Natural Hazards
and Earth System
Sciences

Assessment of subseasonal-to-seasonal (S2S) ensemble extreme precipitation forecast skill over Europe

Pauline Rivoire[1,2,3]**, Olivia Martius**[1,2,4]**, Philippe Naveau**[5]**, and Alexandre Tuel**[1,2]

[1]Institute of Geography, University of Bern, Bern, Switzerland
[2]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[3]Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment,
University of Lausanne, Lausanne, Switzerland
[4]Mobiliar Lab for Natural Risks, University of Bern, Bern, Switzerland
[5]Laboratoire des Sciences du Climat et de l'Environnement, ESTIMR, CNRS-CEA-UVSQ, Gif-sur-Yvette, France

Correspondence: Pauline Rivoire (pauline.rivoire@unil.ch)

Abstract. Heavy precipitation can lead to floods and landslides, resulting in widespread damage and significant casualties. Some of its impacts can be mitigated if reliable forecasts and warnings are available. Of particular interest is the subseasonal-to-seasonal (S2S) prediction timescale. The S2S prediction timescale has received increasing attention in the research community because of its importance for many sectors. However, very few forecast skill assessments of precipitation extremes in S2S forecast data have been conducted. The goal of this article is to assess the forecast skill of rare events, here extreme precipitation, in S2S forecasts, using a metric specifically designed for extremes. We verify extreme precipitation events over Europe in the S2S forecast model from the European Centre for Medium-Range Weather Forecasts. The verification is conducted against ERA5 reanalysis precipitation. Extreme precipitation is defined as daily precipitation accumulations exceeding the seasonal 95th percentile. In addition to the classical Brier score, we use a binary loss index to assess skill. The binary loss index is tailored to assess the skill of rare events. We analyze daily events that are locally and spatially aggregated, as well as 7 d extreme-event counts. Results consistently show a higher skill in winter compared to summer. The regions showing the highest skill are Norway, Portugal and the south of the Alps. Skill increases when aggregating the extremes spatially or temporally. The verification methodology can be adapted and applied to other variables, e.g., temperature extremes or river discharge.

1 Introduction

Extreme precipitation is one of the most impactful weather-related hazards in terms of loss of lives, economic impact and number of disasters (see, e.g., the impact of storms and flood quantified in WMO, 2021). Additionally, if several extreme precipitation events occur in close succession (temporal clustering), flooding becomes more likely (Tuel et al., 2022). The successful mitigation of weather-related hazards depends on our ability to forecast them reliably. It is therefore crucial to quantify the skill of precipitation forecasts and improve the predictability of precipitation extremes for better preparedness (Merz et al., 2020).

Subseasonal-to-seasonal (S2S) prediction refers to forecasting on timescales from about 2 weeks to a season. S2S prediction has a large range of applications (White et al., 2017, 2021), including in the humanitarian sector, public health, energy, water management and agriculture. Forecast skill at this timescale is key to better managing natural hazards (Merz et al., 2020). S2S predictions aim to fill the gap between weather forecasts and seasonal outlooks (White et al., 2017). Providing skillful predictions on subseasonal or monthly timescales is challenging (Hudson et al., 2011). Unlike short-range forecasts and seasonal outlooks that have been operational for many years, the S2S timescale was until recently a "predictability desert" (Vitart et al., 2012). The scientific community working with S2S forecasts has been growing rapidly (Mariotti et al., 2018; Merryfield

et al., 2020; Domeisen et al., 2022). Many research organizations actively contribute to improving S2S forecast skill, for example through the "Challenge to improve Sub-seasonal to Seasonal Predictions using Artificial Intelligence" (S2S-challenge, 2021).

Precipitation is a challenging variable to predict, and, as a result, S2S forecasts of precipitation extremes have limited skill compared to other types of hazards (see, e.g., case studies in Domeisen et al., 2022; Tian et al., 2017; Endris et al., 2021). The analysis of S2S precipitation forecast skill could allow for us to identify regions and seasons with good or bad performance of the forecast. With this information, forecast users can know where and when the forecast information is useful or if it would require further improvement (with for example post-processing, as in Specq and Batté, 2020). Skill information is also useful to identify potential sources of predictability and windows of opportunity (i.e., intermittent time periods with higher skill; Mariotti et al., 2020). Most of the existing research on S2S prediction of precipitation extremes focuses on North America (Zhang et al., 2021; De-Florio et al., 2019), Africa (de Andrade et al., 2021; Olaniyan et al., 2018) and Asia (Yan et al., 2021; Li et al., 2019). However little is known about the skill of S2S extreme precipitation prediction over Europe (Monhart et al., 2018; Domeisen et al., 2022). The present article aims to fill this gap.

S2S forecasts are ensemble forecasts that consist of several equally probable members, i.e., runs of the same numerical model with slightly different initial conditions (World-Climate-Service, 2021). Forecast skill is typically assessed with hindcasts. Hindcasts are forecasts run for past dates over sufficiently long time periods (about 20 years) to assess the quality of the forecast and to identify and correct model biases (e.g., Huijnen et al., 2012; Manrique-Suñén et al., 2020). The goal here is to quantify S2S forecast skill for extreme precipitation events over Europe using the forecast and hindcast data from the European Centre for Medium-Range Weather Forecasts (ECMWF; Vitart, 2020), one of the most frequently used and most skillful S2S modeling systems (de Andrade et al., 2019; Li et al., 2019; Stan et al., 2022; Domeisen et al., 2022).

Common metrics to evaluate the bias and the accuracy – and hence the skill – of ensemble forecasts include the mean absolute error, the probability integral transform, the interquartile range, the continuous ranked probability score (CRPS; Hersbach, 2000; Gneiting et al., 2007; Crochemore et al., 2016; Monhart et al., 2018; Pic et al., 2022), the Brier score (Brier, 1950) and the mean square skill score (Specq and Batté, 2020). However, these metrics capture the mean behavior of a variable: most are not directly suited to verifying extreme events. The CRPS can be adapted to focus on extremes, using the threshold-weighted CRPS (Gneiting and Ranjan, 2011; Allen et al., 2021) or using extreme value theory (Taillardat et al., 2022). Another option to verify extreme events is the relative operating characteristic (ROC): it can be used to measure the ability of the ensemble fore-

cast to identify above-normal precipitation events (Domeisen et al., 2022; Monhart et al., 2018). In this study, we transform precipitation extremes into binary "threshold exceedance events", where the threshold is the daily-precipitation 95th percentile. The Brier score is usually employed to verify the binary forecast. However, it has limitations because of the unbalanced categories in our case. The extreme-event dataset is composed of 95 % zeros and 5 % ones. Large parts of the forecast and observation datasets match because of the large presence of "0" values (daily precipitation lower than the 95th percentile) in both datasets. To address this issue, we also use a binary loss index focusing on extremes ("1" values). We assess the extreme events by proposing and using a simple extension of the binary loss score as introduced by Legrand et al. (2022) to ensemble forecasts. This metric considers only the case of the occurrence of an extreme event in the forecast or in the observation or in both but not the non-events (see Sect. 2.3.2). This has the advantage that the score is not dominated by the correct prediction of non-events. We compare our novel skill score to the classical Brier score (Brier, 1950). To overcome the double-penalty issue (i.e., when a location or timing error in the forecast is penalized by both a false alarm and a missed event), we allow for flexibility by aggregating the forecast information in spatial and temporal windows (Ebert et al., 2013).

This article is structured as follows. Section 2 contains a description of the forecast and verification data and the methods, including the Brier score (Brier, 1950) and a binary loss index (adapted from Legrand et al., 2022). We present the results of the analysis in Sect. 3. We discuss these results, draw conclusions and give an outlook in Sect. 4.

2 Data and methods

2.1 Data

We use ECMWF's S2S precipitation hindcast data (cycle 47r2; ECMWF, 2021; Vitart, 2020; ECMWF, 2022a) from 2001 to 2020. They are composed of 11 ensemble members, initialized twice a week and run for 46 d. We focus on Europe, in the spatial box $[30° \text{N}, 72° \text{N}] \times [-15° \text{E}, 49.5° \text{E}]$. The hindcast period covers 20 years with 2080 forecast initializations between 4 January 2001 and 30 December 2020 (twice a week, on Monday and Thursday). The data were downloaded at the model spectral resolution O320 (ECMWF, 2022b, c) and regridded for the analysis to a $0.5° \times 0.5°$ regular grid using a first-order conservative remapping (Jones, 1999; CDO, 2018).

ERA5 precipitation (Hersbach et al., 2019) is used here as the verification dataset. The choice of a reanalysis dataset is motivated by its continuous spatial and temporal availability and by it avoiding the uncertainties due to the inherent spatial sparsity of weather station networks (Hofstra et al., 2009; Rivoire et al., 2021). Daily precipitation are extracted

over the same time period, from 4 January 2001 to 30 December 2020 plus 46 lead time days, i.e., 14 February 2021, with a spatial resolution of $0.5° \times 0.5°$. For the sake of simplicity, "observation" refers to ERA5 in the remainder.

2.2 Definition of extreme events

We define precipitation extremes as binary exceedances of daily precipitation accumulation above its 95th seasonal all-day percentile (i.e., over all days in March–April–May, MAM; June–July–August, JJA; September–October–November, SON; or December–January–February, DJF). Figure A1 in the Appendix shows the 95th percentile (Q_{95}) in ERA5 in Europe, for the period from 4 January 2001 to 14 February 2021. For the hindcast data, we also compute percentiles separately for each lead time: for a given lead time day and a given season, Q_{95} is computed from daily precipitation of all the ensemble members pooled together. Figure 1 shows the bias in this percentile between the forecast and ERA5 data for four different lead times. In this figure and all the following ones, only values at grid points where Q_{95} in the observations is greater than $5\,\mathrm{mm\,d^{-1}}$ are shown. For a lead time of 1 d, the forecast generally underestimates Q_{95}. For lead times between 2 and 46 d, some regions have a positive bias (central Europe in spring and summer) and some have a negative bias (the Alps in summer, autumn and winter and Norway in spring, autumn and winter; see Fig. 1). Generally over Europe, the bias depends on the lead time and on the season. However, the bias over oceans often has the opposite sign of the bias over land.

2.3 Metrics

We use the Brier score and a binary loss index to assess the forecast skill in extreme events. We compute the Brier score and the binary loss index for the extended winter season (NDJFMA, i.e., November to April) and extended summer season (MJJASO, i.e., March to October). When defining the extremes (see previous section) we used 3-month-long seasons because of the strong seasonal cycle in extreme precipitation (see Fig. A1). The choice of extended seasons for the skill analysis is a compromise between having enough extreme events for a robust analysis and capturing the seasonality of the forecast. As a consequence, the probability of the extreme events is no longer exactly 0.05 if extreme events are not homogeneously distributed within the MAM and SON seasons.

2.3.1 Brier score

The Brier score B is defined as the mean square difference between forecast probability and binary observations (Brier, 1950): $B = \frac{1}{n_\mathrm{D}} \sum_{i=1}^{n_\mathrm{D}} (f_i - Y_i)^2$, where n_D is the total number of days (i.e., the number of initializations in the given extended season: about 1040 per lead time, which is half the number

of initializations per year); Y_i the binary observation of extremes for day i ($Y_i = 1$ if the daily precipitation exceeds the 95th percentile, and $Y_i = 0$ otherwise); and f_i is the forecast probability of extreme occurrence for day i, i.e., the mean of the ensemble members $f_i = \frac{1}{M} \sum_{m=1}^{M} F_{(i,m)}$, with M the number of ensemble members (here $M = 11$) and $F_{(i,m)}$ the binary forecast for a given ensemble member m for day i.

B is negatively oriented (the lower, the better). The climatological Brier score B_clim is used as a reference value for the skill calculation:

$$B_\mathrm{clim} = \frac{1}{n_\mathrm{D}} \sum_{i=1}^{n_\mathrm{D}} (p - Y_i)^2, \tag{1}$$

where p is the extreme-climatological-event probability. Note that the value of this probability is not exactly 0.05, as two of the 3-month seasons are split to form the extended seasons. p is therefore computed empirically.

The forecast is skillful if its Brier score is lower than the climatological Brier score. These scores can be compared using the Brier skill score (BSS):

$$\mathrm{BSS} = 1 - \frac{B_\mathrm{hind}}{B_\mathrm{clim}}. \tag{2}$$

BSS varies between $(-\infty, 1]$ and is positively oriented (the closer to 1, the better). For a given lead time day, a forecast has skill if $\mathrm{BSS} > 0$. From here on, the expression "the last skillful day" refers to the largest lead time day with skill.

2.3.2 Binary loss index

Legrand et al. (2022) studied in detail a risk function defined as the ratio between (i) the empirical probability of having an extreme event in either the observation dataset or the forecast dataset and (ii) the empirical probability of having an extreme event in the observations or the forecast (including having an event in both datasets). In our context, the risk function can be written as follows:

$$R^{(u)}(X) = \frac{\mathbb{P}(X^{(u)} \neq Y^{(u)})}{\mathbb{P}(Y^{(u)} = 1 \text{ or } X^{(u)} = 1)}, \tag{3}$$

where $Y^{(u)}$ is the binary observation with $Y^{(u)} = 0\,(Y^{(u)} = 1)$ if the observed daily precipitation is lower (greater) than a certain threshold u and $X^{(u)}$ is the binary forecast with $X^{(u)} = 0\,(X^{(u)} = 1)$ if the predicted daily precipitation is lower (greater) than u.

The risk function $R^{(u)}$ focuses on how well the 1 values (extreme-event days) match between observation and forecast. It does not take into account steps when neither the forecast nor the observation experiences an extreme event. $R^{(u)}(X)$ varies between $[0, 1]$ and is negatively oriented (the closer to zero, the better the forecast is). The strength of $R^{(u)}(X)$ is its asymptotic behavior: even for very rare events, both the over-optimistic and over-pessimistic forecasts will
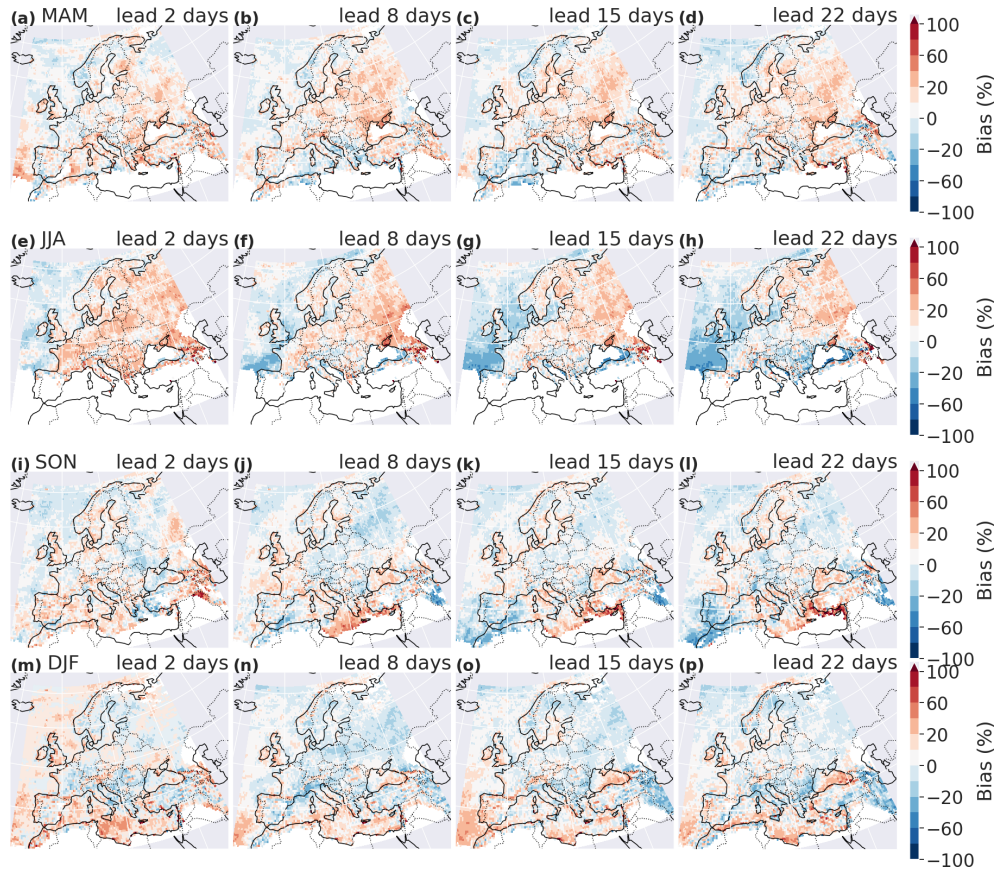
Figure 1. Forecast bias in the 95th percentile (Q_{95}) compared to ERA5, for spring (MAM, **a–d**), summer (JJA, **e–h**), autumn (SON, **i–l**) and winter (DJF, **m–p**) at a 2 d (**a, e, i, m**), 8 d (**b, f, j, n**), 15 d (**c, g, k, o**) and 22 d (**d, h, l, p**) lead time. Grid points with $Q_{95} \leq 5 \, \text{mm} \, \text{d}^{-1}$ are displayed in white.

be penalized. In other words, even for a very large threshold u, i.e., $Y = 1$ for very rare occasions (but at least once), if the forecast is too optimistic and $X = 0$ for all time steps, then $R^{(u)}(X) = 1$ ("naive" classifier; Legrand et al., 2022). A very pessimistic forecast will be penalized the same way ("crying-wolf" classifier; see Legrand et al., 2022). The commonly used Brier score rather assesses the average behavior, with a very weak penalty for underrepresented classes. Because all days are compared, the assessment of rare extreme events (missed, false alarm or hit) by the Brier score is lost among the huge number of correctly predicted 0 values.

The function $1 - R^{(u)}(X)$ can be understood as a critical success index for rare events (Schaefer, 1990; Legrand et al., 2022), with asymptotic properties proven by Legrand et al. (2022), such as the link to the extremal index (we refer the reader to their article for more details).

The risk function $R^{(u)}(X)$ is initially designed for deterministic forecasts. We extend it here to an index for ensemble forecasts by comparing the observed exceedances with the median member of the forecast exceedances F^{med}. There are 11 members in the ECMWF precipitation hindcast data: for a given location, a given initialization date and a given

lead time, $F^{\text{med}} = 1$ if at least six ensemble members predict extreme precipitation, and $F^{\text{med}} = 0$ otherwise. We take here the median forecast across members, but in practice $F^{\text{med}} = 1$ could be set to 1 only if fewer or more than six members forecast extreme precipitation. The choice depends on the risk aversion of the users and is discussed in Sect. 4.

This adapted index is later on called the binary loss index (BLI_m, m indicating the median of the ensemble members). It is defined by

$$\text{BLI}_m = \frac{N_1^{\text{med}}}{N_2^{\text{med}}}, \tag{4}$$

where N_1^{med} is the number of days when the observation and the ensemble median disagree, i.e., $N_1^{\text{med}} = \#\{j \mid F_j^{\text{med}} \neq Y_j\}$, and N_2^{med} is the number of days when an extreme event occurs in either the observation or the ensemble median or in both of them, i.e., $N_2^{\text{med}} = \#\{j \mid \left(F_j^{\text{med}} = 1 \text{ or } Y_j = 1\right)\}$. In other words, N_1^{med} is the number of false positives and false negatives and N_2^{med} is the number of true positives, false positives and false negatives.
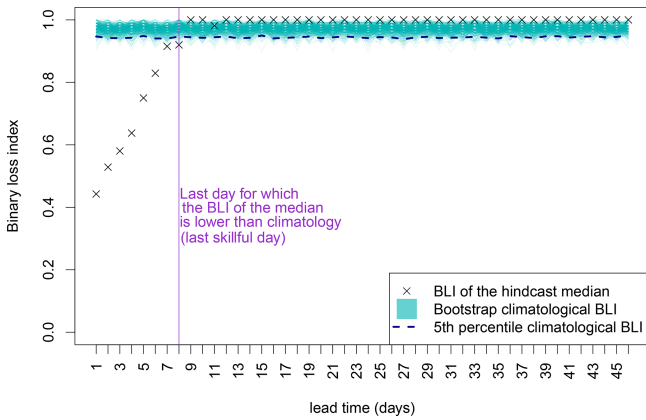
Figure 2. Definition of the last skillful day for BLI: example for one grid point and one season.

To measure the lead time dependence of the skill, BLI is computed for each lead time day. Note that if the forecast F and the observation Y are independent (i.e., the forecast has no skill) and if $\mathbb{P}[F = 1] = \mathbb{P}[Y = 1] = \alpha$, then BLI $= \frac{2-2\alpha}{2-\alpha}$ (here, $\alpha = 0.05$ for daily exceedances in a given season). In our case, $\mathbb{P}[F = 1]$ is not exactly equal to $\mathbb{P}[Y = 1]$ because the index is computed on extended seasons and not on 3-month seasons.

The climatological value of BLI, referred to as BLI_{clim}, is used as reference value. We compute confidence intervals for BLI_{clim} with a bootstrap procedure to determine if the forecast is skillful, i.e., if BLI is significantly lower than BLI_{clim}. For a given bootstrap step, a random time series is formed by drawing values in the observation time series. BLI is computed with this random time series as the forecast. For a given lead time day, a forecast is deemed to be significantly skillful if the BLI of the median member of the forecast (F^{med}) is lower than the 5th percentile of the confidence interval on BLI_{clim}. Like for the Brier score, we compute the "last skillful day" for BLI, with the same definition (largest lead time day with skill; see Fig. 2 for an example).

2.3.3 Spatio-temporal extension of the metrics

Requiring an exact match of events in the forecast and the observations on the same day and at the same grid point is very strict. Indeed, precipitation is a complex variable to forecast precisely in space and time. A forecast may contain useful information, even if the forecast does not predict the event exactly on the same day or at the same location as in the observation but in a close neighborhood. Moreover, a temporal lag or a spatial shift between the observation and the forecast is penalized twice, by (1) a missed event at the observed time/location of the event and (2) a false alarm at the erroneously predicted time/location of the event (double-penalty issue; see, e.g., Ebert et al., 2013). We therefore also compute skill scores on data aggregated in space and time, which

allows for some flexibility in the exact location or exact timing of the events. The spatial and the temporal aggregations are conducted independently to analyze the individual impact of each aggregation. Both the spatial and the temporal neighborhoods are non-overlapping to consider each extreme event only once. The spatio-temporal extensions are applied before computing the median member.

The temporal aggregation consists in counting the number of extreme events N^t in a 7 d window. We then translate it into a binary series E_n^t, given a minimum number of events n in the window ($n = \{1, \ldots, 7\}$):

$$E_n^t = \begin{cases} 1 & \text{if } N^t \geq n \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Figure 3 provides an example for the definition of E_n^t. To the various binary series (one for each n value), we apply the Brier score and BLI to quantify forecast skill. We estimate climatological skill in a way that conserves the temporal structure of the climatology. We randomly select the beginning of the 7 d time windows in the observation. The following 6 d is not randomly selected; it is the 6 d actually following the beginning of time window in the time series of observations.

The spatial aggregation is performed by counting extreme precipitation events in neighborhoods. Like for the temporal aggregation, we define two categories, depending on whether or not the count of events N^s in the spatial neighborhood exceeds some threshold n (see Fig. 4 for an example):

$$E_n^s = \begin{cases} 1 & \text{if } N^s \geq n \\ 0 & \text{otherwise.} \end{cases}$$

Precipitation includes some spatial structure, i.e., spatial dependence between points in a neighborhood. When computing the climatology for both scores, the spatial structure is conserved: for one step of the bootstrap, only the date is randomly chosen, and the spatial neighborhood is the observed neighborhood for that day. We define the neighborhoods as square boxes of about 150 km × 150 km, i.e., boxes with a latitudinal extent of 1.5° N (three grid boxes) and with a longitudinal grid extent that depends on the latitude: from 1.5° E at 30° N (three grid boxes) to 4.5° E at 70° N (nine grid boxes); see Fig. D1 in the Appendix for an illustration.

3 Results

3.1 Daily and local comparison

We begin by discussing the forecast skill at the daily and grid-point scale. BLI indicates more skill during the extended winter (skill for up to 11 d and many regions with a last skillful day greater than 7 d) than during the extended summer (last skillful day below 6 d for most grid points); see Fig. 5.
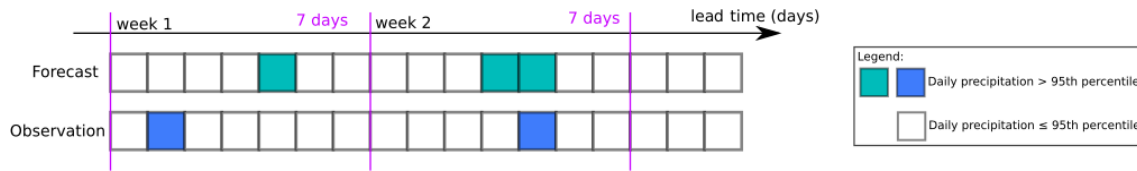
Figure 3. Illustration of the weekly aggregation of extremes at one grid point. During week 1, the forecast predicts one extreme and one extreme is observed. For both the forecast and the observation, the number of extreme events in the 7 d window is greater than or equal to 1: $E_1^t = 1$ for the two datasets. For both datasets, the number of events in the 7 d window is lower than n for $n \geq 2$: $E_n^t = 0$ for the two datasets. During week 2, one extreme is observed but the forecast predicts two events. For both datasets, the number of extreme events in the 7 d window is greater than or equal to 1: $E_1^t = 1$ for the two datasets. For the observation, the number of events in the 7 d window is lower than 2 ($E_2^t = 0$), and this number is greater than or equal to 2 for the forecast ($E_2^t = 1$). For the configuration with $n \geq 3$, $E_n^t = 0$ for both datasets.
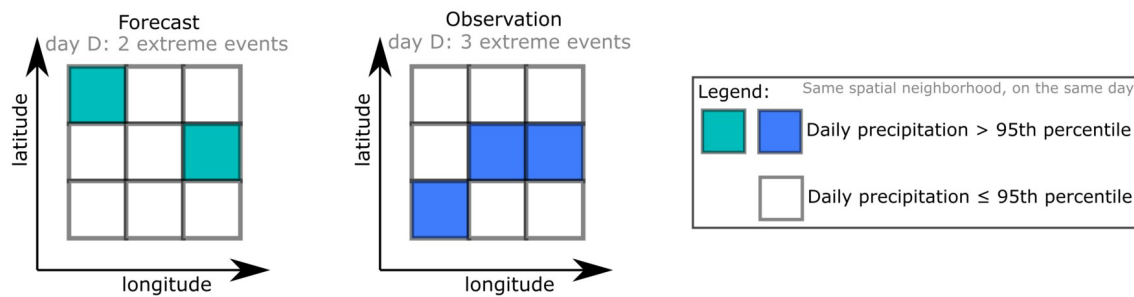
Figure 4. Illustration of the spatial aggregation of extremes in one neighborhood. The forecast indicates two extremes in the spatial neighborhood, and three events are observed. For both datasets, the number of extreme events in the neighborhood is greater than or equal to 1 ($E_1^s = 1$) and greater than or equal to 2 ($E_2^s = 1$). For three events or more, $E_3^s = 0$ for the forecast and $E_3^s = 1$ for the observation. For four events or more, $E_n^s = 0$ for both datasets.

Regions with high skill are Norway, the Alps and the western half of the Iberian Peninsula in the extended winter and the Bay of Biscay, the south of France, Norway, central Europe and the south of the Alps in the extended summer. The BLI skill score is less conservative than the Brier skill score; however the spatial patterns are similar for the two metrics (Fig. B1 in the Appendix). That is, the last skillful day for the Brier skill score is overall smaller than the last skillful day for BLI, but both the Brier score and BLI show the same regions with high and low skill of the forecast for precipitation extremes, in summer and winter.

3.2 Temporal aggregation

The 7 d extreme-precipitation-event counts are also better predicted during the extended winter than during the extended summer (Fig. 6). For the category "one event or more occurred during the 7 d", the forecasts at most grid points still have skill for lead times into the second week, i.e., days 8–14, in extended winter. BLI decreases as the number of events per week increases; however, the spatial patterns remain the same. The regions where temporal clustering is more skillfully forecasted are the Iberian Peninsula, Norway and the northern Mediterranean coastline (especially in winter). The Brier score confirms these results, with similar patterns (see Fig. C1 in the Appendix).

3.3 Spatial aggregation

Extended winter forecasts for spatially aggregated extremes are also more skillful than summer ones (see Fig. 7). The last skillful day is greater when spatially aggregating that for the local analysis, but the two configurations have a similar spatial pattern. In extended winter, for one event or more in the neighborhood, the last skillful lead time reaches up to 11 d in many regions: the western Iberian Peninsula, the Norway coast and the west-facing coasts in general. In extended summer, the last skillful lead time is between 8 and 11 d on the Atlantic coast of France, Italy, western Europe and the coasts of the Iberian Peninsula. The spatial skill pattern remains similar with an increasing number of events per neighborhood, but the skill decreases.
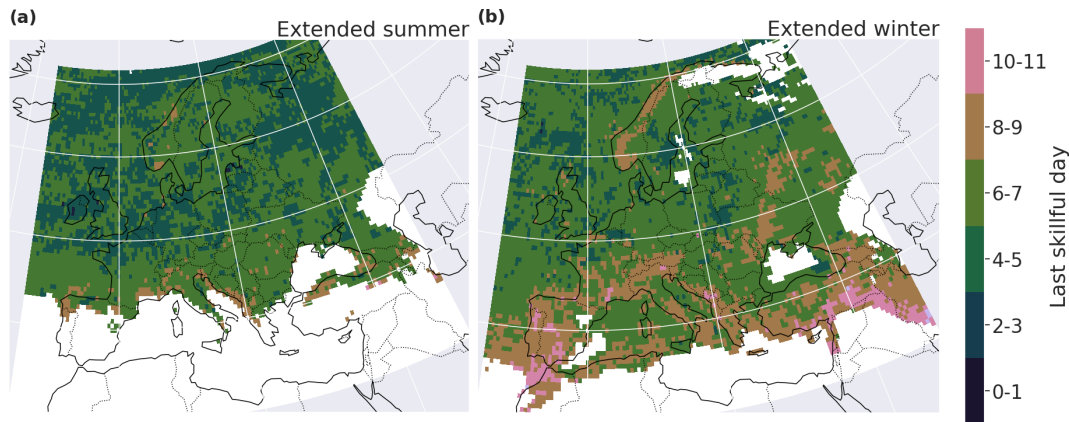
Figure 5. Last skillful day for BLI for a local and daily comparison, in extended summer **(a)** and extended winter **(b)**.
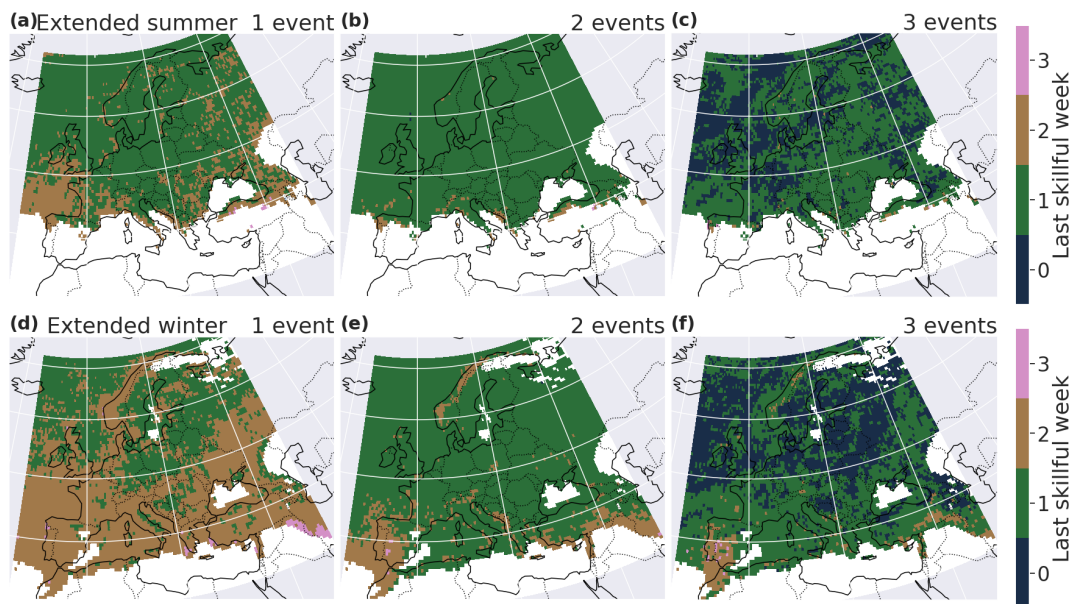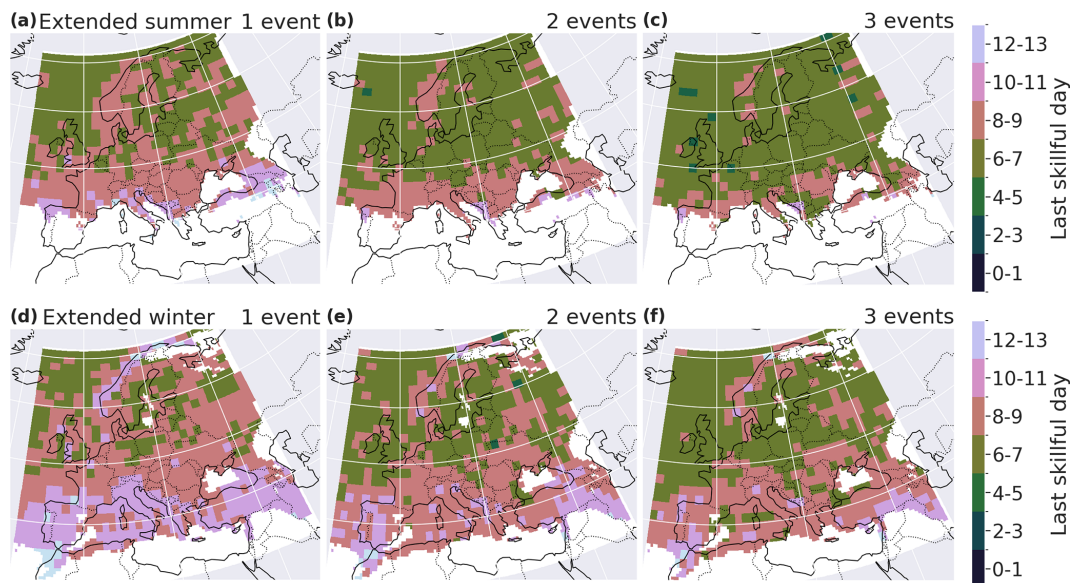

Figure 6. Last week of skill for BLI in extended summer **(a–c)** and extended winter **(d–e)** for a minimum of one **(a, d)**, two **(b, e)** and three **(c, f)** events in a 7 d window. A last skillful week equal to 0 means that, for the count of extremes during the first-week lead time, BLI of the forecast is not significantly lower than BLI$_{clim}$.

Figure D2 in Appendix shows maps of the last lead time day with a positive Brier skill score, for different numbers of events aggregated spatially, in extended summer and extended winter. The regions with higher skill are the same for the Brier score and for BLI. The spatial pattern of the skill also remains similar with an increasing number of events per neighborhood.

4 Discussion and conclusion

In this paper, we assess forecast skill of extreme precipitation occurrence over Europe in the ECMWF S2S model. Extremes are defined as exceedances over the seasonal 95th percentile. We conduct a verification against ERA5 precipi-

tation with the binary loss index (BLI) and the Brier score. We extend the binary loss score studied by Legrand et al. (2022), which was designed for deterministic forecasts only, to ensemble forecasts. We define BLI as the binary loss score calculated for the ensemble median member of binary exceedances. The choice of the median member was motivated by a trade-off between false alarms and missed events. The skill will be different when one chooses a lower percentile of the ensemble members to compute BLI (risk-averse setting) or when one chooses a high percentile of the ensemble members (risk-loving). BLI has the advantage of focusing exclusively on extreme-event occurrence (hit, false alarm or miss) and is not biased by the high counts of extreme-event non-occurrence. BLI is qualitatively compared with the Brier

Figure 7. Last day of skill for BLI in extended summer **(a–c)** and extended winter **(d–f)** for a minimum of one **(a, d)**, two **(b, e)** and three **(c, f)** events in neighborhoods of 150 km × 150 km.

score; the skill scores of the two metrics agree very well over Europe. Despite the great importance of accurately forecasting rare extremes, the Brier score does not give a special weight to underrepresented classes. Therefore, by design, BLI should be preferred to the Brier score when assessing the forecast skill for very rare events. With further research, a probability score for ensemble forecasts could be developed from BLI.

The S2S forecasts have overall higher skill in predicting extreme precipitation events in winter than in summer. A likely explanation resides in the fact that precipitation over Europe mainly results from large-scale processes during winter but from small-scale convective events in summer. Predicting small-scale events is indeed more challenging than predicting large-scale ones (Haylock and Goodess, 2004; Kenyon and Hegerl, 2010). This result is in agreement with the existing literature on S2S prediction in other regions (Tian et al., 2017; Kolachian and Saghafian, 2019). Norway, Portugal and the south of the Alps are regions with the most skill. The orography seems to be a source of skill (like in Norway, the Pyrenees and the south of the Alps): the forecast seems to better capture precipitation events where the complex topography acts as a forcing for precipitation. The Mediterranean region exhibits relatively good skill in winter. Similarly, coastal regions in general have a higher skill compared to continental regions. A potential explanation for this difference is that the water transported from the ocean first rains next to the coast; it is more challenging to predict where the remaining water in the atmosphere will rain down on continental regions because land–atmosphere interactions introduce uncertainty. A follow-up study could further inves-

tigate these hypotheses on the physical reasons behind the spatial and seasonal heterogeneity of the skill.

Allowing for temporal or spatial flexibility in the evaluation of the forecast extremes confirms the skill patterns, bringing robustness to the analysis. The skill for the spatially aggregated precipitation is slightly larger than for the local analysis, as it is easier for the forecast to have an event that matches with observation on a larger grid. The spatial aggregation conducted here could be adapted for an impact-oriented analysis, by aggregating, e.g., over catchments to evaluate the predictability of heavy precipitation that can potentially result in floods or by analyzing multi-day heat waves.

We additionally investigated the effect of European weather regimes on the forecast skill (as defined in Grams et al., 2017), as the forecast skill of the weather regimes themselves can largely differ (Büeler et al., 2021). We computed the forecast skill independently for positive phases and negative phases of the North Atlantic Oscillation (NAO). The forecast skill does not exhibit a strong dependence on the NAO phase, although the data were also spatially aggregated to increase robustness (not shown). This absence of signal should be confirmed with a deeper analysis, by considering some time lag or seasonality for the influence of the teleconnection patterns (Tabari and Willems, 2018) or by aggregating over larger spatio-temporal neighborhoods, to increase the robustness. Other teleconnection patterns could be investigated, such as Scandinavian and eastern Atlantic patterns, El Niño–Southern Oscillation, the Atlantic Multidecadal Oscillation (Casanueva et al., 2014), or the state of the stratosphere (Domeisen et al., 2019).

An assessment focused on the precipitation intensity could extend the verification; the precipitation forecast data would then require calibration (Gneiting et al., 2007; Specq and Batté, 2020; Crochemore et al., 2016; Monhart et al., 2018; Huang et al., 2022). An extension of the CRPS would be an option to measure the intensity forecast skill with a focus on heavy precipitation, like the threshold-weighted CRPS (see, e.g., Pantillon et al., 2018; Allen et al., 2021) or using extreme value theory (Taillardat et al., 2022). Post-processing the hindcast data and analyzing the paradigm of "maximizing the sharpness of the predictive distributions subject to calibration" could also be an extension of this work (Gneiting et al., 2007); the usual evaluation metrics – the probability integral transform histogram, marginal calibration plots, the sharpness diagram – could be applied with a focus on extremes.

Note that for practical applications, one needs caution to interpret the skill in an absolute way for two reasons: (i) a skillful forecast does not mean that the forecast is also a useful forecast for practical applications and (ii) the absolute last skillful day depends on the choice of the member for the daily predictor (here, the median member). (i) If BLI is equal to 0.8 but is outside of the climatological confidence interval, the forecast is better than the climatology and therefore skillful. However, it also means that only 25 % of the extremes are caught by the forecast (by simple transformation of $BLI = \frac{FN+FP}{TP+FN+FP} = 0.8$, where FN denotes the false negatives, FP the false positives and TP the true positives). For 75 % of the time, the forecast either erroneously predicted an extreme (false alarm, FP) or did not predict an extreme that occurred (miss, FN). The definition of the last skillful day can be adapted depending on the usage of the forecast. The definition can be more conservative, e.g., using the last lead time day for which at least 75 % of the extreme events are caught (rather than a comparison to the climatology) or using a smaller percentile of the members rather than the median member. (ii) The last skillful day is larger when choosing the maximum member as the daily predictor (i.e., $F^{\mathrm{med}} = 1$ if at least 1 ensemble member predicts extreme precipitation, and $F^{\mathrm{med}} = 0$ otherwise). This is due to the number of TPs not collapsing to zero with increasing lead times: the condition "at least one member predicts daily precipitation is greater than the 95th percentile" is very easily satisfied. By "chance", the maximum member predicts many TPs, still compensating a bit for the FPs for large lead times. For the choice of the member, we recommend considering a good balance between FNs and FPs. However, it is important to note that the spatial pattern of skill does not depend on the choice of the member. The regions with a relatively large skill (e.g., Norway, Portugal, west coasts in winter) remain the same, independently of the choice of predictor (minimum, median or maximum member). Following these two remarks, we emphasize that our aim here was to provide a robust qualitative assessment by identifying regions of skill

and challenging regions for the forecast model to predict precipitation extremes on the S2S timescale.

Checking if a value of BLI is significant is a kind of hypothesis test that is repeated for a large number of grid points. One could argue that some regional significance should be investigated. However, when displaying the local significance as "largest lead time day with skillful forecast", the results are continuous rather than a strict "yes or no" response. Moreover, the spatial coherence of the results confirms the robustness of the method.

Our method to assess extremes can also be applied to other variables, such as consecutive days of high temperature, river discharge, etc. Considering the other end of extremes, evaluating the skill of forecasts to predict droughts is also of crucial importance. For droughts, the persistence of dry periods matters, rather than the occurrence of precipitation. The method could be adapted accordingly, e.g., by adjusting the definition of temporal aggregation introduced in this study.
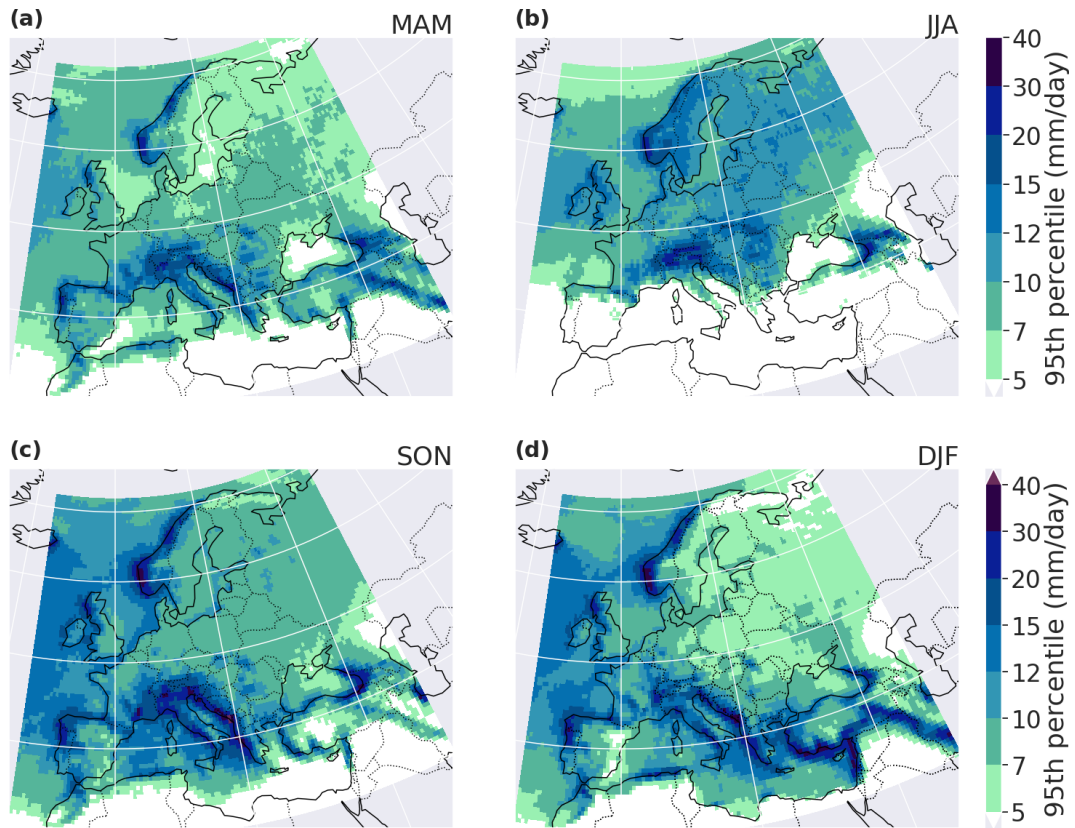
Appendix A: The 95th percentile

Figure A1. The 95th percentile of daily precipitation in ERA5, 2001–2021.

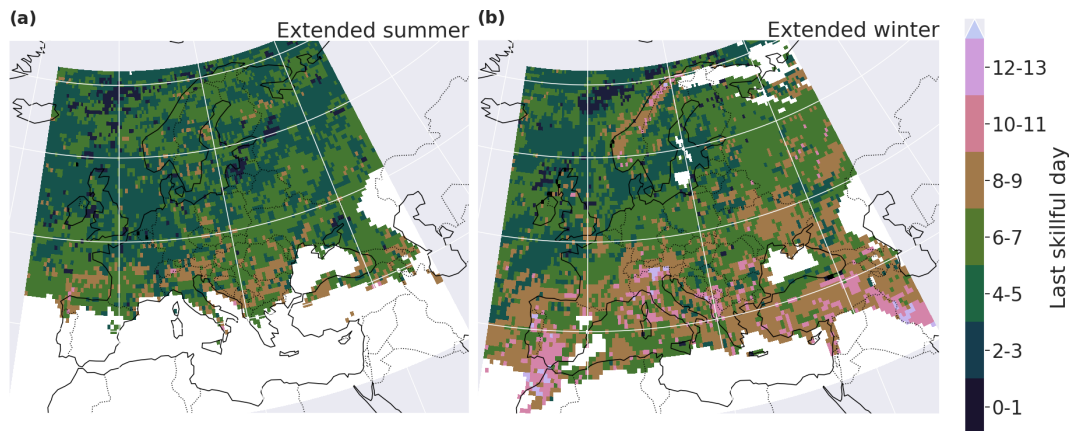## Appendix B: Local and daily comparison of extremes

Figure B1. Last skillful day for the Brier skill score for local and daily comparison, in extended summer **(a)** and extended winter **(b)**.
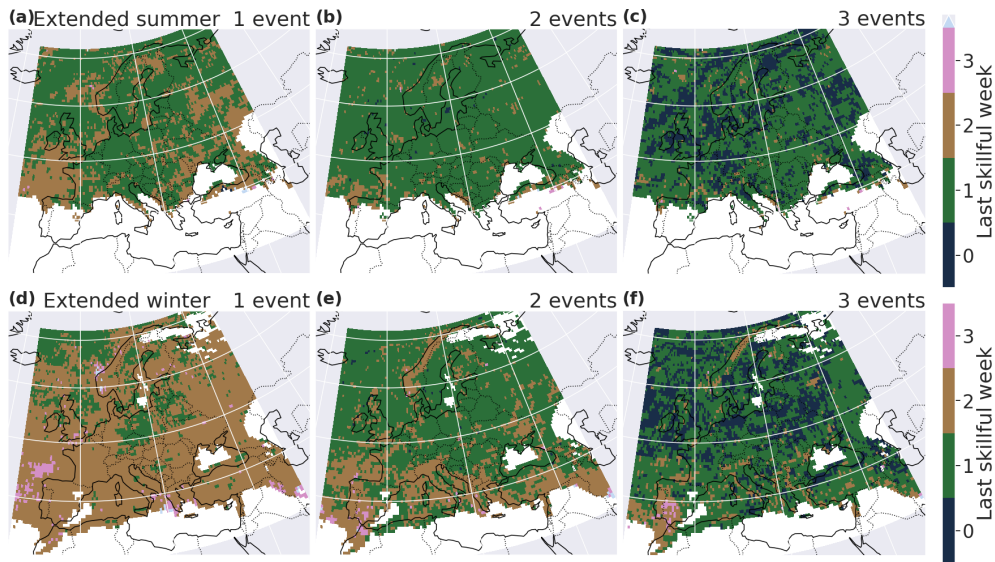
Appendix C: Temporally accumulated extremes

Figure C1. Last week of skill for the Brier skill score in extended summer **(a–c)** and extended winter **(d–f)** for a minimum of one **(a, d)**, two **(b, e)** and three **(c, f)** events in a 7 d window.

Appendix D: Spatially accumulated extremes
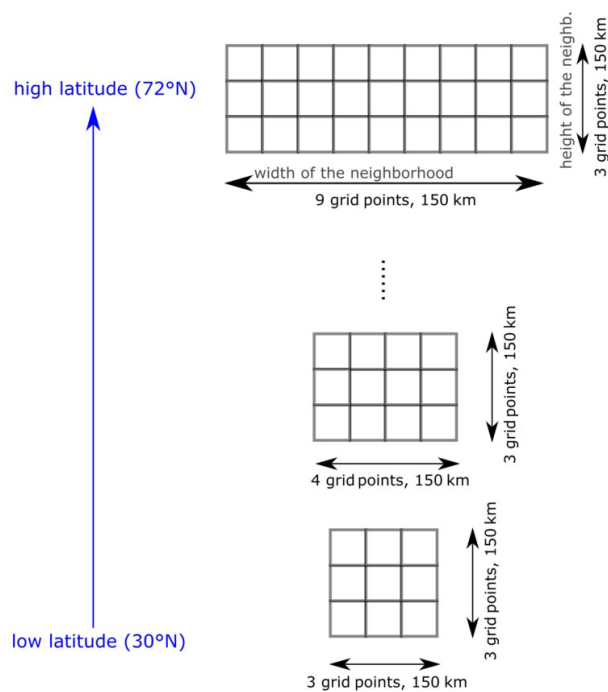
Figure D1. Illustration of the width of the spatial neighborhood, in terms of grid points, depending on the latitude for a constant width in kilometers (and for a constant area).
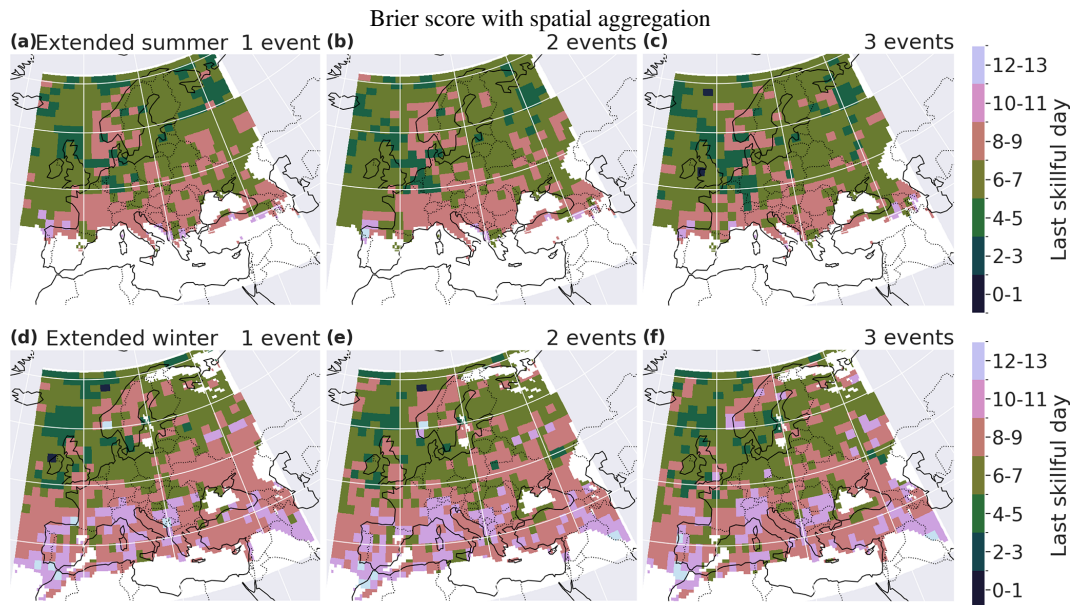
Figure D2. Last day of skill for the Brier skill score in extended summer **(a–c)** and extended winter **(d–f)** for a minimum of one **(a, d)**, two **(b, e)** and three **(c, f)** events in neighborhoods of $150\,\mathrm{km} \times 150\,\mathrm{km}$.

References

Allen, S., Evans, G. R., Buchanan, P., and Kwasniok, F.: Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts, Q. J. Roy. Meteor. Soc., 147, 1403–1418, https://doi.org/10.1002/qj.3983, 2021.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

Büeler, D., Ferranti, L., Magnusson, L., Quinting, J. F., and Grams, C. M.: Year-round sub-seasonal forecast skill for Atlantic–European weather regimes, Q. J. Roy. Meteor. Soc., 147, 4283–4309, https://doi.org/10.1002/qj.4178, 2021.

Casanueva, A., Rodríguez-Puebla, C., Frías, M. D., and González-Reviriego, N.: Variability of extreme precipitation over Europe and its relationships with teleconnection patterns, Hydrol.

Nat. Hazards Earth Syst. Sci., 23, 2857–2871, 2023

https://doi.org/10.5194/nhess-23-2857-2023

Earth Syst. Sci., 18, 709–725, https://doi.org/10.5194/hess-18-709-2014, 2014.

CDO: Regridding with CDO, Running the Remapping (Conservative Method), https://www.climate-cryosphere.org/wiki/index.php?title=Regridding_with_CDO#Running_the_Remapping_.28Conservative_Method.29 (last access: 19 May 2022), 2018.

Crochemore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 20, 3601–3618, https://doi.org/10.5194/hess-20-3601-2016, 2016.

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

de Andrade, F. M., Young, M. P., Macleod, D., Hirons, L. C., Woolnough, S. J., and Black, E.: Subseasonal precipitation prediction for africa: Forecast evaluation and sources of predictability, Weather Forecast., 36, 265–284, https://doi.org/10.1175/WAF-D-20-0054.1, 2021.

DeFlorio, M. J., Waliser, D. E., Guan, B., Ralph, F. M., and Vitart, F.: Global evaluation of atmospheric river subseasonal prediction skill, Clim. Dynam., 52, 3039–3060, https://doi.org/10.1007/s00382-018-4309-x, 2019.

Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E. P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S. W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 2. Predictability Arising From Stratosphere-Troposphere Coupling, J. Geophys. Res.-Atmos., 125, e2019JD030923, https://doi.org/10.1029/2019JD030923, 2019.

Domeisen, D. I., White, C. J., Afargan-Gerstman, H., Muñoz, Á. G., Janiga, M. A., Vitart, F., Wulf, C. O., Antoine, S., Ardilouze, C., Batté, L., Bloomfield, H. C., Brayshaw, D. J., Camargo, S. J., Charlton-Pérez, A., Collins, D., Cowan, T., del Mar Chaves, M., Ferranti, L., Gómez, R., González, P. L., González Romero, C., Infanti, J. M., Karozis, S., Kim, H., Kolstad, E. W., LaJoie, E., Lledó, L., Magnusson, L., Malguzzi, P., Manrique-Suñén, A., Mastrangelo, D., Materia, S., Medina, H., Palma, L., Pineda, L. E., Sfetsos, A., Son, S.-W., Soret, A., Strazzo, S., and Tian, D.: Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe, B. Am. Meteorol. Soc., 103, E1473–E1501, https://doi.org/10.1175/bams-d-20-0221.1, 2022.

Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Nurmi, P., Gill, P., Göber, M., Joslyn, S., Brown, B., Fowler, T., and Watkins, A.: Progress and challenges in forecast verification, Meteorol. Appl., 20, 130–139, https://doi.org/10.1002/met.1392, 2013.

ECMWF: Summary of cycle 47r2, https://www.ecmwf.int/en/forecasts/about-our-forecasts/evolution-ifs/cycles/summary-cycle-47r2 (last access: 24 September 2022), 2021.

ECMWF: Changes in ECMWF model, https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model (last access: 19 May 2022), 2022a.

ECMWF: Re-forecast for medium and extended forecast range, https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range/re-forecast-medium-and-extended-forecast-range (last access: 27 April 2022), 2022b.

ECMWF: S2S, ECMWF, Reforecasts, Daily averaged, https://apps.ecmwf.int/datasets/data/s2s-reforecasts-daily-averaged-ecmf/levtype=sfc/type=cf/ (last access: 27 April 2022), 2022c.

ECMWF: S2S, ECMWF, Reforcasts, Instantaneous and Accumulated, ECMWF [data set], https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-ecmf/levtype=sfc/type=cf/ (last access: 21 August 2023), 2023.

Endris, H. S., Hirons, L., Segele, Z. T., Gudoshava, M., Woolnough, S., and Artan, G. A.: Evaluation of the skill of monthly precipitation forecasts from global prediction systems over the greater horn of africa, Weather Forecast., 36, 1275–1298, https://doi.org/10.1175/WAF-D-20-0177.1, 2021.

Gneiting, T. and Ranjan, R.: Comparing Density Forecasts Using Threshold-and Quantile-Weighted Scoring Rules, J. Bus. Econ. Stat., 29, 411–422, http://www.jstor.org/stable/23243806 (last access: last access: 18 August 2023), 2011.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. R. Stat. Soc. B, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., and Wernli, H.: Balancing Europe's wind-power output through spatial deployment informed by weather regimes, Nat. Clim. Change, 7, 557–562, https://doi.org/10.1038/NCLIMATE3338, 2017.

Haylock, M. R. and Goodess, C. M.: Interannual variability of European extreme winter rainfall and links with mean large-scale circulation, Int. J. Climatol., 24, 759–776, https://doi.org/10.1002/joc.1033, 2004.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Sabater, J. M., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Dee, D.: Global reanalysis: goodbye ERA-Interim, hello ERA5, ECMWF Newsletter, 146, 17–24, https://doi.org/10.21957/vf291hehd7, 2019.

Hofstra, N., Haylock, M., New, M., and Jones, P. D.: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature, J. Geophys. Res.-Atmos., 114, D21101, https://doi.org/10.1029/2009JD011799, 2009.

Huang, Z., Zhao, T., Xu, W., Cai, H., Wang, J., Zhang, Y., Liu, Z., Tian, Y., Yan, D., and Chen, X.: A seven-parameter Bernoulli-Gamma-Gaussian model to calibrate subseasonal to seasonal precipitation forecasts, J. Hydrol., 610, 127896, https://doi.org/10.1016/j.jhydrol.2022.127896, 2022.

Hudson, D., Alves, O., Hendon, H. H., and Marshall, A. G.: Bridging the gap between weather and seasonal forecasting: intraseasonal forecasting for Australia, Q. J. Roy. Meteor. Soc., 137, 673–689, https://doi.org/10.1002/qj.769, 2011.

Huijnen, V., Flemming, J., Kaiser, J. W., Inness, A., Leitão, J., Heil, A., Eskes, H. J., Schultz, M. G., Benedetti, A., Hadji-Lazaro, J., Dufour, G., and Eremenko, M.: Hindcast experiments of tropospheric composition during the summer 2010 fires over western Russia, Atmos. Chem. Phys., 12, 4341–4364, https://doi.org/10.5194/acp-12-4341-2012, 2012.

Jones, P. W.: First- and second-order conservative remapping schemes for grids in spherical coordinates, Mon. Weather Rev., 127, 2204–2210, https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2, 1999.

Kenyon, J. and Hegerl, G. C.: Influence of modes of climate variability on global precipitation extremes, J. Climate, 23, 6248–6262, https://doi.org/10.1175/2010JCLI3617.1, 2010.

Kolachian, R. and Saghafian, B.: Deterministic and probabilistic evaluation of raw and post processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes, Theor. Appl. Climatol., 137, 1479–1493, https://doi.org/10.1007/s00704-018-2680-5, 2019.

Legrand, J., Naveau, P., and Oesting, M.: Evaluation of binary classifiers for asymptotically dependent and independent extremes, arXiv [preprint], https://doi.org/10.48550/arXiv.2112.13738, 2022.

Li, W. E., Chen, J. I., Li, L. U., Chen, H. U., Liu, B., Xu, C. Y., and Li, X.: Evaluation and bias correction of S2S precipitation for hydrological extremes, J. Hydrometeorol., 20, 1887–1906, https://doi.org/10.1175/JHM-D-19-0042.1, 2019.

Manrique-Suñen, A., Gonzalez-Reviriego, N., Torralba, V., Cortesi, N., and Doblas-Reyes, F. J.: Choices in the verification of s2s forecasts and their implications for climate services, Mon. Weather Rev., 148, 3995–4008, https://doi.org/10.1175/MWR-D-20-0067.1, 2020.

Mariotti, A., Ruti, P. M., and Rixen, M.: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort, npj Climate and Atmospheric Science, 1, 2–5, https://doi.org/10.1038/s41612-018-0014-z, 2018.

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., Dirmeyer, P. A., Ferranti, L., Johnson, N. C., Jones, J., Kirtman, B. P., Lang, A. L., Molod, A., Newman, M., Robertson, A. W., Schubert, S., Waliser, D. E., and Albers, J.: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond, B. Am. Meteorol. Soc., 101, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1, 2020.

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., Domeisen, D. I., Ferranti, L., Ilynia, T., Kumar, A., Müller, W. A., Rixen, M., Robertson, A. W., Smith, D. M., Takaya, Y., Tuma, M., Vitart, F., White, C. J., Alvarez, M. S., Ardilouze, C., Attard, H., Baggett, C., Balmaseda, M. A., Beraki, A. F., Bhattacharjee, P. S., Bilbao, R., De Andrade, F. M., DeFlorio, M. J., Díaz, L. B., Ehsan, M. A., Fragkoulidis, G., Grainger, S., Green, B. W., Hell, M. C., Infanti, J. M., Isensee, K., Kataoka, T., Kirtman, B. P., Klingaman, N. P., Lee, J. Y., Mayer, K., McKay, R., Mecking, J. V., Miller, D. E., Neddermann, N., Ng, C. H. J., Ossó, A., Pankatz, K., Peatman, S., Pegion, K., Perlwitz, J., Recalde-Coronel, G. C., Reintges, A., Renkl, C., Solaraju-Murali, B., Spring, A., Stan, C., Sun, Y. Q., Tozer, C. R., Vigaud, N., Woolnough, S., and Yeager, S.: Current and emerging developments in subseasonal to decadal prediction, B. Am. Meteorol. Soc., 101, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1, 2020.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I., Feser, F., Koszalka, I., Kreibich, H., Pantillon, F., Parolai, S., Pinto, J. G., Punge, H. J., Rivalta, E., Schröter, K., Strehlow, K., Weisse, R., and Wurpts, A.: Impact Forecasting to Support Emergency Management of Natural Hazards, Rev. Geophys., 58, 1–52, https://doi.org/10.1029/2020RG000704, 2020.

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., and Liniger, M. A.: Skill of Subseasonal Forecasts in Europe: Effect of Bias Correction and Downscaling Using Surface Observations, J. Geophys. Res.-Atmos., 123, 7999–8016, https://doi.org/10.1029/2017JD027923, 2018.

Olaniyan, E., Adefisan, E. A., Oni, F., Afiesimama, E., Balogun, A. A., and Lawal, K. A.: Evaluation of the ECMWF subseasonal to seasonal precipitation forecasts during the peak of West Africa Monsoon in Nigeria, Front. Environ. Sci., 6, 1–15, https://doi.org/10.3389/fenvs.2018.00004, 2018.

Pantillon, F., Lerch, S., Knippertz, P., and Corsmeier, U.: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble, Q. J. Roy. Meteor. Soc., 144, 1864–1881, https://doi.org/10.1002/qj.3380, 2018.

Pic, R., Dombry, C., Naveau, P., and Taillardat, M.: Mathematical Properties of Continuous Ranked Probability Score Forecasting, arXiv [preprint], https://doi.org/10.48550/arXiv.2205.04360, 2022.

Rivoire, P.: Codes for Assessment of subseasonal-to-seasonal (S2S) ensemble extreme precipitation forecast skill over Europe, Zenodo [code], https://doi.org/10.5281/zenodo.8260765, 2023.

Rivoire, P., Martius, O., and Naveau, P.: A Comparison of Moderate and Extreme ERA-5 Daily Precipitation With Two Observational Data Sets, Earth and Space Science, 8, e2020EA001633, https://doi.org/10.1029/2020EA001633, 2021.

S2S-challenge: Challenge to improve Sub-seasonal to Seasonal Predictions using Artificial Intelligence, https://s2s-ai-challenge.github.io/ (last access: 5 March 2022), 2021.

Schaefer, J. T.: The Critical Success Index as an Indicator of Warning Skill, Weather Forecast., 5, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2, 1990.

Specq, D. and Batté, L.: Improving subseasonal precipitation forecasts through a statistical–dynamical approach : application to the southwest tropical Pacific, Clim. Dynam., 55, 1913–1927, https://doi.org/10.1007/s00382-020-05355-7, 2020.

Stan, C., Zheng, C., Chang, E. K. M., Domeisen, D. I. V., Garfinkel, C. I., Jenney, A. M., Kim, H., Lim, Y.-K., Lin, H., Robertson, A., Schwartz, C., Vitart, F., Wang, J., and Yadav, P.: Advances in the prediction of MJO-Teleconnections in the S2S forecast systems, B. Am. Meteorol. Soc., 103, E1426–E1447, https://doi.org/10.1175/BAMS-D-21-0130.1, 2022.

Tabari, H. and Willems, P.: Lagged influence of Atlantic and Pacific climate patterns on European extreme precipitation, Sci. Rep., 8, 1–10, https://doi.org/10.1038/s41598-018-24069-9, 2018.

Taillardat, M., Fougères, A.-L., Naveau, P., and de Fondeville, R.: Extreme events evaluation using CRPS distributions, arXiv [preprint],https://doi.org/10.48550/arXiv.1905.04022, 2022.

Tian, D., Wood, E. F., and Yuan, X.: CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States, Hydrol. Earth Syst. Sci., 21, 1477–1490, https://doi.org/10.5194/hess-21-1477-2017, 2017.

Tuel, A., Schaefli, B., Zscheischler, J., and Martius, O.: On the links between sub-seasonal clustering of extreme precipitation and high discharge in Switzerland and Europe, Hydrol. Earth Syst. Sci., 26, 2649–2669, https://doi.org/10.5194/hess-26-2649-2022, 2022.

Vitart, F.: List of parameters of S2S reforecast data from ECMWF, https://confluence.ecmwf.int/display/S2S/Parameters (last access: 11 March 2022), 2020.

Vitart, F., Robertson, A. W., and Anderson, D. L.: Subseasonal to Seasonal Prediction Project: Bridging the Gap between Weather

and Climate, WMO Bull., 61, 23–28, https://library.wmo.int/index.php?lvl=notice_display&id=13609 (last access: 18 August 2023), 2012.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., 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., Pegion, K. V., Holbrook, N. J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T. A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., and Zebiak, S. E.: Potential applications of subseasonal-to-seasonal (S2S) predictions, Meteorol. Appl., 24, 315–325, https://doi.org/10.1002/met.1654, 2017.

White, C. J., Domeisen, D. I. V., Acharya, N., Adefisan, E. A., Anderson, M. L., Aura, S., Balogun, A. A., Bertram, D., Bluhm, S., Brayshaw, D. J., Browell, J., Büeler, D., Charlton-perez, A., Christel, I., Coelho, C. A. S., Deflorio, M. J., Monache, D., Giuseppe, F. D., García-solórzano, A. M., Gibson, P. B., Goddard, L., Romero, C. G., Graham, R. J., Graham, R. M., Grams, C. M., Halford, A., Huang, W. T. K., Jensen, K., Kilavi, M., Lawal, K. A., Lee, W., Macleod, D., Manrique-suñén, A., Martins, E. S. P. R., Carolyn, J., Merryfield, W. J., Muñoz, Á. G., Olaniyan, E., Otieno, G., Oyedepo, A., Palma, L., Pechlivanidis, I. G., Pons, D., Ralph, F. M., Dirceu Jr., S. R., Remenyi, T. A., Risbey, J. S., Robertson, D. J. C., Andrew, W., Smith, S., Soret, A., Sun, T., Todd, M. C., Tozer, C. R., Vasconcelos Jr., F. C., Vigo, I., Waliser, D. E., Wetterhall, F., and Wilson, G.: Advances in the application and utility of subseasonal-to-seasonal predictions, B. Am. Meteorol. Soc., 103, 1–57, https://doi.org/10.1175/bams-d-20-0224.1, 2021.

WMO: Weather-related disasters increase over past 50 years, causing more damage but fewer deaths, https://public.wmo.int/en/media/press-release/weather-related-disasters-increase-over-past-50-years-causing-more-damage-fewer (last access: 25 April 2022), 2021.

World-Climate-Service: The Difference Between Deterministic and Ensemble Forecasts, https://www.worldclimateservice.com/2021/10/12/difference-between-deterministic-and-ensemble-forecasts/ (last access: 18 August 2023), 2021.

Yan, Y., Liu, B., Zhu, C., Lu, R., Jiang, N., and Ma, S.: Subseasonal forecast barrier of the North Atlantic oscillation in S2S models during the extreme mei – yu rainfall event in 2020, Clim. Dynam., 58, 0123456789, https://doi.org/10.1007/s00382-021-06076-1, 2021.

Zhang, L., Kim, T., Yang, T., Hong, Y., and Zhu, Q.: Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous U.S., J. Hydrol., 603, 127058, https://doi.org/10.1016/j.jhydrol.2021.127058, 2021.