the forecasts initialized in April-May (top) and June-July (bottom).

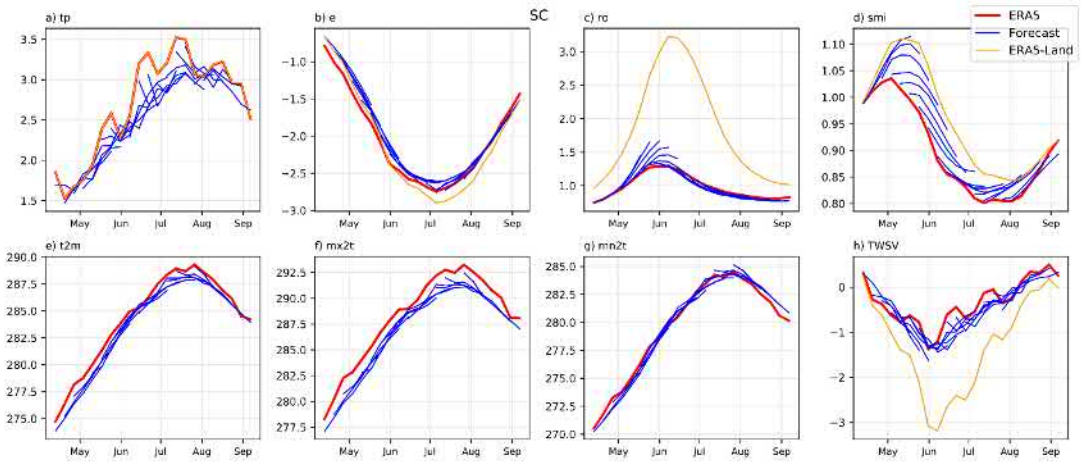

Figure A 4- As Figure 6 but for the SC region.

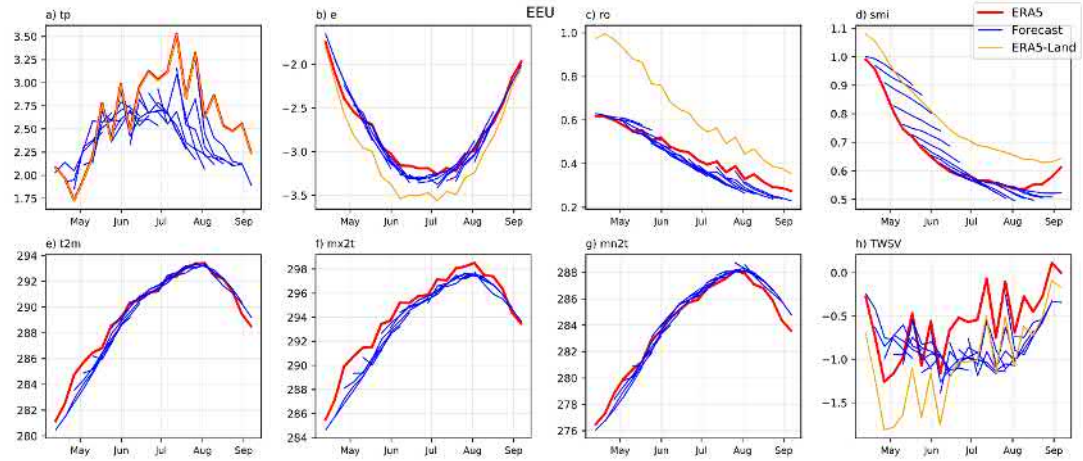

Figure A 5- As Figure 6 but for the EEU region.

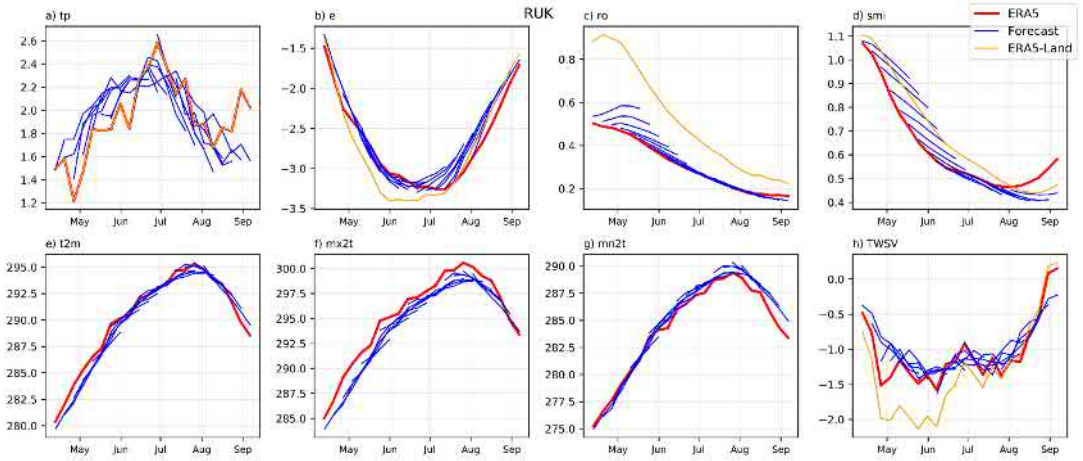

Figure A 6- As Figure 6 but for the RUK region.

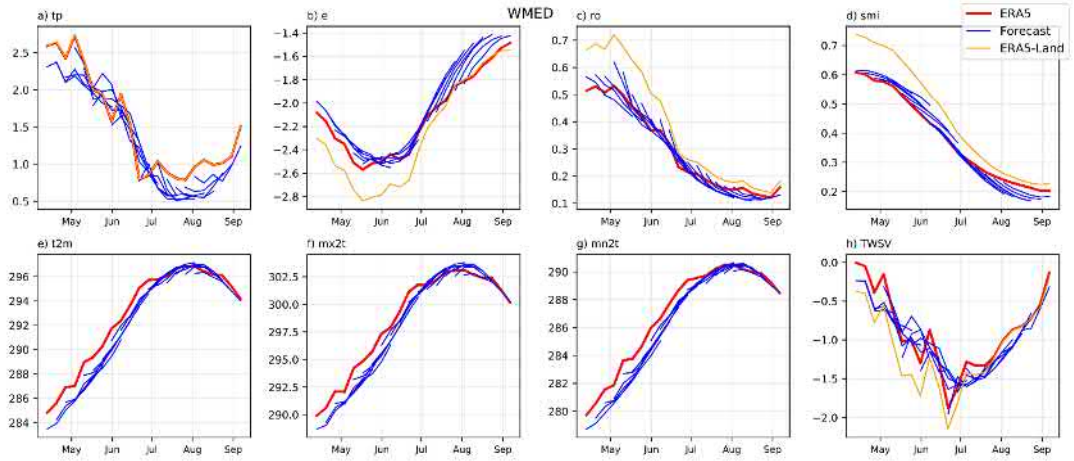

Figure A 7- As Figure 6 but for the WMED region.

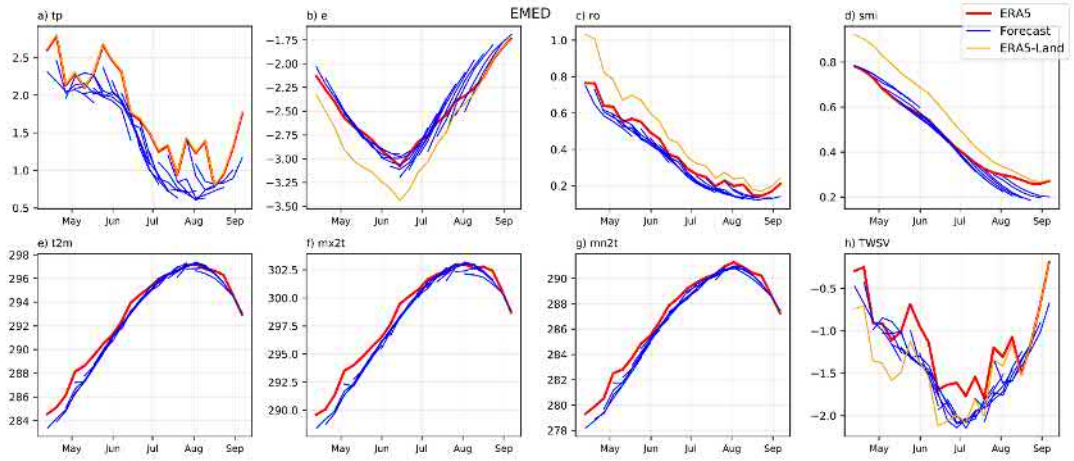

Figure A 8- As Figure 6 but for the EMED region.

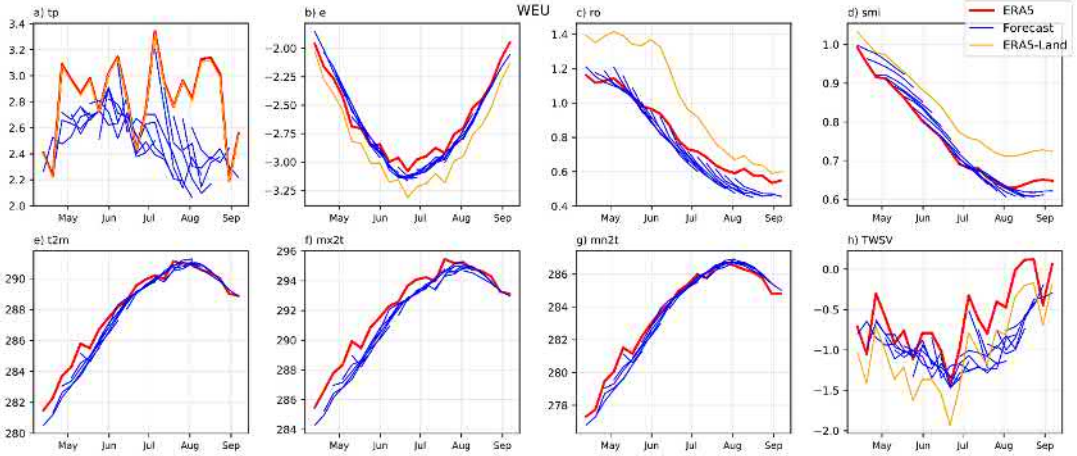

Figure A 9- As Figure 6 but for the WEU region.

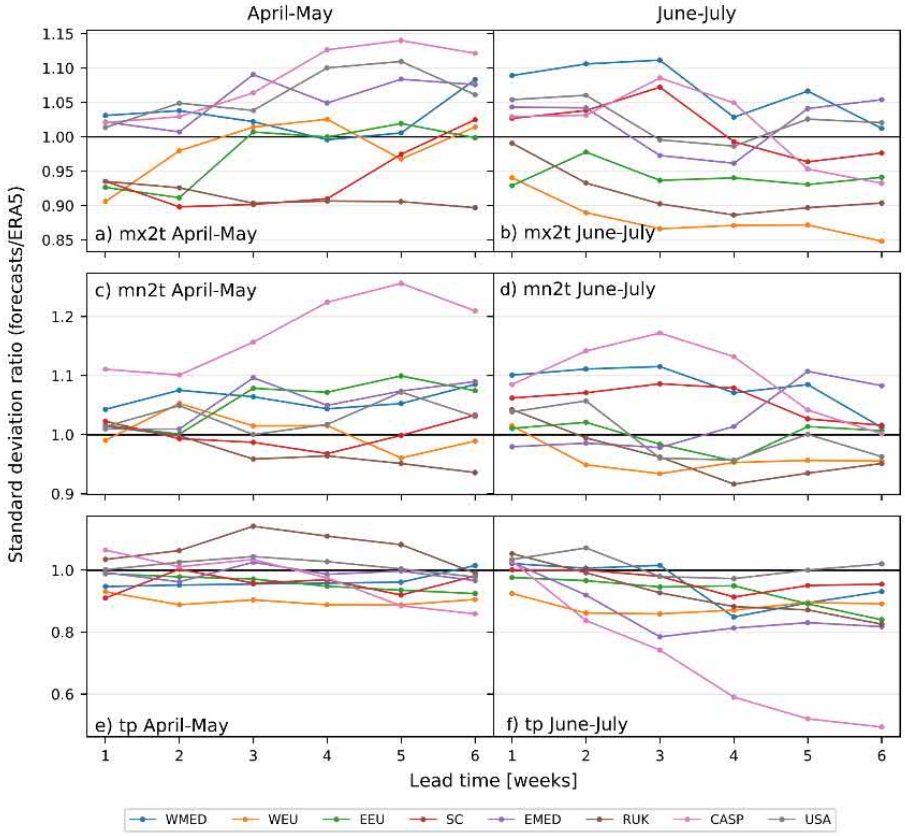

Figure A 10- Standard deviation ratio between forecast and ERA5 as function of lead time (weeks) in each region (color lines) for the April-May and June-July starting dates for mx2t (a,b), mn2t (c,d) and tp (e,f).

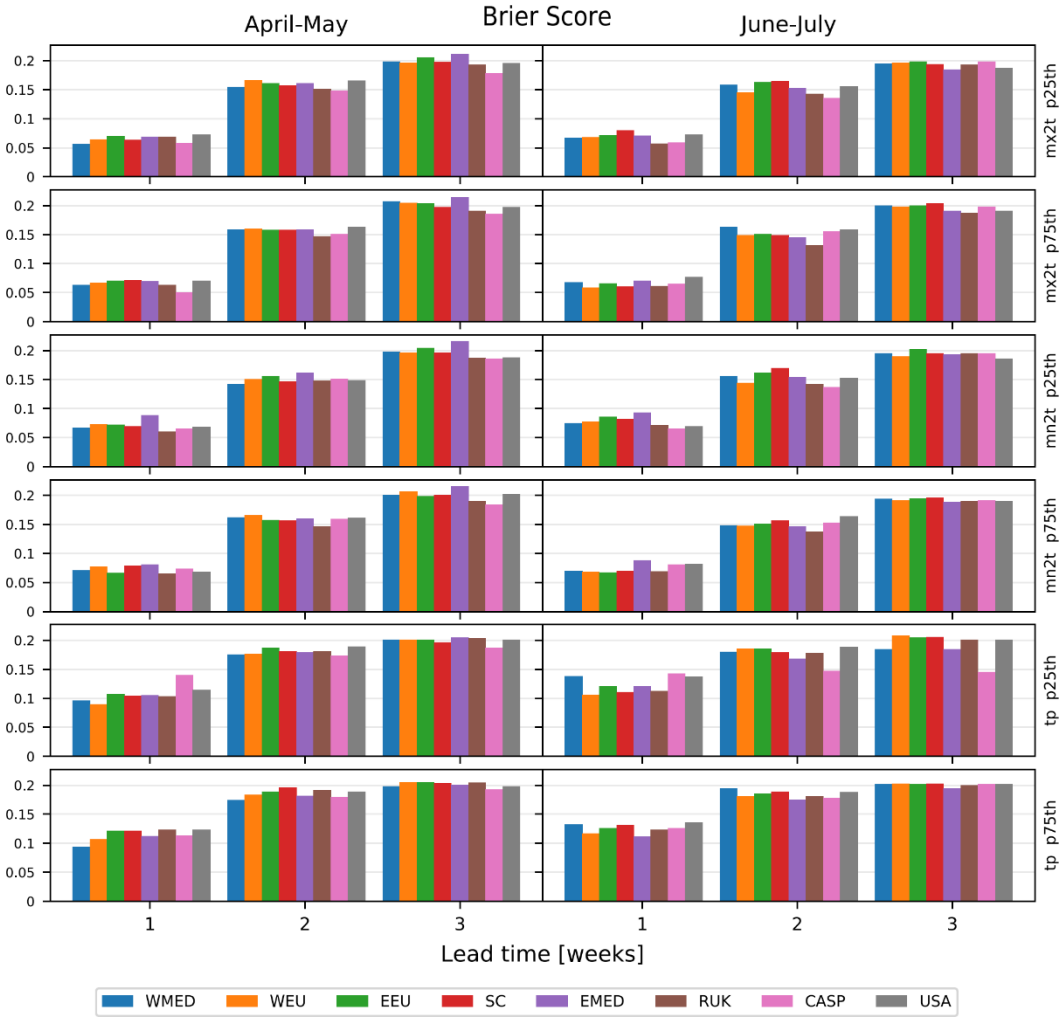

Figure A 11- Brier score as function of lead time (week) for the forecasts initialized in April-May (left panel) and June-July (right panels) from top to bottom: mx2t below percentile 25, mx2t above percentile 75, mn2t below percentile 25, mn2t above percentile 75, tp below percentile 25, and tp above percentile 75.

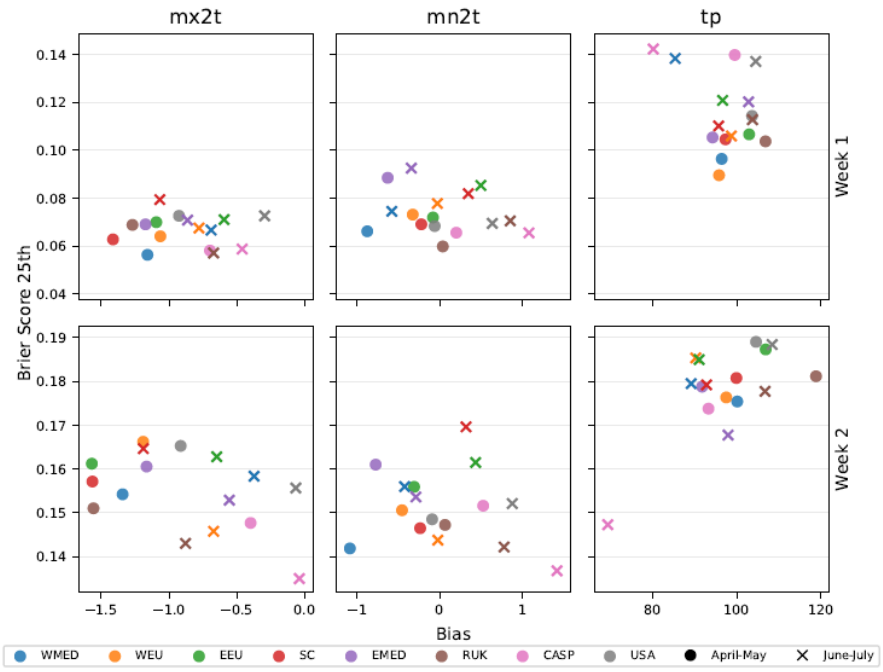

Figure A 12 – As Figure 5 for but the brier score of the forecasts below the 25th percentile.

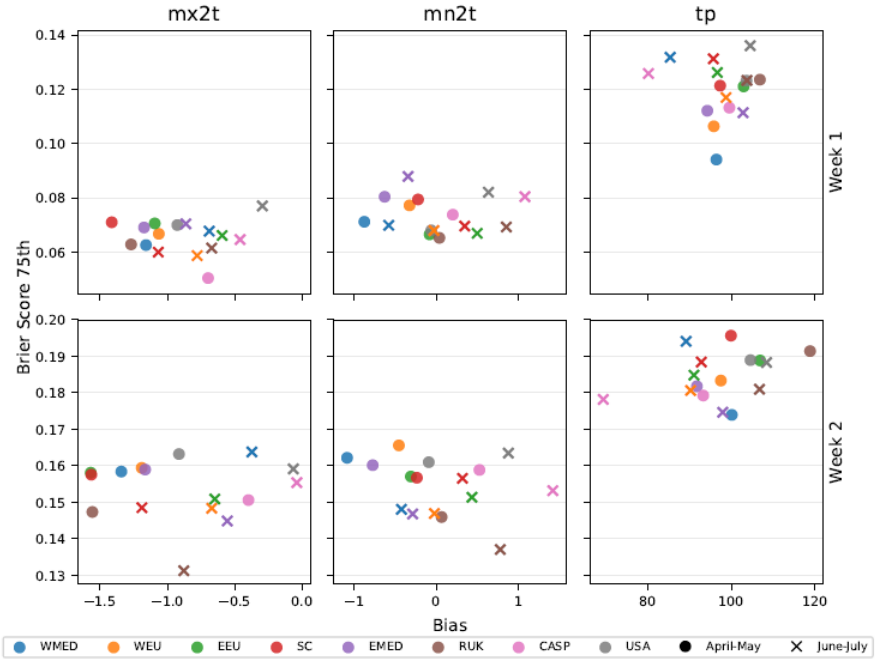

Figure A 13- As Figure 5 for but the brier score of the forecasts above the 75th percentile.

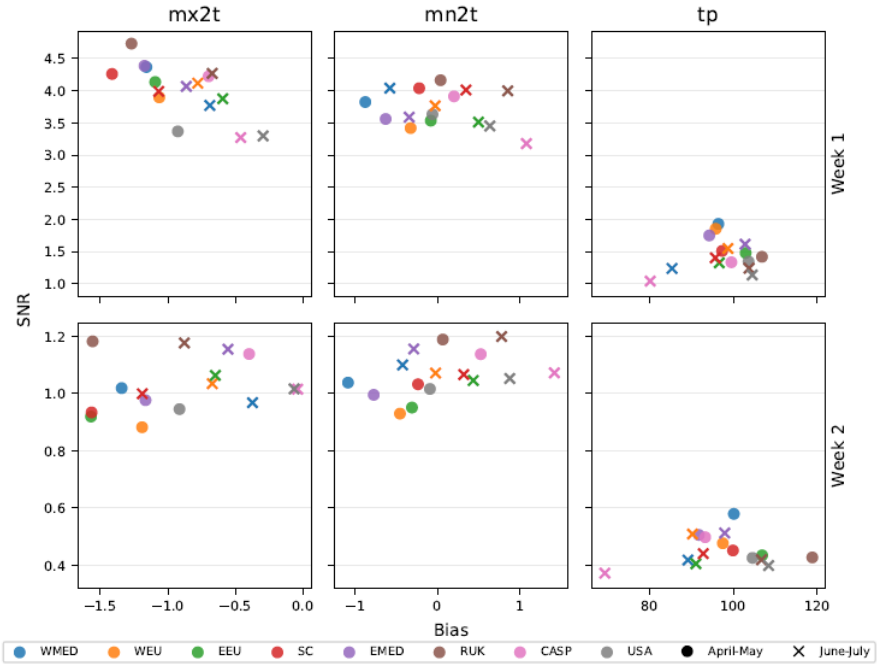

Figure A 14 - As Figure 5 for but the signal-to-noise ratio.