Exploiting the signal to noise ratio in multi-system predictions of boreal summer precipitation and temperature

Juan C. Acosta Navarro[1] and Andrea Toreti[1]

[1]European Commission, Joint Research Centre, Ispra, Italy.

Correspondence to: Juan C Acosta Navarro (juan.acosta-navarro@ec.europa.eu)

Abstract

Droughts and heatwaves are among the most impactful climate extremes. Their co-occurrence can have adverse consequences on natural and human systems. Early information on their possible occurrence on seasonal timescales is beneficial for many stakeholders. Seasonal climate forecasts have become openly available to the community but a wider use is currently hindered by limited skill in certain regions and seasons. Here we show that a simple forecast metric from a multi-system ensemble, the signal to noise ratio, can help overcome some limitations. Forecasts of mean daily near surface air temperature and precipitation in boreal summers with high signal to noise ratio tend to coincide with observed larger deviations from the mean than summers with small signal to noise ratio. The signal to noise ratio of the ensemble predictions may serve as a complementary measure of forecast reliability that could benefit users of climate predictions.

1. Introduction

Droughts are typically slow onset climate extreme events (Mishra and Singh, 2010), yet they can be disruptive and affect millions of people every year (Below et al., 2007; Enekel et al., 2020). Heatwaves can intensify and trigger a faster drought evolution (Bevacqua et al., 2022). Compound drought and heatwaves can strongly impact socio-economic and ecological systems, and may even compromise our ability to reach the UN sustainable development goal on climate action while strongly reducing the Earth system's current natural capacity to absorb and store carbon (Yin et al., 2023). The use of seasonal climate forecasts can provide actionable information to reduce the risks and the impacts of these events on key sectors like agriculture, energy, transport, water supply (Buontempo et al 2018; Ceglar and Toreti 2021).

In the last couple of decades, climate predictions have shown important progress in anticipating the evolution of various components of the climate system across the subseasonal to decadal time range (Merryfield et al., 2020; Meehl et al., 2021). A combination of multiple forecast systems has shown overall benefits as compared with single systems, and can improve forecast quality up to a certain extent (Hagedorn et al., 2005; Mishra et al., 2019). In spite of the recent progress, climate predictions still exhibit low to moderate skill in many regions and seasons (e.g. European summer; Mishra et al. 2019), something that limits their use and represents a barrier for stakeholders. Furthermore, multiple studies have shown that large ensembles are required to achieve skillful predictions, something that seems to be related to the forecast systems being more skillful at predicting real climate

39 than at predicting their own realizations (i.e. ensemble members). This odd phenomenon has been called the signal
40 to noise paradox (Eade et al., 2014; Scaife and Smith, 2018; Smith et al., 2020). It is particularly evident in the
41 Euro Atlantic region during winter both on seasonal and decadal timescales. However boreal summer predictions
42 have been generally overlooked. A recent study based on a single forecasting system has shown that sampling
43 years with high SNR results in more skillful predictions of monthly temperatures in Japan throughout the year
44 (Doi et al., 2022).

46 In this study we exploit multi-system ensembles to test whether specific boreal summers with higher than normal
47 predictability can be detected through the local relation between skill and SNR. We explore this for near surface
48 air temperature and precipitation, both locally and on large aggregated mid-latitude regions of the Northern
49 Hemisphere.

51 **2. Methods**

53 This analysis is based on seasonal re-forecasts (also known as hindcasts) of mean boreal summer precipitation
54 and 2-meter mean daily temperature (T2m) for the period 1993-2016 from ECMWF SEAS5 (S5, Johnson et al.,
55 2019), UKMO GloSea6 (S600, MacLachlan et al., 2015), MeteoFrance (S8, Batté et al., 2021), CMCC (S35,
56 Gualdi et al., 2020) and DWD (S21, Baehr et al., 2015), available from the Copernicus C3S Climate Data Store.
57 The observationally based datasets to evaluate the re-forecasts are ERA5 (Hersbach al., 2020) for T2m and GPCC
58 (Schnider et al., 2011) for precipitation. The use of summer mean T2m is not intended to characterize single
59 heatwaves, but to estimate average daily deviations from the mean on a seasonal scale. In a climatological sense,
60 more intense, more frequent or longer heatwaves than usual generally define hot summers and hence average T2m
61 may be seen as a seasonal integrator of heatwave activity. Forecast skill is evaluated with the anomaly correlation
62 coefficient (ACC) between the ensemble mean and the observational reference. To complement the skill estimates
63 of ACC, two additional deterministic skill metrics are computed: the mean squared skill score (MSSS, Murphy,
64 1988) and the Gilbert skill score (GSS, WMO, 2014). The mean squared skill score compares the mean square
65 error of the forecasts with the mean square error of the climatological value. It ranges from minus infinity to 1
66 and values above 0 indicate skill in the predictions. The GSS measures the fraction of correctly predicted events
67 over the total number of predicted events plus misses, and takes into consideration the randomly predicted events.
68 The thresholds to define event/non event are the top and bottom 25% summers for T2m (hot) and precipitation
69 (dry), respectively. Standardization of the anomalies of each ensemble member and the observational reference
70 data is performed prior to the analysis. This step guarantees that each member from each system has a comparable
71 year-to-year variability to the observed one. Additionally, the standardized T2m anomalies are linearly detrended
72 at the grid level and for each member of the re-forecasts and in ERA5 to isolate as much as possible the impact of
73 the long term warming.

75 Following Doi et al. (2022), the SNR is calculated as: $SNR = \frac{\mu_e}{\sigma_e}$, where μ_e is the multi-system ensemble mean
76 and σ_e is the multi-system standard deviation after standardization, computed across ensemble members for every
77 summer (June - August) and for each gridbox. 25 members per system are used to have an equal contribution from
78 each system.

79

3. Signal to noise ratio and forecast skill

80

81

Figure 1 displays spatial maps of mean (boreal) summer T2m ACC, time averaged SNR, and a scatter plot which shows the local relation between ACC and SNR. On average, skill values over land increase with higher SNR values. Negative values of ACC are nearly non-existent when the threshold of SNR exceeds the value of about 0.5 in the same gridbox. Statistically significant skill in T2m is mostly confined to the tropics and sub-tropics. However, significant skill is also found in western North America, the eastern Mediterranean, central Asia and southern South America. Notable exceptions in the tropics are Congo and parts of the Amazon rainforests. The patterns of SNR largely mirror those of ACC. Generally, there is a good agreement between areas of high skill (ACC) and areas with high SNR, something that is further confirmed by the local relation between ACC and SNR (Fig. 1c).

91

92

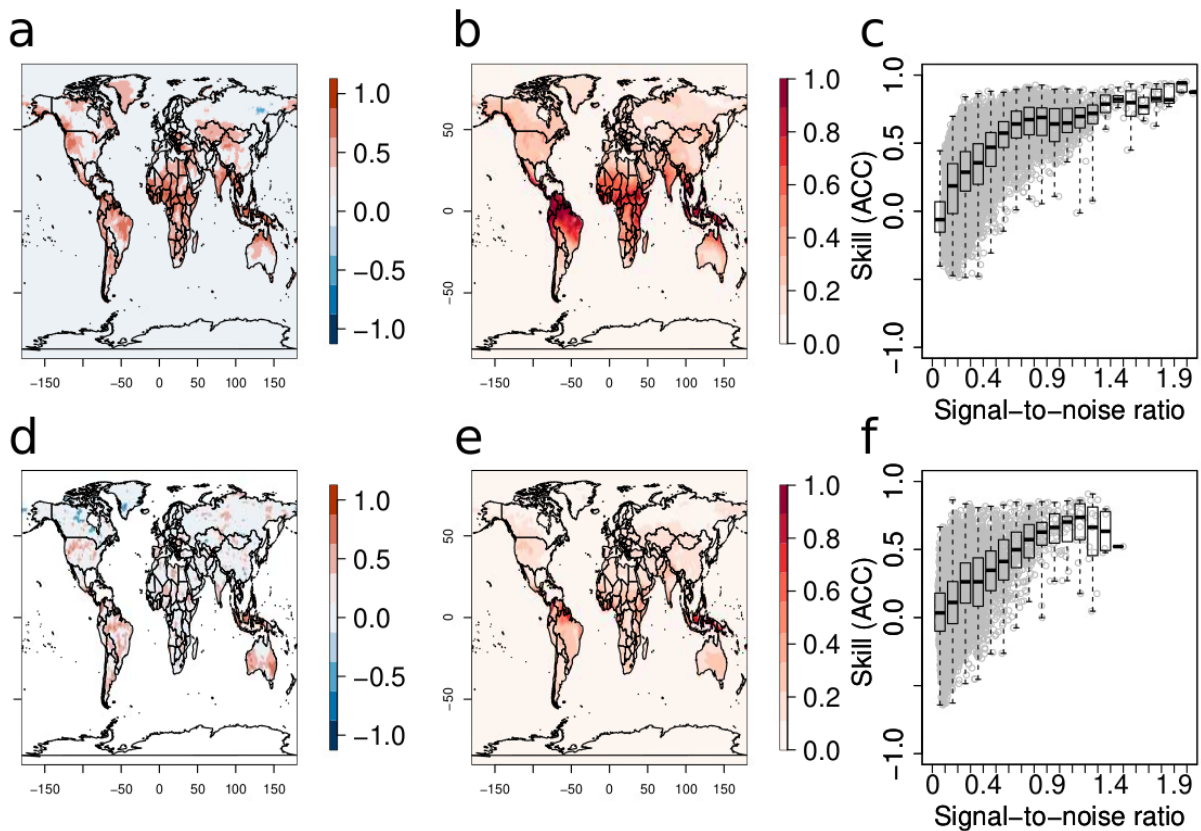



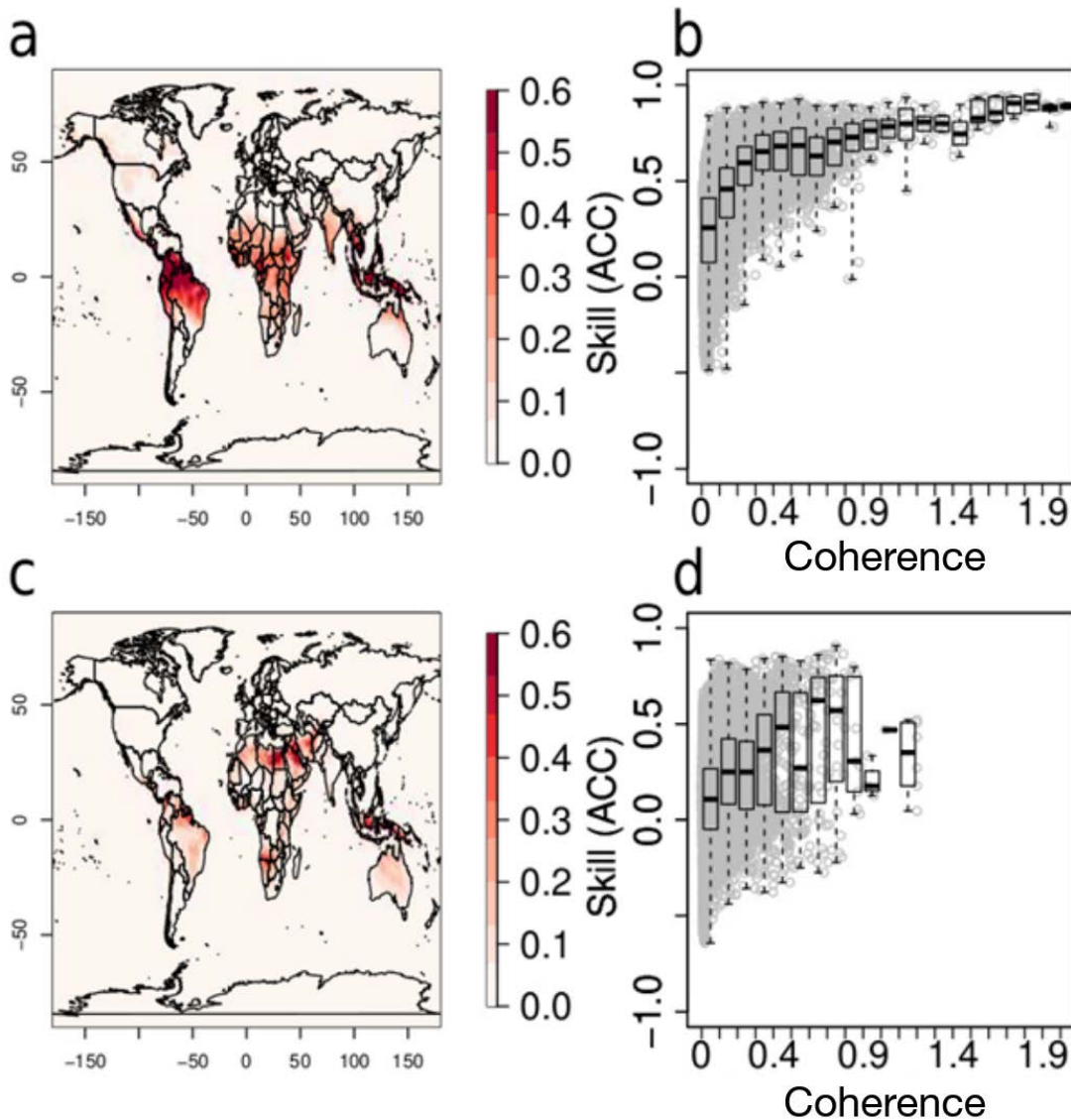
Figure 1: June-August Skill (ACC), time averaged SNR and scatterplots of local relation between ACC and SNR for T2m (a-c) and precipitation (d-f). Each gray dot in c,f represents the values of ACC and SNR at each gridbox. Only statistically significant values with a 90% confidence based on a t-test are displayed in (a,d). The re-forecasts are initialized every May.

97

Precipitation follows a similar behavior in terms of ACC and SNR, although statistically significant skill is less widespread (Fig. 1d-f). Areas under the influence of El Niño Southern Oscillation (ENSO; Lenssen et al., 2020) appear as regions with significant ACC and high SNR. Skillful values are mostly located in the Americas, the

101   Maritime continent and Australia. Precipitation skill and SNR in Africa and Asia are much lower, making these

102   the regions with the largest qualitative differences between the two variables.

103

105 **Figure 2: June-August time averaged ensemble coherence and scatterplots of the local relation between ACC and**

106 **ensemble coherence for T2m (a-b) and precipitation (c-d). Each gray dot in b,d represents the values of ACC and**

107 **ensemble coherence at each gridbox. The re-forecasts are initialized every May.**

108

109   In Figure 2 we show the effect of the ensemble coherence on skill. Ensemble coherence is defined as the inverse

110   of the ensemble standard deviation (σ_e), minus one. The spatial distribution of time averaged ensemble coherence

111   displays many similarities to the SNR for both T2m and precipitation, although the signal is clearly dominated by

112   the tropics and subtropics with virtually no contribution from the extra-tropics, except for a minor one from T2m

113   in western North America and from precipitation in the Middle East (Fig. 2a,c). In terms of the local relation

114   between ensemble coherence and skill, T2m displays a clear increase in skill with higher values of coherence (Fig.

115   2b). Skill is virtually always positive when coherence values exceed 0.3, implying that ensemble spread may also

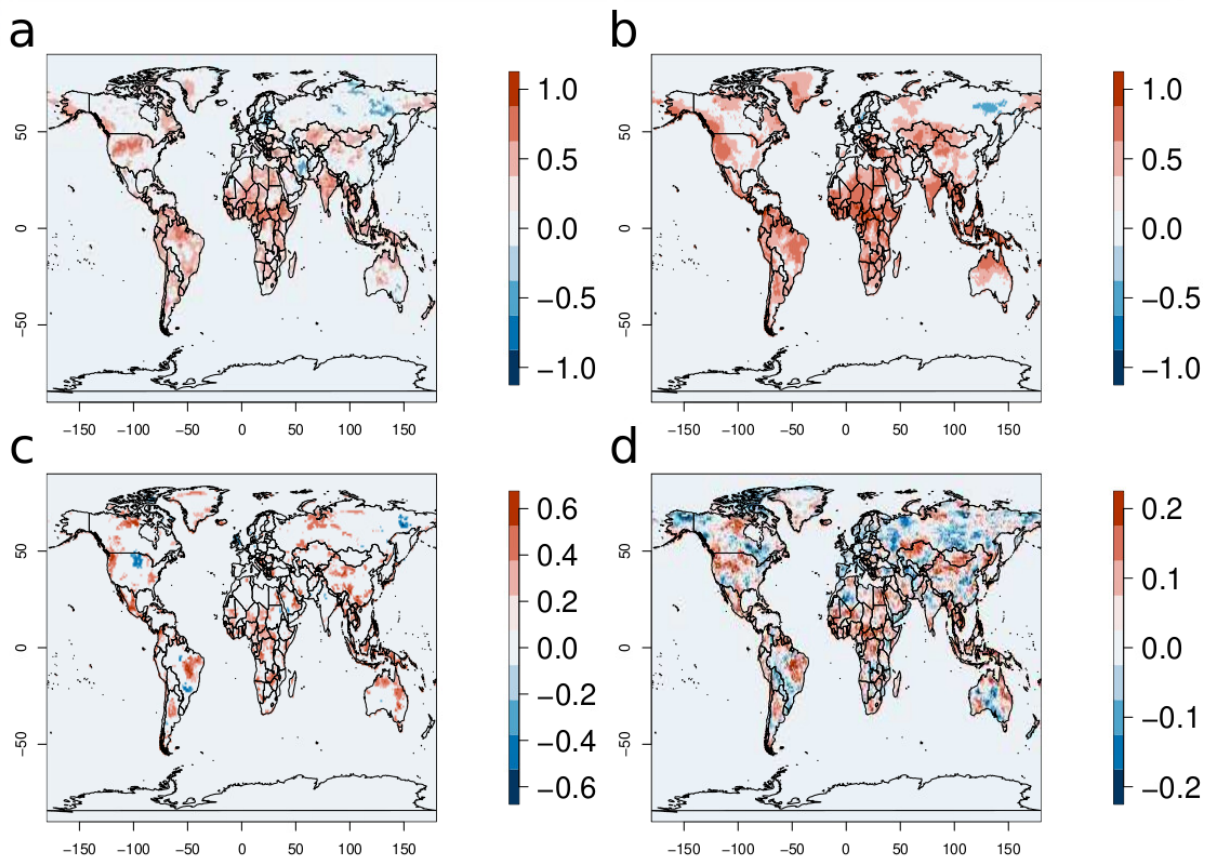
116   be a good indicator of skill for T2m, similar to SNR. For precipitation there is weaker relation between skill and

117   ensemble coherence than for T2m as there appear to be as many locations of high coherence with little skill as

118 locations with high skill and high coherence (Fig. 2d). This can be a result of a weaker relation between skill and
119 ensemble coherence than between skill and SNR, but may also be at least partially a result of the large uncertainty
120 in observed precipitation in many regions.

121

122 Based on the observed link between skill and SNR, we use the latter one as the single criterion to exclude from
123 the re-forecasts years with very low and very high values to understand their impact on skill. When 25% of the
124 years (6 in total) with the highest SNR (Fig. 3a) are excluded, the results overall show much lower values of ACC
125 than when only 25% of the years with the lowest SNR are excluded (Fig. 3b). Furthermore, differences between
126 the latter and the former result (in many cases) in higher statistically significant values than the ACC computed
127 when selecting only years without the highest SNR (Fig. 3a,c). This result highlights the importance that these
128 extreme SNR years can have on skill. In fact, only skill values that are computed by excluding the bottom 25% of
129 SNR years (Fig. 3b) are comparable to the ones estimated when all years are used for the computation (Fig. 1a).

130

131

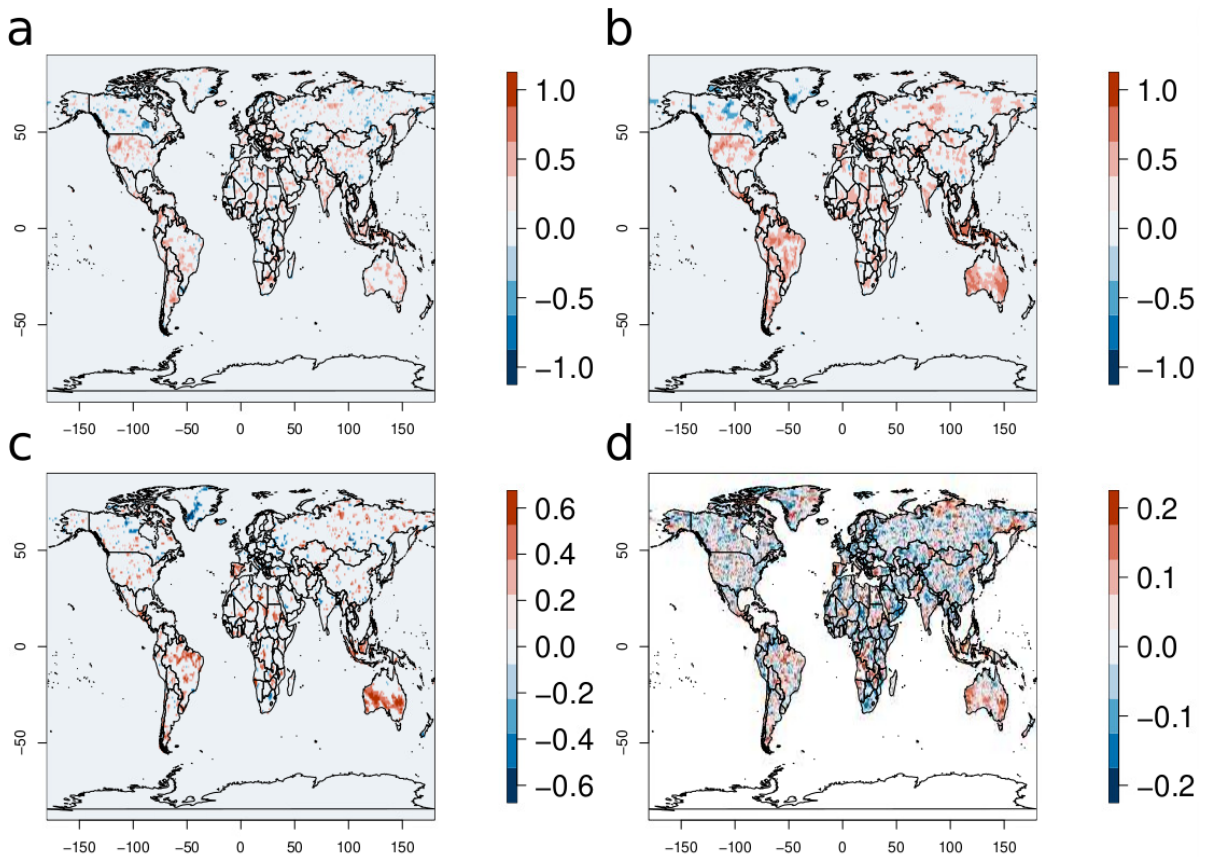
132 **Figure 3: Skill (ACC) of T2m predictions excluding 25% of the years with highest (a) and lowest (b) local SNR. (c)**
133 **Difference between (a) and (b). (d) Difference in the time-averaged absolute deviation from the mean in ERA5 T2m,**
134 **excluding years having 25% of the lowest and highest local SNR, respectively. Only statistically significant values with**
135 **a 90% confidence based on a t-test are displayed in (a-c). The re-forecasts are initialized every May.**

136

137 Interestingly, using the same criterion to select ERA5 T2m values reveals that in general, excluding years with
138 high ensemble SNR results in lower absolute deviations from the mean than when the low SNR years are excluded
139 (Fig. 3d). Additionally, these differences overall coincide with regions with significant skill differences (Fig.
140 3c,d). This implies that years with more extreme deviations from the mean (in the observations/reanalysis) may

141 be identified a priori by calculating the ensemble SNR of the forecast, and that forecast systems are in general

142 more skillful when large deviations from the mean occur.

143

145 **Figure 4: Same as Figure 3, but for precipitation.**

146

147 Similar to T2m, the exclusion of years with high SNR also results in lower overall precipitation skill values than

148 the one obtained when excluding low SNR years (Fig. 4a,b). Important skill differences appear in the Iberian

149 Peninsula, Brazil, Australia and Indonesia (Fig. 4c), and in most cases imply a shift from non-significant to

150 significant skill (Fig. 4 a and b, respectively). Contrasting with T2m, the relation between ACC and mean absolute

151 deviation from the mean in the observations is not obvious for precipitation (Fig. 4c,d). To further investigate this

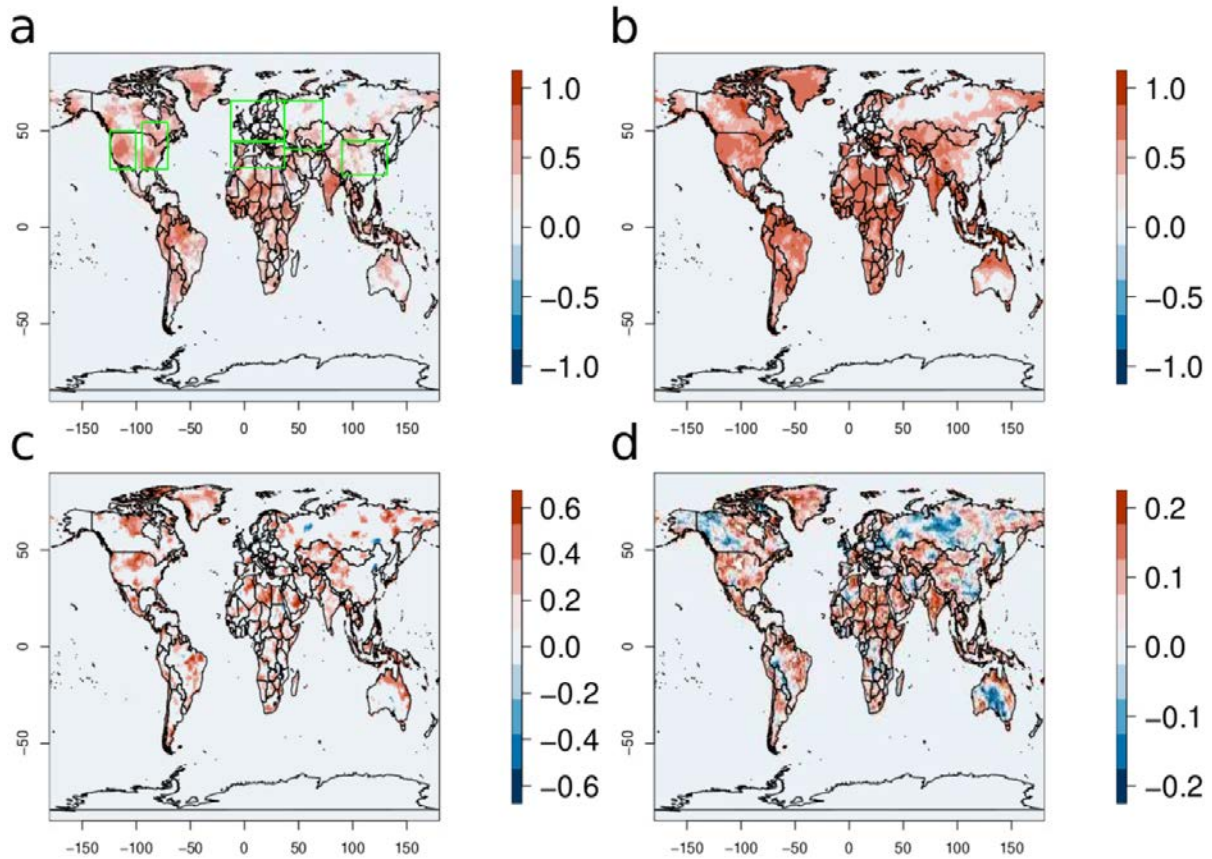
152 behavior, we analyzed the relationship between skill differences and the differences in absolute deviation from

153 the mean for T2m and precipitation, as usual by using the re-forecasts that exclude the 25% of the years with the

154 lowest and the highest SNR, respectively. This analysis (not shown) confirms a statistically robust relationship

155 between skill and large deviations from mean observed precipitation, but still weaker than for T2m.

156

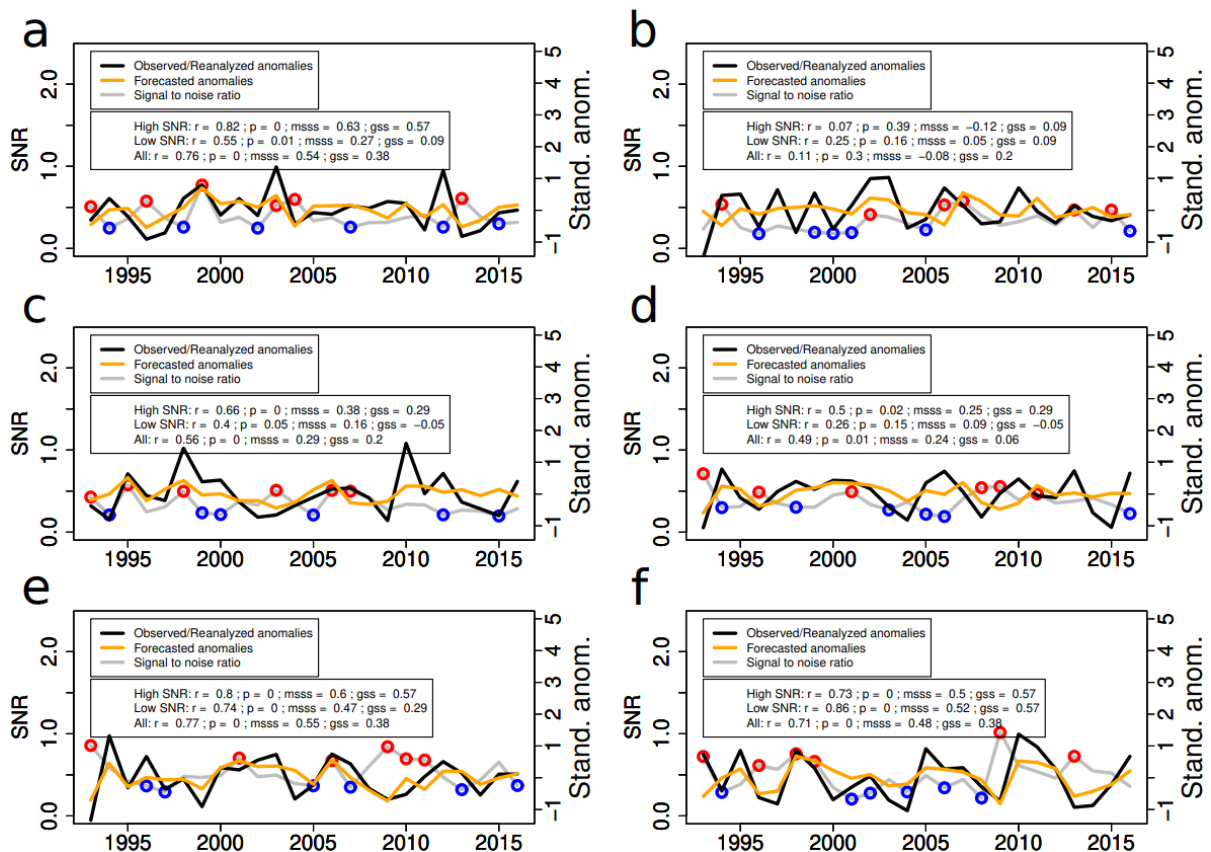
157

158
159 **Figure 5: The same as Figure 3, but for re-forecasts initialized every June. Boxes in (a) show the areas used in Figures**
160 **6 and 7.**

161

162 Figure 5 shows a clearer relation between the impact on skill of the most extreme years in terms of SNR and the
163 absolute T2m anomalies in ERA5, as compared with Figure 3. There is a good correspondence in all continents,
164 including parts of Europe (Fig. 5 c,d). The only difference between the two Figures is that they show the results
165 from re-forecasts with different initialization dates. Both target the boreal summer months (June-August), but
166 Figure 3 shows the results from the May initialization while Figure 5 shows the results from the June initialization.
167 Similar qualitative conclusions can be made for precipitation (not shown).

168

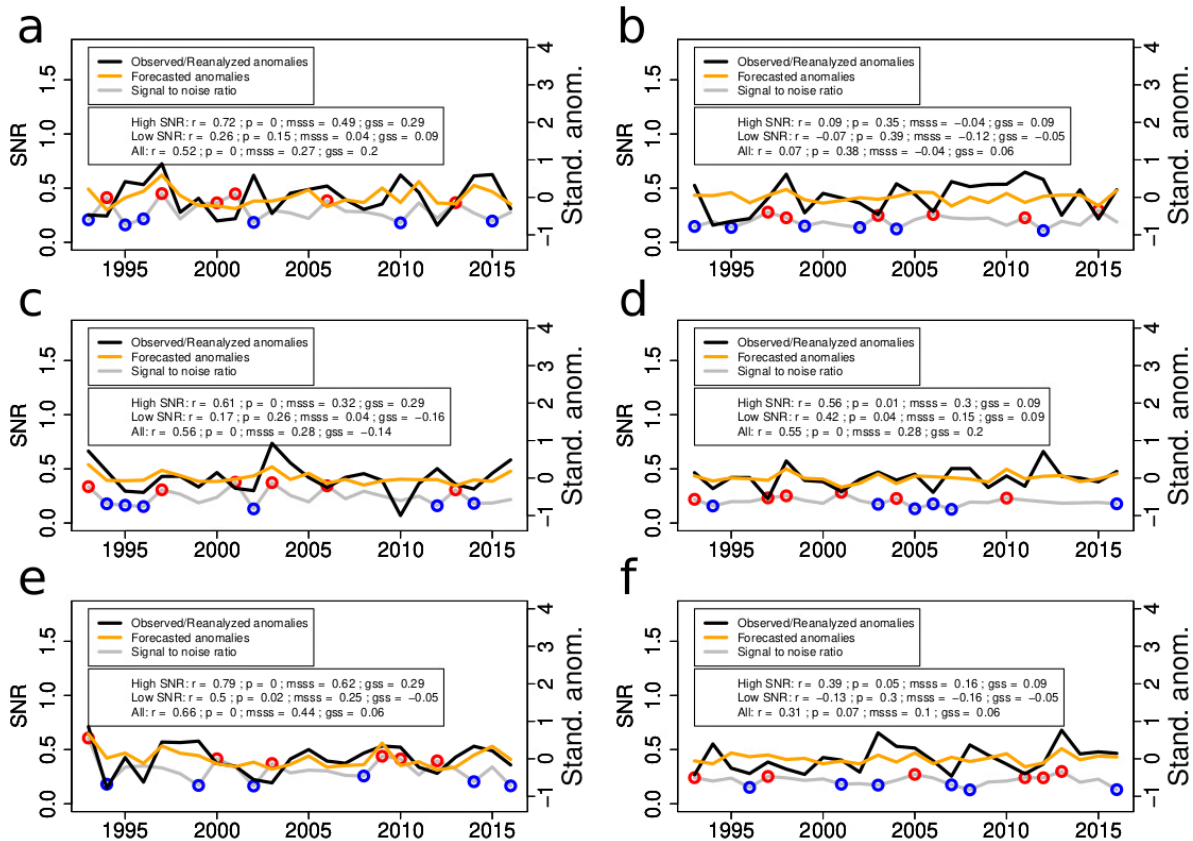
169 In Figure 6 we use the same methodology to sample years based on T2m SNR, but applied to specific northern
170 hemisphere mid-latitude regions: the Mediterranean, North and Central Europe, north western Asia, east Asia,
171 western North America and eastern North America. All the three skill metrics computed show that sampling the
172 18 years with highest SNR, generally results in more skillful T2m predictions than when sampling all 24 years or
173 the 18 years with lowest SNR. The only exceptions are observed in North and Central Europe, where there is
174 basically no skill, as well as in eastern North America, where all the three selection methods show similar skill
175 levels. Examples of successful prediction of extreme (high) T2m years and high SNR are 1999 and 2003 in the
176 Mediterranean, 2002 in northern/central Europe, 1998 in northwestern Asia, 2006 and 1998 in western and eastern
177 North America, respectively. There are also some examples of extreme (high) T2m and low SNR, such as 2012
178 in the Mediterranean, or 1994 and 2016 in East Asia. However, higher overall GSS for the top T2m positive

179 anomalies indicates that on average, sampling years with high SNR results in better prediction of the extreme
180 events.
181
182 A similar analysis on precipitation is shown in Figure 7. The results of precipitation qualitatively agree with those
183 of T2m. Precipitation skill is highest for years with highest SNR and lowest for years with lowest SNR, the only
184 exception being northern/central Europe, again a region with no skill in either precipitation or T2m predictions.
185 Years of successful predictions of low precipitation and high SNR are 1994 and 2000 in the Mediterranean, 2015
186 in northern/central Europe, 1997 and 2001 in East Asia, 2003 in western North America, and 2011 in eastern
187 North America. Similar to T2m, GSS for low precipitation summers is generally higher for the top 18 years (in
188 terms of SNR) than for the bottom 18 years or for all 24 years. It is worth noting that skill scores for precipitation
189 are generally lower than those of T2m. This is primarily due to the lower overall predictability of precipitation
190 compared to T2m.   . Note also that the same conclusions are obtained for both T2m and precipitation when
191 separately sampling only the half of years with highest and lowest SNRs and/or when varying the threshold to
192 define the most extreme years used in the GSS calculations (not shown).
193
194
195

196

197 **Figure 6: Area-averaged time series of observed and predicted, detrended and standardized mean summer T2m (right**
198 **axis) and SNR (left axis) in (a) the Mediterranean (10W-35E, 30-45N), (b) North and Central Europe (10W-35E, 45-**
199 **65N), (c) northwestern Asia (35-70E, 40-65N), (d) East Asia (90-130E, 25-45N), (e) western North America (123-100W,**
200 **30-50N) and (f) eastern North America (90-70W, 30-55N). Skill metrics are provided separately for the 18 years with**
201 **highest SNR (excluding blue circles), the 18 years with the lowest SNR (excluding red circles) and for all 24 years. The**

202 **skill metrics are linear correlation, mean square skill score and Gilbert skill score (See methods). The p-values of the**
203 **linear correlation coefficients are also displayed for each region. The results are from the re-forecasts initialized in**
204 **June.**

205

206

208 **Figure 7: The same as Figure 6, but for precipitation.**

209

210 **4. Discussion**

211

212 The SNR measures the relative weight of the ensemble mean anomalies with respect to the ensemble coherence.

213 Its close resemblance in terms of spatial patterns with a skill metric like ACC indicates that it can provide

214 complementary information related to seasonal climate predictability. We have shown that in regions where the

215 forecasts are skillful, years with high SNR exhibit on average larger observed deviations from the mean than years

216 with low SNR, for both T2m and precipitation. This means that forecast systems are on average more reliable at

217 predicting extremes when there is a higher coherence. This has been further demonstrated for several Northern

218 Hemisphere mid-latitude regions during boreal summer. Ensemble coherence is also a good indicator of T2m and

219 precipitation predictability, although appears to be only suitable for tropical and subtropical locations.

220

221 Despite the well-known limitations of climate forecast systems (e.g. the signal to noise paradox), we have shown

222 that in a multi-system ensemble, the SNR may provide valuable information as it represents an intrinsic measure

223 of reliability for T2m and precipitation forecast. The short span of 24 years defining the common hindcast period

is a limitation of this study. Hence, longer hindcasts would be necessary to obtain more robust results, but are currently unavailable for most of the multiple systems analyzed.

References

Acosta Navarro, J. C., García-Serrano, J., Lapin, V., and Ortega, P. (2022). Added value of assimilating springtime Arctic sea ice concentration in summer-fall climate predictions. Environmental Research Letters, 17(6), 064008.

Baehr, J., Fröhlich, K., Botzet, M., Domeisen, D. I., Kornblueh, L., Notz, D. and Müller, W. A. (2015). The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model. Climate Dynamics, 44(9), 2723-2735.

Below, R., Grover-Kopec, E., and Dilley, M. (2007). Documenting drought-related disasters: A global reassessment. The Journal of Environment & Development, 16(3), 328-344.

Batté L., L. Dorel, C. Ardilouze, and J.-F. Guérémy, (2021). Documentation of the METEO- FRANCE seasonal forecasting system 8. Météo-France, 36 pp., https://www.umr-cnrm.fr/IMG/pdf/system8-technical.pdf.

Bevacqua, E., Zappa, G., Lehner, F., and Zscheischler, J. (2022). Precipitation trends determine future occurrences of compound hot–dry events. Nature Climate Change, 12(4), 350-355.

Buontempo, C., Hanlon, H. M., Soares, M. B., Christel, I., Soubeyroux, J. M., Viel, C. and Liggins, F. (2018). What have we learnt from EUPORIAS climate service prototypes?. Climate Services, 9, 21-32.

Doi, T., Nonaka, M. and Behera, S. (2022). Can signal-to-noise ratio indicate prediction skill? Based on skill assessment of 1-month lead prediction of monthly temperature anomaly over Japan. Front. Clim. 4:887782.

Ceglar, A., and Toreti, A. (2021). Seasonal climate forecast can inform the European agricultural sector well in advance of harvesting. npj Climate and Atmospheric Science, 4(1), 1-8.

Eade, R. et al. (2014). Do seasonal to decadal climate predictions underestimate the predictability of the real world? Geophys. Res. Lett. 41, 5620–5628.

Enenkel, M., Brown, M. E., Vogt, J. V., McCarty, J. L., Reid Bell, A., Guha-Sapir, D. and Vinck, P. (2020). Why predict climate hazards if we need to understand impacts? Putting humans back into the drought equation. Climatic Change, 162(3), 1161-1176.

Gualdi, S., A. Sanna, A. Borrelli, A. Cantelli, M. del Mar Chaves Montero, S. Tibaldi. (2020). The new CMCC Operational Seasonal Prediction System SPS3.5. Centro Euro-Mediterraneo sui Cambiamenti Climatici. CMCC Tech. Note RP0288, 26pp.

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A: Dynamic Meteorology and Oceanography, 57(3), 219-233.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. and Thépaut, J. N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999-2049.

263 Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., and Monge-Sanz,
264 B. M. (2019). SEAS5: the new ECMWF seasonal forecast system. Geoscientific Model Development, 12(3),
265 1087-1117.

266 Lenssen, N. J., Goddard, L., and Mason, S. (2020). Seasonal forecast skill of ENSO teleconnection maps. Weather
267 and Forecasting, 35(6), 2387-2406.

268 MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A. and Madec, G. (2015).
269 Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. Quarterly
270 Journal of the Royal Meteorological Society, 141(689), 1072-1084.

271 Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F. and Xie, S. P. (2021). Initialized
272 Earth System prediction from subseasonal to decadal timescales. Nature Reviews Earth & Environment, 2(5),
273 340-357.

274 Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A. & Yeager, S. (2020). Current
275 and emerging developments in subseasonal to decadal prediction. Bulletin of the American Meteorological
276 Society, 101(6), E869-E896.

277 Mishra, A. K., and Singh, V. P. (2010). A review of drought concepts. Journal of hydrology, 391(1-2), 202-216.

278 Mishra, N., Prodhomme, C., and Guemas, V. (2019). Multi-model skill assessment of seasonal temperature and
279 precipitation forecasts over Europe. Climate Dynamics, 52(7), 4207-4225.

280 Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation
281 coefficient. Monthly weather review, 116(12), 2417-2424.

282 Scaife, A.A., Smith, D. (2018). A signal-to-noise paradox in climate science. npj Clim Atmos Sci 1, 28

283 Smith, D.M., Scaife, A.A., Eade, R. et al. (2020). North Atlantic climate far more predictable than models imply.
284 Nature 583, 796–800.

285 Schnider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., and Ziese, M. (2011). GPCC Full Data
286 Reanalysis Version 6.0 at 1.0∘: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and
287 Historic Data.

288 Toreti, A., Belward, A., Perez-Dominguez, I., Naumann, G., Luterbacher, J., Cronie, O. and Zampieri, M. (2019).
289 The exceptional 2018 European water seesaw calls for action on adaptation. Earth's Future, 7(6), 652-663.

290 Yin, J., Gentine, P., Slater, L., Gu, L., Pokhrel, Y., Hanasaki, N. and Schlenker, W. (2023). Future socio-
291 ecosystem productivity threatened by compound drought–heatwave events. Nature Sustainability, 1-14.

292 World Meteorological Organization (WMO, 2014). Forecast verification for the African severe weather
293 forecasting demonstration projects; No. 1132. Geneva, Switzerland: World Meteorological Organization.

294
295