# Assessment of S2S ensemble extreme precipitation ~~forecasts~~ forecast skill over Europe

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

[1]Institute of Geography and Oeschger Centre for Climate Change Research, University of Bern, Switzerland
[2]Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, Switzerland
[3]Mobiliar Lab for Natural Risks, University of Bern, Switzerland
[4]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~~ sub-seasonal 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 ~~introduce a new methodology to assess the~~ 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 $95^{\text{th}}$ 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 analyse daily events locally and spatially aggregated, as well as 7-day 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)~~(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 a better preparedness ~~Merz et al. (2020).~~ (Merz et al., 2020).

Subseasonal-to-seasonal (S2S) prediction refers to forecasting on timescales from about two weeks to a season. S2S prediction has a large range of applications ~~White et al. (2017, 2021)~~(White et al., 2017, 2021), including the humanitarian sector, public health, energy, water management and agriculture. Forecast skill at this time scale is key to better manage natural hazards ~~Merz et al. (2020)~~(Merz et al., 2020). S2S predictions aim to fill the gap between weather forecasts and seasonal outlooks ~~White et al. (2017)~~(White et al., 2017). Providing skillful predictions on subseasonal or monthly timescales is challenging ~~Hudson et al. (2011)~~(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)~~(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)~~(Mariotti et al., 2018; Merryfield et al., 2020; Domeisen et al., 2022). Many research organisations actively contribute to improving S2S forecast skill, for example the *Challenge to improve Sub-seasonal to Seasonal Predictions using Artificial Intelligence* ~~S2S-challenge (2021)~~(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) )~~(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 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 processingSpecq and Batté (2020))~~(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))~~(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); DeFlorio et al. (2019), Africa de Andrade et al. (2021); Olaniyan et al. (2018) and Asia Yan et al. (2021); Li et al. (2019)~~(Zhang et al., 2021; DeFlorio et al., 201 , 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)~~(Monhart et al., 2018; Do . 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)~~(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~~typically~~) to assess the quality of the forecast and to identify and correct model biases ~~(e.g.Huijnen et al. (2012); Manrique-Suñen et al )~~(e.g. Huijnen et al., 2012; Manrique-Suñen 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))~~(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)~~(de Andrade et al., 2019; Li et al., 2019; Stan et al., 20 .

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)~~(CRPS, Hersbach, 2000; Gneiti , the Brier score ~~Brier (1950)~~(Brier, 1950), and the mean square skill score ~~Specq and Batté (2020)~~(Specq and Batté, 2020).

However, these metrics capture the mean behavior of a variable: most are not directly suited to verify extreme events. The CRPS can be adapted to focus on extremes, using the threshold-weighted CRPS ~~Gneiting and Ranjan (2011); Allen et al. (2021)~~ (Gneiting and Ranjan, 2011; Allen et al., 2021) or using extreme value theory ~~Taillardat et al. (2022)~~(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 forecast to identify above-normal precipitation events ~~Domeisen et al. (2022); Monhart et al. (2018)~~(Domeisen et al., 2022; Mon . In this study, we transform precipitation extremes into binary “threshold exceedance events~~"~~”, where the threshold is the daily precipitation $95^{\mathrm{th}}$ 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 events dataset is composed of 95% zeros and 5% ones. A large part of the forecast and observation datasets are matching because of the large presence of “0s” (daily precipitation lower than the $95^{\mathrm{th}}$ percentile) in both datasets. To address this issue, we also use a binary loss index focusing on extremes (“1s”). We assess the extreme events by proposing and using a simple extension of the binary loss score as introduced by ~~Legrand et al. ?~~ 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 section 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)~~ (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)~~(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)~~ (Brier, 1950) and a binary loss index ~~(adapted from Legrand et al. ?).~~ (adapted from Legrand et al., 202 . We present the results of the analysis in Section 3. We discuss these results, draw conclusions and give an outlook in Section 4.

## 2 Data and Methods

### 2.1 Data

We use ECMWF’s S2S precipitation hindcast data ~~(cycle 47r2, ECMWF (2021); Vitart (2020); ECMWF (2022a))~~ (cycle 47r2, , ECMWF, from 2001 to 2020. It is composed of 11 ensemble members, initialized twice a week and run for 46 days. We focus on Europe, in the spatial box [30°N ; 72°N] × [-15 °E; 49.5 °E]. The hindcast period covers 20 years with 2080 forecast initializations between 2001-01-04 and 2020-12-30 (twice a week, on Monday and Thursday). The data were downloaded at the model spectral resolution O320 ~~ECMWF (2022b, c)~~ (ECMWF, 2022b, c) and regridded for the analysis to a 0.5°×0.5° regular grid using a first-order conservative remapping ~~Jones (1999); CDO (2018)~~(Jones, 1999; CDO, 2018).

ERA5 precipitation ~~Hersbach et al. (2019)~~ (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 to avoid the uncertainties due to the inherent spatial sparsity of weather station networks ~~Hofstra et al. (2009); Rivoire et al. (2021)~~(Hofstra et al., 2009; Rivoire et al., 2021)

. Daily precipitation are extracted over the same time period, from 2001-01-04 to 2020-12-30 plus 46 lead time days i.e. 2021-02-14, with a spatial resolution of 0.5°×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 $95^{\mathrm{th}}$ 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 $95^{\mathrm{th}}$ percentile ($Q_{95}$) in ERA5 in Europe, for the period from 2001-01-04 to 2021-02-14. 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 5mm per day are shown. For a lead time of one day, the forecast generally underestimates $Q_{95}$. For lead times between 2 and 46 days, 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, Norway in spring, autumn and winter, see Figure 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 figure 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 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)~~ (Brier, 1950):

$$B = \frac{1}{n_D}\sum_{i=1}^{n_D}(f_i - Y_i)^2,$$

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, i.e. half the number of initializations per year); $Y_i$ the binary observation of extreme for day $i$ ($Y_i$=1 if the daily precipitation exceeds the $95^{\mathrm{th}}$ and $Y_i = 0$ otherwise); $f_i$ 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 $i$.

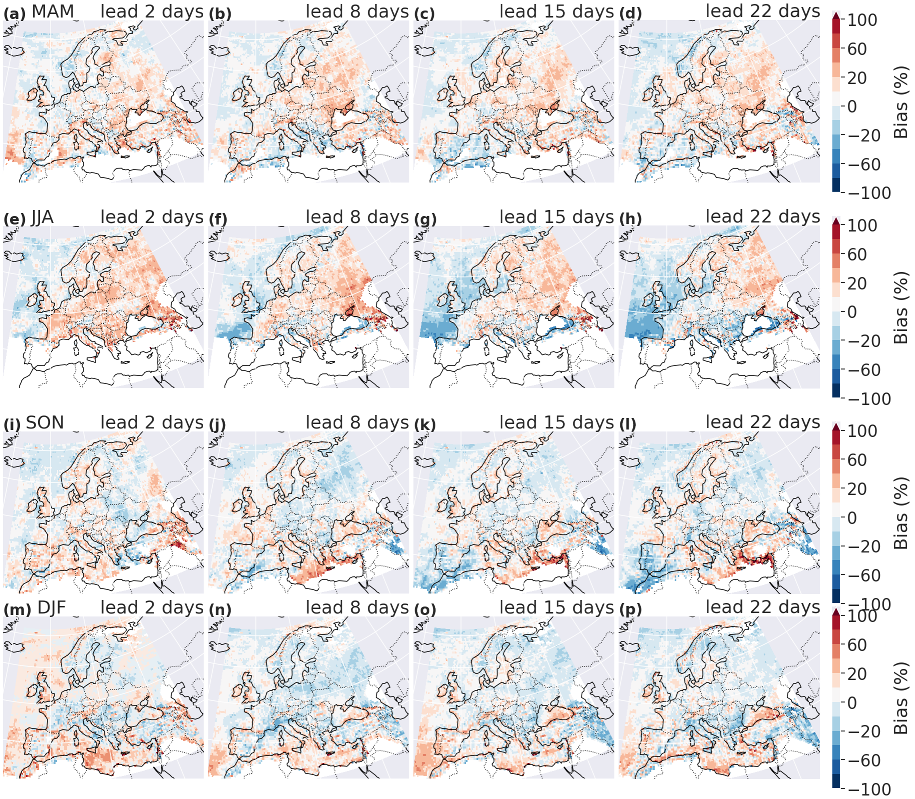

**Figure 1.** Forecast bias in the $95^{\mathrm{th}}$ percentile ($Q_{95}$) compared to ERA5, for spring (MAM, (a)-(d)), summer (JJA, (e)-(h)), autumn (SON, (i)-(l)) and winter (DJF, (m)-(p)) at a 2-day (first column), 8-day (second column), 15-day (third column) and 22-day (last column) lead time. Grid points with $Q_{95}$ ≤5mm per day are displayed in white.

$B$ is negatively oriented (the lower the better). The climatological Brier score $B_{\text{clim}}$ is used as a reference value for the skill calculation:

$$B_{\text{clim}} = \frac{1}{n_D}\sum_{i=1}^{n_D}(p - Y_i)^2,$$

where $p$ is the climatological extreme 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. 120

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 ($BSS$):

$$BSS = 1 - \frac{B_{\text{hind}}}{B_{\text{clim}}}.$$

$BSS$ varies between $]-\infty; 1]$ and is positively oriented (the closer to one, the better). For a given lead time day, a forecast has skill if $BSS > 0$. 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. ? introduced~~ Legrand et al. (2022) studied in detail a risk function defined as the ratio between the empirical probability of having an extreme event in either the observation dataset or the forecast dataset, and the empirical probability of having an extreme event in the observations or the forecast (including having an event in both datasets). ~~This risk function~~ In our context, the risk function can be written:

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

where $Y^{(u)}$ is the binary observation, $Y^{(u)} = 0$ (resp. $Y^{(u)} = 1$) if the observed daily precipitation is lower (greater) than a certain threshold $u$; and $X^{(u)}$ is the binary forecast $X^{(u)} = 0$ (resp. $X^{(u)} = 1$) 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 experience 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 be penalized. In other words, even for 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 under-represented 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 amount of correctly predicted 0s.

$1 - R^{(u)}(X)$ 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 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 initialisation date and a lead time, $F^{med} = 1$ if at least 6 ensemble members predict extreme precipitation and $F^{med} = 0$ otherwise. We take here the median forecast across members, but in practice $F^{med} = 1$ could be set to 1 only if fewer or more than 6 members forecast extreme precipitation. The choice depends on the risk aversion of the users and is discussed in Section 4.

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

$$BLI_m = \frac{N_1^{med}}{N_2^{med}},$$

where $N_1^{med}$ is the number of days when the observation and the ensemble median disagree, i.e. $N_1^{med} = \#\{j \mid F_j^{med} \neq Y_j\}$, and $N_2^{med}$ is the number of days when an extreme event occurs in either or both the observation and the ensemble median, i.e.

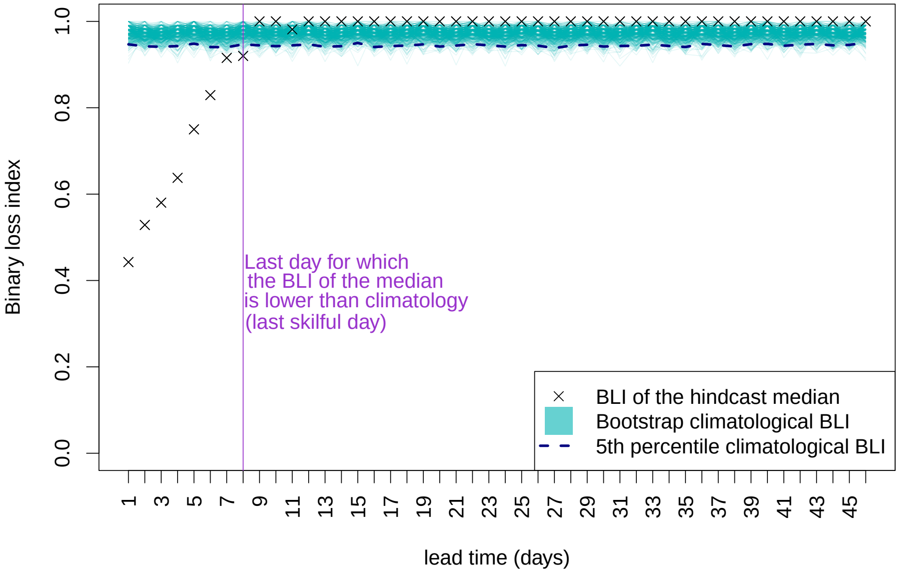

**Figure 2.** Definition of the last skillful day for BLI: example for one grid point and one season.

$N_2^{med} = \#\{j \mid \big(F_j^{med} = 1 \text{ or } Y_j = 1\big)\}$. 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.

~~This metric focuses on how well the "1" values (extreme event days) match between the two datasets. It doesn't take into account steps when neither the forecast nor the observation experience an extreme event. 1-$BLI$ can be understood as the critical success index Schaefer (1990); ?.~~ To measure the lead time dependence of the skill, $BLI$ is computed for each lead time day.

~~$BLI$ varies between [0;1] and is negatively oriented (the closer to zero, the better).~~ 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. The $BLI$ is computed with this random time series as 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 5% percentile of the confidence interval on $BLI_{clim}$. Like for the Brier score, we compute the "last skillful day" for the BLI, with the same definition (largest lead time day with skill, see figure 2 for an example).

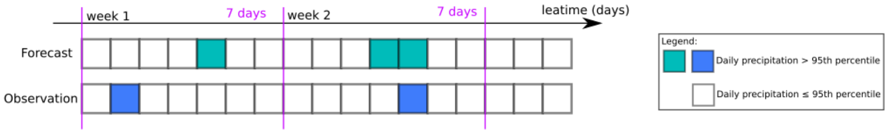
week 1
7 days
week 2
7 days
leatime (days)
Forecast
Observation
Legend:
Daily precipitation > 95th percentile
Daily precipitation ≤ 95th percentile

**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-day window is greater or equal to 1: $E_1^t = 1$ for the two datasets. For both datasets, the number of events in the 7-day 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 7day window is greater or equal to 1: $E_1^t = 1$ for the two dataset. For the observation, the number of events in the 7-day window is lower than 2 ($E_2^t = 0$) and this number is greater or equal to 2 for the forecast ($E_2^t = 1$). For the configuration with $n$ events, $n \geq 3$, $E_n^t = 0$ for both datasets.

### 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))~~(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 analyse the individual impact of each aggregation. Both the spatial and 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-day window. We then translate it into a binary series $E_n^t$: given a minimum number of events $n$ in the window ($n = \{1,...,7\}$) :

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

Figure 3 provides an example for the definition of $E_n^t$. To the various binary series (one for each $n$), 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-day time windows in the observation. The 6 following days are not randomly selected, they are the 6 days 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 the count of events $N^s$ in the spatial neighborhood exceeds or

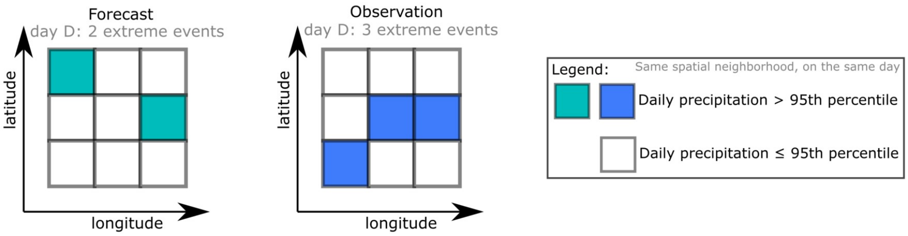
Forecast
day D: 2 extreme events
latitude
longitude
Observation
day D: 3 extreme events
latitude
longitude
Legend:
Same spatial neighborhood, on the same day
Daily precipitation > 95th percentile
Daily precipitation ≤ 95th percentile

**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 or equal to 1 ($E_1^s = 1$) and greater or equal to 2 ($E_2^s = 1$). For 3 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 for $n$ ≥4.

not some threshold $n$ (see figure 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, the spatial neighborhood is the observed neighborhood for that day. We define the neighborhoods as square boxes of about 150km*150km, i.e. boxes with a latitudinal extent of 1.5°N (3 gridboxes) and with a longitudinal grid extent that depends on the latitude: from 1.5°E at 30°N (3 gridboxes) to 4.5°E at 70°N (9 gridboxes), see figure D1 in 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. The BLI indicates more skill during the extended winter (skill for up to 11 days, and many regions with a last skillful day greater than 7 days) than during the extended summer (last skillful day below 6 days for most grid points), see figure 5. 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 those of the Brier skill score~~confirms these patterns, see .~~, however the spatial patterns are similar for the two metrics (figure B1 in the appendix). That is, the last skillful day for the Brier skill score is overall smaller than the last skillful day for the BLI, but both the Brier score and the BLI show the same regions with high and low skill of the forecast for precipitation extremes, in summer and winter. ~~However, the BLI skill estimates are more conservative than those of the Brier skill score .~~

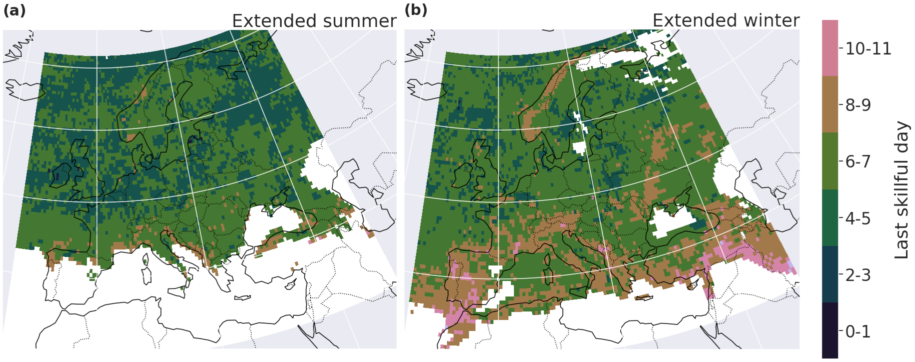

**Figure 5.** Last skillful day for the BLI for a local and daily comparison, in extended summer (a) and extended winter (b).

### 3.2 Temporal aggregation

7-day extreme precipitation event counts are also better predicted during the extended winter than during the extended summer (figure 6). For the category “one event or more occurred during the 7 days", the forecasts at most grid points still have skill for lead times into the second week, i.e. days 8 -14, in extended winter. The 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 figure C1 in appendix).

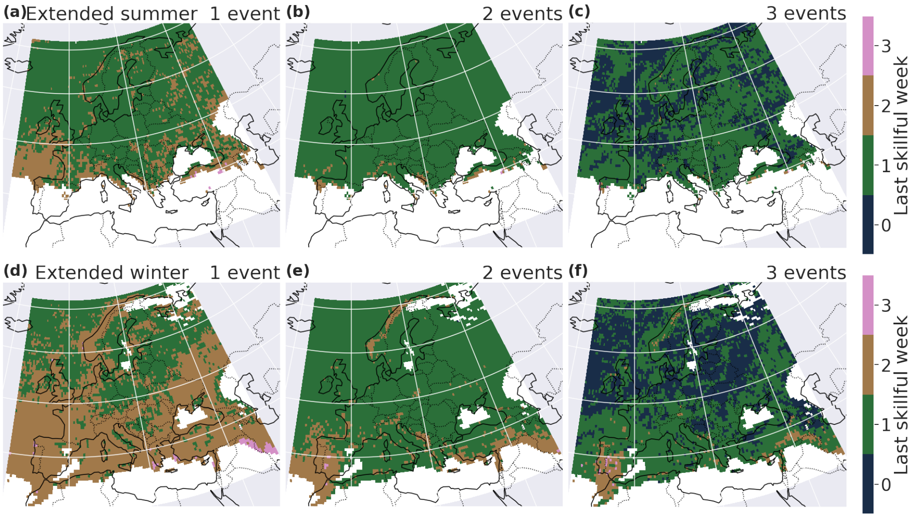

**Figure 6.** Last week of skill for the BLI in extended summer (a-c) and extended winter (d-e) for a minimum of 1 (a,d), 2 (b,e) and 3 (c,f) events in a 7-day window. A last ~~skilful~~ skillful week equal to 0 means that, for the count of extremes during the first week lead time, the $BLI$ of the forecast is not significantly lower than $BLI_{clim}$.

### 3.3 Spatial aggregation

Extended winter forecasts for spatially aggregated extremes are also more ~~skillfull~~ skillful than summer ones (see figure 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 days 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 days on the Atlantic coast of France, Italy, Western Europe and the coasts of the Iberian Peninsula . The spatial skill pattern remains similar with increasing number of events per neighborhood but the skill decreases.

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 the BLI. The spatial pattern of the skill also remains similar with 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 $95^{\text{th}}$ percentile. We conduct a verification against ERA5 precipitation with the binary loss index ($BLI$) and the Brier score. We ~~introduce the $BLI$ as an extension to ensemble forecasts of the~~ extend the binary loss score ~~of Legrand et al. ?~~studied by Legrand et al. (2022), which was designed for deterministic forecasts only, to ensemble

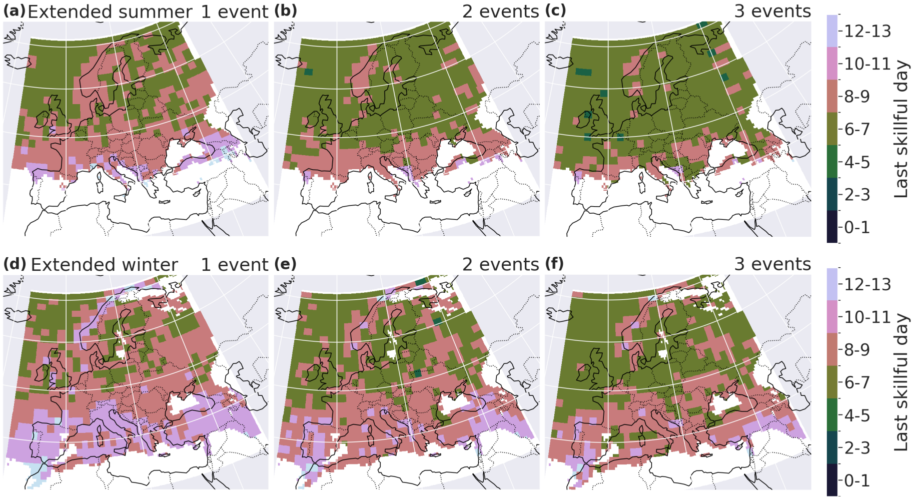

**Figure 7.** Last day of skill for the BLI in extended summer (a-c) and extended winter (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in ~~neighborhood~~ neighborhoods of 150×150km.

forecasts. We define the $BLI$ 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 the $BLI$ (risk averse setting) or when one chooses a ~~very~~ high percentile of the ensemble members (risk loving). The $BLI$ has the advantage of focusing exclusively on extreme event occurrence (hit, false alarm or miss) and ~~are~~ is not biased by the high counts of extreme event non-occurrence. The ~~BLI~~ $BLI$ is qualitatively compared with the Brier score; the skill scores of 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, the $BLI$ 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 the ~~$BLF$~~$BLI$.

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 large scale ones ~~Haylock and Goodess (2004); Kenyon and Hegerl (2010)~~(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)~~ (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 out 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 investigate 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 a matching event 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 analysing multi-day heatwaves.

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 NAO. The forecast skill does not exhibit a strong dependence on the NAO phase, although the data was 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 East 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 to be calibrated ~~Gneiting et al. (2007); Specq and Batté (2020); Crochemore et al. (2016); Monhart et al. (2018); Huang et al. (202~~ (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))~~ (see e.g. Pantillon et al., 2018; Allen et al., 2021) or using extreme value theory ~~Taillardat et al. (2022)~~(Taillardat et al., 2022). Post-processing the hindcast data and analysing 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)~~(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 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 the 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 ~~20~~25% of the extremes are caught by the forecast ~~. 80~~(by simple transformation of $BLI = \frac{FN+FP}{TP+FN+FP} = 0.8$, where FN are the false negatives, FP are the false positives and TP are the true positives). 75% of the time, either the forecast ~~predicted erroneously~~ 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. 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 ~~quantile~~ 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. $F^{med} = 1$ if at least 1 ensemble member predicts extreme precipitation and $F^{med} = 0$ otherwise). This is due to the number of TP 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 TP, still compensating a bit for the FP for large lead times. For the choice of the member, we recommend considering a good balance between FN and FP. 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 larger 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 the 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 side of extremes, evaluating the skill of forecasts to predict droughts would also be 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. adjusting the definition temporal aggregation introduced in this study.

*Code and data availability.* The codes used for the data analysis are available on github (https://github.com/PauRiv/S2S_verif_precip). The ECMWF’s S2S hindcast data are available on the ECMWF platform (https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-ecmf/levtype=sfc/type=cf/, cycle 47r2).

## Appendix A: 95th percentile

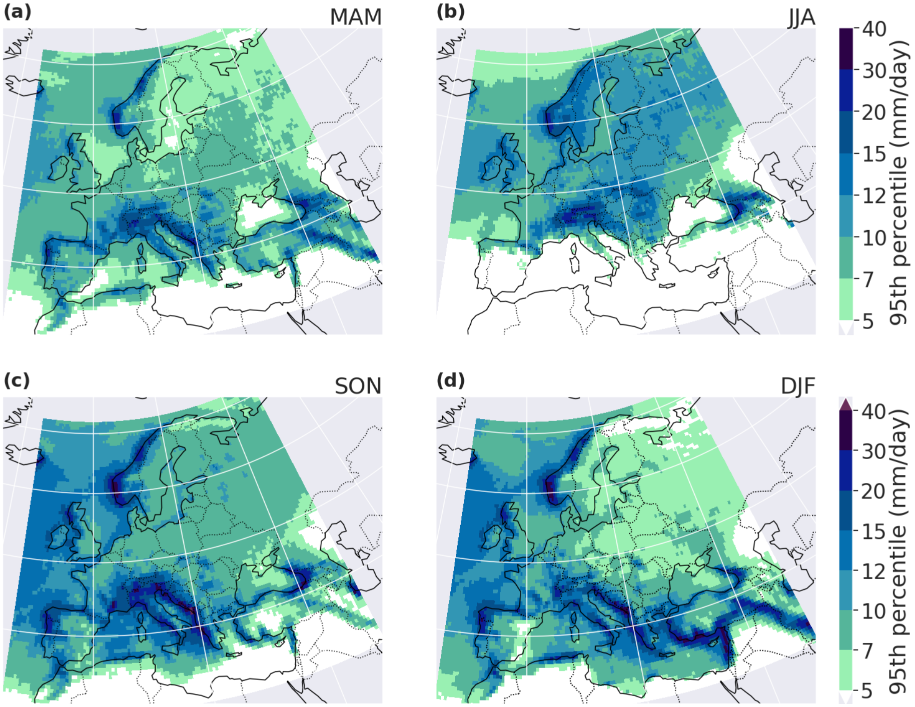

**Figure A1.** $95^{\text{th}}$ percentile of daily precipitation in ERA5, 2001-2021.

## Appendix B: Local and daily comparison of extremes

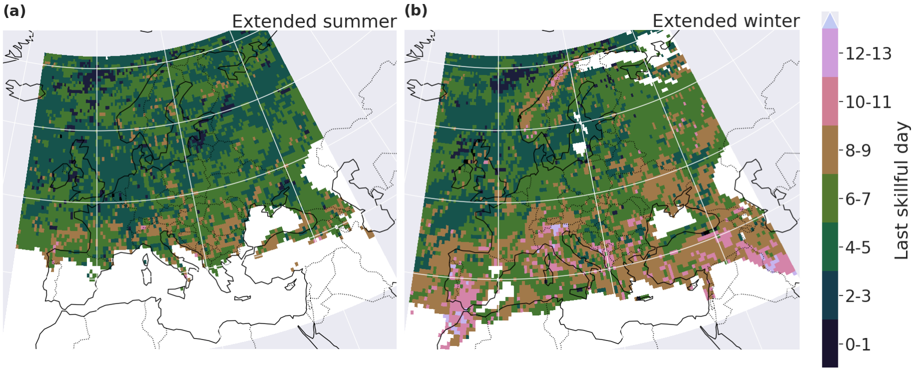

**Figure B1.** Last skillful day for the Brier skill score for local and daily comparison, in extended summer (a) and extended winter (b).

## Appendix C: Temporally accumulated extremes

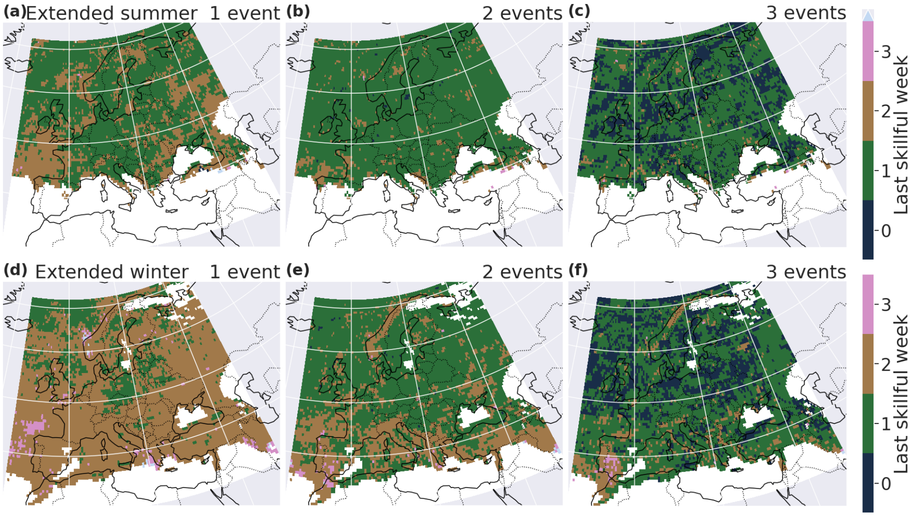

**Figure C1.** Last week of skill for the Brier skill score in extended summer (a-c) and extended winter (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in a 7 days window.

## Appendix D: Spatially accumulated extremes

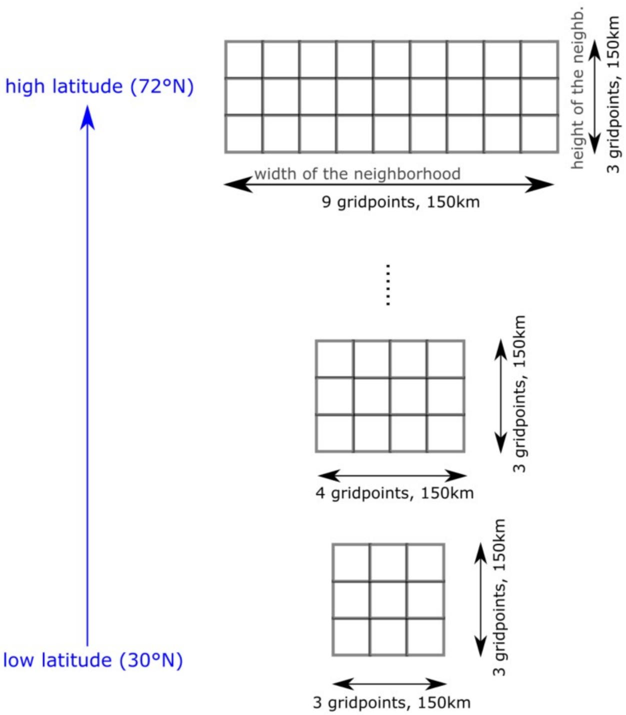

**Figure D1.** Illustration of the width of the spatial neighborhood, in terms of grid points, depending on the latitude for a constant width in kilometers (and for a constant area).

*Author contributions.* PR, OM and PN designed the research. PR executed the analysis, created the figures and wrote the draft. OM and PN contributed to the formal analysis. OM, PN and AT contributed to the data analysis and revised the manuscript. A.T. collected the data. All authors contributed to the final version of the manuscript.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* P.R. and O.M. acknowledge funding from the Swiss National Science Foundation (grant number 178751). We thank Jonas Bhend for helpful discussions. Part of P.N. work was supported by three French national programs (80 PRIME CNRS-INSU, ANR T-REX under reference ANR-20-CE40-0025-01, ANR Melody (ANR-19-CE46-0011), and the European H2020 XAIDA (Grant agreement ID: 101003469). The support of DAMOCLES-COST-ACTION on compound events is also acknowledged.

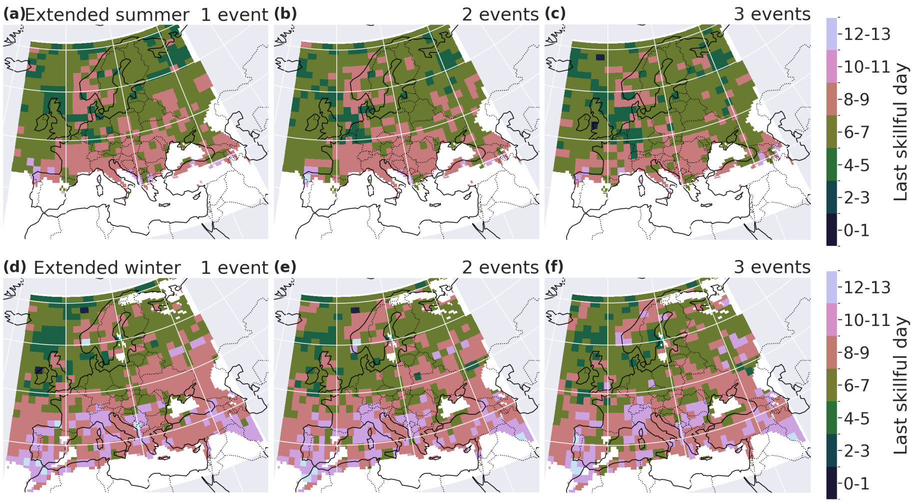

**Figure D2.** Last day of skill for the Brier skill score in extended summer (a-c) and extended winter (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in ~~neighborhood~~ neighborhoods of 150×150km.

## 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, Quarterly Journal of the Royal Meteorological Society, 147, 1403–1418, https://doi.org/10.1002/qj.3983, 2021.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Monthly Weather Review, 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, Quarterly Journal of the Royal Meteorological Society, 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, Hydrology and Earth System Sciences, 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, accessed: 2022-05-19, 2018.

Crochemore, L., Ramos, M. H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrology and Earth System Sciences, 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, Climate Dynamics, 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 and Forecasting, 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, Climate Dynamics, 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, Journal of Geophysical Research: Atmospheres, 125, 1–20, 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, Bulletin of the American Meteorological Society, pp. 1–76, 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, Meteorological Applications, 20, 130–139, https://doi.org/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, accessed: 2022-09-24, 2021.

ECMWF: Changes in ECMWF model, https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model, accessed: 2022-05-19, 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, accessed: 2022-04-27, 2022b.

ECMWF: S2S, ECMWF, Reforecasts, Daily averaged, https://apps.ecmwf.int/datasets/data/s2s-reforecasts-daily-averaged-ecmf/levtype=sfc/type=cf/, accessed: 2022-04-27, 2022c.

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 and Forecasting, 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, Journal of Business & Economic Statistics, 29, 411–422, http://www.jstor.org/stable/23243806, 2011.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 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, Nature Climate 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, International Journal of Climatology, 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 and Forecasting, 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, Journal of Geophysical Research Atmospheres, 114, 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, Journal of Hydrology, p. 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, Quarterly Journal of the Royal Meteorological Society, 137, 673–689, https://doi.org/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, Atmospheric Chemistry and Physics, 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, Monthly Weather Review, 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, Journal of 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, Theoretical and Applied Climatology, 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, https://doi.org/10.48550/ARXIV.2112.13738v2, 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, Journal of Hydrometeorology, 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, Monthly Weather Review, 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, Bulletin of the American Meteorological Society, 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, Bulletin of the American Meteorological Society, 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, Reviews of Geophysics, 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, Journal of Geophysical Research: Atmospheres, 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 sub-seasonal to seasonal precipitation forecasts during the peak of West Africa Monsoon in Nigeria, Frontiers in Environmental Science, 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, Quarterly Journal of the Royal Meteorological Society, 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, http://arxiv.org/abs/2205.04360, 2022.

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, 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/, accessed: 2022-03-05, 2021.

Schaefer, J. T.: The Critical Success Index as an Indicator of Warning Skill, Weather and Forecasting, pp. 570–575, 1990.

Specq, D. and Batté, L.: Improving subseasonal precipitation forecasts through a statistical–dynamical approach : application to the southwest tropical Pacific, Climate Dynamics, 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, Bulletin of the American Meteorological Society, -1, 2022.

Tabari, H. and Willems, P.: Lagged influence of Atlantic and Pacific climate patterns on European extreme precipitation, Scientific Reports, 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, http://arxiv.org/abs/1905.04022, 2022.

Tian, D., Wood, E., and Yuan, X.: CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States, Hydrology and Earth System Sciences, 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, Hydrology and Earth System Sciences, 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, accessed: 2022-03-11, 2020.

Vitart, F., Robertson, A. W., and Anderson, D. L.: Subseasonal to Seasonal Prediction Project: Bridging the Gap between Weather and Climate, WMO Bulletin, 61, 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, Meteorological Applications, 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, S., Jr, 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., Jr, F. C. V., Vigo, I., Waliser, D. E., Wetterhall, F., and Wilson, G.: Advances in the application and utility of subseasonal-to-seasonal predictions, Bulletin of the American Meteorological Society, pp. 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, accessed: 2022-04-25, 2021.

World-Climate-Service: The Difference Between Deterministic and Ensemble Forecasts, https://www.worldclimateservice.com/2021/10/12/difference-between-deterministic-and-ensemble-forecasts/, 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, Climate Dynamics, 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., Journal of Hydrology, 603, 127 058, https://doi.org/10.1016/j.jhydrol.2021.127058, 2021.