Page 3 -> merge the mini-paragraph on definitions with the next or possibly the first three paragraphs. The link of these definitions to the work are not clear to me. Why do other definitions than meteo drought matter for this paper? -> Tie in better

Done as suggested. We consider it important to indicate that we focus on meteorological droughts. As the definition of a drought is different from one discipline to the other it appears necessary to clarify this in the introductory section.

Both reviewers and myself asked to discuss the issue of only looking at SPI- 1 in the Discussion section. The condensed bit of text added to the introduction is not sufficient to address this major concern. It also does not fully reflect your more detailed response letter. -> Revise this aspect also in the discussion section! There is still also not much embedding into other work. If you argue with the usefulness for operational application, you should review and point to, which operational systems in the world already use a one month forecast and how your study may influence progress in this respect. Quickly browsing, I found E.g. http://droughtmonitor.unl.edu/SupplementalInfo/Forecasts.aspx.
And I am sure a few of these global applications also do this, no? So how exactly will your study help them improve?

We have added a discussion section.
Most drought studies use SPI with three- to six- months or even longer accumulation periods for drought monitoring and characterization. To forecast droughts over such long-term periods a very accurate and reliable atmospheric model is required. Since it is well known that the current reliability of precipitation forecasts decreases drastically after the first month, the benefit of using a lead time of two months or more is, however, not obvious (Dutra et al. 2013).

This paper, therefore, looks as a first step at the possibilities to provide a reliable one month forecast over the European continent. This information, in combination with monitoring data such as satellite or in-situ measurements that provide an accurate characterization of ongoing drought conditions (e.g. during the last two months), can provide the best estimate of near future conditions. However, also such a combination of monitoring and forecasting data will not allow looking more than one month ahead and an amalgamation of both information types would bias the testing of the forecast skill, which is the intention of this paper. Several meteorological services or agencies, such as the Bureau of Meteorology in Australia or the United States National Drought Mitigation Center, provide relevant monitoring data as well as a one month outlook. For the case of Europe, the European Drought Observatory (EDO) at the Joint Research Centre of the European Commission provides relevant monitoring data, but up to now lacks the forecast beyond 7 days.

A one month forecast with a good reliability is considered to be a very valuable product for decision makers as it provides information on the probability of occurrence of a dry spell (in case of ongoing normal conditions) and of the probable persistence or end of a drought (in case of an ongoing precipitation deficit). Before providing such information, it is however necessary to assess the quality of the forecasts, which was the first aim of this study. The second objective was to define the most robust (Boolean) method to activate alert levels for the end users of the forecast information. Both steps are essential in an operational early warning environment.

New equations for skill scores:
->Add references to the equation numbers in the text
Done as suggested

Page 19 new paragraph:
a) first sentence:-> remove “global” or find another descriptor, but the application was neither ‘global’ in geographic sense, nor in the sense of an all-encompassing assessment (but rather narrowly limited to SPI-1).
Removed as suggested

b) should be ‘other methods’
Corrected as suggested

c) Since it is rather unclear what is meant by ‘other methods’ and other ‘atmospheric predictors’ (which, what for – should it be drought indicators?), I suggest to swap the order of the two sentences somehow to make this relation to other approaches clearer.
In fact, since the last part then goes back to SPI-1 (after mentioning other predictors (of drought?), it may be more useful to have the new part last. Please revise this last paragraph for better logic and more clarity.
'other methods' has been replaced by 'statistical weather prediction methods'.
As mentioned, atmospheric predictors are not used as drought indicators but as drought predictors. We consider that the definition and the possible relationships between predictors and droughts are beyond the scope of this paper.

Manuscript prepared for Hydrol. Earth Syst. Sci. Discuss.
with version 2014/07/29 7.12 Copernicus papers of the LaTeX class copernicus.cls.
Date: 25 June 2015

Early warning of drought in Europe using the monthly ensemble system from ECMWF

C. Lavaysse[1], J. Vogt[1], and F. Pappenberger[2]

[1]European Commission, Joint Research Centre, Ispra (Va), Italy
[2]European Centre for Medium-range Weather Forecasts, Reading, UK

Correspondence to: C. Lavaysse (christophe.lavaysse@jrc.ec.europa.eu)

Abstract

Timely forecasts of the onset or possible evolution of droughts are an important contribution to mitigate their manifold negative effects. In this paper we therefore analyse and compare the performance of the first month of the probabilistic extended range forecast and of the seasonal forecast from ECMWF in predicting droughts over the European continent. The Standardized Precipitation Index (SPI-1) is used to quantify the onset or likely evolution of ongoing droughts for the next month.

It can be shown that on average the extended range forecast has greater skill than the seasonal forecast whilst both outperform climatology. No significant spatial or temporal patterns can be observed but the scores are improved when focussing on large-scale droughts. In a second step we then analyse several different methods to convert the probabilistic forecasts of SPI into a Boolean drought warning. It can be demonstrated that methodologies which convert low percentiles of the forecasted SPI cumulative distribution function into warnings are superior in comparison with alternatives such as the mean or the median of the ensemble. The paper demonstrates that up to 40 % of droughts are correctly forecasted one month in advance. Nevertheless, during false alarms or misses, we did not find significant differences in the distribution of the ensemble members that would allow for a quantitative assessment of the uncertainty.

1 Introduction

Droughts can impact many human activities and environmental processes including agriculture, water resources management, inland water transport, energy production and freshwater ecology (Fraser et al., 2013). They often spread over vast geographical regions and last for many months or even years (Lloyd-Hughes and Saunders, 2002). The spatial extent and manifold impacts makes them one of the costliest natural disasters (Below et al., 2007). Given this situation, continuous monitoring as well as forecasting the onset or likely evolution of an ongoing drought over the next few weeks are important to trigger actions

for mitigating negative impacts in the mentioned fields. To do so, decision makers and end users require simple and robust forecast indicators which are capable of informing about the onset, possible duration and end of drought conditions.

Droughts can be classified in several categories (Wilhite and Glantz, 1985): (i) meteorological drought which is defined as a rainfall deficit over a certain space and period of time; (ii) agricultural or soil moisture drought, which describes the propagation of precipitation deficits to soil moisture deficits resulting in plant water stress; and (iii) finally hydrological drought, which is associated with the effects of precipitation deficits on surface and subsurface water supplies. In this study we focus on meteorological droughts using monthly precipitation forecasts from the ECMWF ensemble systems. This timescale is considered as a challenge because located between the medium-range forecasting, which is strongly related to initial conditions, and the seasonal time-scale, mainly driven by oceanic variabilities (Vitart, 2014). The goal is to test the possibilities to provide to decision makers a forecast of the onset or likely evolution of a drought during the next month.

It has been demonstrated that droughts can be forecasted using stochastic or neural networks (Kim and Valdés, 2003; Mishra et al., 2007). While Mishra and Desai (2005) demonstrated that these forecasts can provide "reasonably good agreement for forecasting with 1 to 2 months lead times", they do not quantify the improvement of these methods with respect to using probabilistic forecasts of the precipitation fields. Forecasts of droughts can also be produced using deterministic Numerical Weather Prediction Models. Such forecasts are highly uncertain due to the chaotic nature of the atmosphere, which is particularly strong on a sub-seasonal time scale (Stockdale et al., 1998; Vitart, 2014). Therefore, ensemble prediction systems have been developed that forecast multiple scenarios of future weather. Probabilistic forecasts become particularly important to assess the risks associated with high-impact and rare weather events such as tropical cyclones or droughts (Hamill et al., 2012; Dutra et al., 2013, 2014) as well as for identifying uncertainties in the forecasts (Buizza et al., 2005).

Forecasts on the sub-seasonal time-scale and seasonal forecasts from dynamical models have considerably evolved over recent years and demonstrate potential usefulness

to predict large-scale features and teleconnections (Barnston et al., 2012; Arribas et al., 2011). The latter can be used in statistical downscaling methods using weather types. Eshel et al. (2000), for example, used the North Atlantic sea level pressure precursors to forecast drought over the eastern Mediterranean. However, while their forecasts are statistically significant for several months lead time, this region represents a relatively small part of Europe known to be one of the most sensitive to weather types. In general, the published literature indicates that the skill of the precipitation fields produced by Numerical Weather Predictions over Europe is low (Richardson et al., 2013; Weisheimer and Palmer, 2014; Singleton, 2012) even though there are considerable spatial variations. However, these analyses tend to be performed from the point of view of weather forecasting and do not incorporate specific properties that are relevant for drought forecasting such as persistence.

Drought forecasts can be based on different lead times, ranging from a few weeks to several months and the accuracy of any forecast will decrease with increasing lead times. Nevertheless, so far, there is no reference study providing a general assessment of meteorological drought forecasting over Europe. Such a study is necessary to provide a base for researchers that develop new forecast methods. It is also necessary for decision makers and end users to assess the uncertainties of the warning provided by forecast services.

The European Centre for Medium-range Weather Forecasts (ECMWF) provides two different types of forecasts for this time range: an extended range forecast, with lead times up to 32 days which is issued twice a week and a seasonal forecast, with lead times of up to 12 months issued once a month. The extended range forecast incorporates more recent model developments and is usually of higher resolution (Vitart et al., 2008). The seasonal forecasting system is based on an older model cycle (Molteni et al., 2011), among other significant differences. Analysing the potential of both products requires understanding the property and skill differences between the two systems for the particular application. For the case of droughts such an analysis needs to include both the numerical forecasting skill and the possibilities for binary decisions to issue drought warnings. In particular, the latter is challenging if such decisions are based on probabilistic forecasts.

The objectives of this paper are to analyse the possibilities for issuing 30 day forecasts of drought conditions based on Ensemble Prediction Systems and the Standardized Precipitation Index (SPI, McKee et al., 1993). The latter is a normalized quantification of the precipitation anomalies (Vicente-Serrano, 2006; Dutra et al., 2013) and considered as a good indicator for analyzing meteorological droughts over different time scales (WMO, 2012). Considering the difficulties to predict drought, in this study, we focus on the evolution of the precipitation for the next month, calculating the rainfall anomaly for the same time period (SPI-1). This product, which provides the trend of precipitation for the next month in relation to the climatology, could be combined with routine drought monitoring to create more robust and useful information for stakeholders. To do so, the extended range and seasonal forecasting systems are compared directly but also within the setting of a decision-making framework. Multiple scores as well as multiple methodologies which allow the transformation of probabilistic forecasts into binary decisions are developed and tested.

Underlying issues are : what is the predictability of a drought based on the SPI for a 1 month rainfall accumulation period (SPI-1), what is the most useful model between the Seasonal (SEAS) and the monthly ENSemble system (ENS) for forecasting 30 day cumulative precipitation; and what are the spatial and temporal variabilities of the model's ability? Adapted skill scores provide information about the ability of the probabilistic models to accurately forecast such kind of extreme events. The paper is organized as follows; the tools and methods used will be detailed in Sect. 2 and the results will be discussed in Sect. 3. Final conclusions are drawn in Sect. 4.

2 Data and methods

2.1 Precipitation

2.1.1 Observations

In this study, the combined gridded precipitation dataset from the ENSEMBLES project and ECA & D (Haylock et al., 2008; Van den Besselaar et al., 2011, E-OBS Version 5) was used which is available from 1950 onwards and is continuously updated. The spatial resolution of the dataset is 0.25° by 0.25°, which was up-scaled by averaging the cumulative precipitation to a 1° by 1° grid as this analysis focuses on large-scale droughts.

Validation of the original datasets has been performed by Pereira et al. (2013) and Sunyer et al. (2013), who found that datasets from ECA & D show higher values for extreme precipitation, and E-OBS tends to over-smooth the data. This can generate some problems when analysing intense precipitation events but appears of secondary importance in drought analysis. Daily precipitation values have been aggregated to monthly values to provide comparison with monthly forecasts. To be consistent with the data provided by the ensembles from ECMWF, a common period of the hindcast that covers the period from 1992 to 2013 is used to calculate the precipitation anomalies.

2.1.2 Forecasts

Two sets of coupled ensemble forecasting systems are provided by ECMWF to forecast one month ahead: an extended range monthly forecast and a seasonal forecast.

The ECMWF monthly (32 day) extended range ensemble forecasting system (ENS hereafter; Vitart, 2004), has been routinely issued twice a week since October 2011. This model is the latest version of the ECMWF Integrated Forecasting System. For lead times up to day 10 the model is not coupled to the ocean and has a resolution of $\sim 32\,\mathrm{km}$ (T639). It is forced by persistent sea-surface temperature anomalies. Beyond a lead time of 10 days the resolution of the model is coarser (T319, 64 km), however, it is coupled to an ocean model. The vertical resolution remains unchanged during the entire simulation at 62 vertical levels.

ECMWF provides a back statistic (hindcasts) for ENS which is a 5-member ensemble starting on the same day and month as each Thursday's real-time forecast for each of the past 20 years. For a more detailed description see Vitart (2014).

The second ECMWF ensemble system used in this study is the seasonal forecast called System 4 (Molteni et al., 2011; SEAS hereafter), which is launched once a month (on the first day of the month). It has lead times up to 13 months and a resolution of T255 (80 km). This model is the 2011 version of the Integrated Forecast System, with 91 vertical levels. SEAS provides a back statistic, which is a 15/51 member ensemble (number depends on month) identical to SEAS for every month from 1980 onwards. In this study, only the first forecast month is used.

SEAS and ENS are composed of 50 members, which are generated by perturbing initial conditions and physical tendency (Molteni et al., 1996; Weisheimer et al., 2014) and one unperturbed member. Both datasets were re-gridded to a one square degree resolution using a mass conservative interpolation. The two systems will be compared over their hindcast periods as well as over a forecast period as can be seen in Table 1. This allows for a larger sample size and enables a more significant comparison.

However, despite this technique being robust and frequently used, it also has a few disadvantages: the ensemble size of the reforecasts is only five members instead of 51 members for the real-time forecasts. Ensemble size can have an impact on skill scores, which needs to be corrected for. Weigel et al. (2008) faced the same issue when they scored the ECMWF reforecasts produced in 2006 and used a correction of the probabilistic skill score which takes into account the ensemble size.

2.2 Drought detection

In this study the Standardized Precipitation Index (SPI) is used to detect droughts. It was developed by McKee et al. (1993) and is currently used in many scientific studies or operational systems (Guttman, 1999; Khan et al., 2008; Dutra et al., 2013, 2014). SPI has the advantage that it is provides easily understandable information about the precipitation anomaly. In addition it is also very flexible, allowing calculations aggregated over

different spatial scales (from station data to large-scale area) as well as temporal domains (from 10 day to several month's cumulative precipitation, Mishra and Desai, 2006; Cacciamani et al., 2007).

This study focuses on the monthly timescale and therefore the SPI was calculated using monthly accumulated precipitation (SPI-1). The SPI is usually computed by fitting a probability density function (often a Gamma distribution) to the data (Lloyd-Hughes and Saunders, 2002; Edossa et al., 2010; Dutra et al., 2013; Guy Merlin and Kamga, 2014). Through the application of an inverse normal (Gaussian) function, data are transformed into normal space with a mean equal to 0 and a standart deviation (SD) equal to 1. It is important that the hypothesis that the data can be approximated by a Gamma distribution is tested to ensure that all conclusions are valid. The Gamma function cannot be fitted when only a low number of data points (events) or very low data values (precipitation) exist because numerical convergence of the optimization process cannot be achieved. Therefore, the SPI methodology cannot be applied in very arid regions.

The SPI value can be broken down into different classes (WMO, 2012): normal conditions from -1 to 1; moderate drought with SPI < -1; severe droughts with SPI < -1.5; and extreme drought for SPI < -2. The time series of the analysed forecasts in this paper are too short to justify any focus on an SPI lower than -2 (last 2.3 % of the distribution). Therefore, this study focuses on moderate and severe droughts only. One strong advantage of this method is that it produces an unbiased product with a homogeneous rank histogram (Talagrand Diagram) of the observed precipitation onto the forecasted precipitation (not shown).

2.3 Deriving a decision from probabilistic forecasts

One of the main objectives of this work is to provide decision makers and end users with a simple and robust Boolean index to forecast a drought based on a probabilistic forecasting system. Several methods to select the Boolean solution are tested and are compared with a deterministic model (defined here as the unperturbed member of the Ensemble). Also, a comparison against a climatological forecast will be performed. Methods to derive this

index are given in Table 2 and can be categorized into three types: individual, where the index is based on an individual member or percentile; partially integrative, where the sum of particular individual members or percentiles are used; and integrative which is represented by the ensemble mean. The individual types should be seen as providing complementary information giving information about the intensity of the SPI-1, but also the distribution of the members.

The individual types have been subdivided into 5 classes representing dry members (Q13, Q23), wet ones (Q77, Q88) or the median. The extreme members of the distribution are not used to avoid outliers generally associated with ensemble systems (Lavaysse et al., 2013). For each method, a threshold was defined. A SPI lower than -1 or -1.5 will select 16 % and 6.7 % respectively of the normalized series. Therefore, to be coherent, the thresholds have been defined to select the same number of events.

2.4 Evaluation scores

A plethora of scores to evaluate probabilistic forecasts exist (Nurmi, 2003) and in this study we have chosen scores which are suitable for drought forecasting.

The Relative Operating Characteristic (ROC) score was proposed by Mason (1982) and is plotting the false alarm rate against the hit rate. The objective of that score is to calculate the ability of the forecast to discriminate between events and non-events. This score is not bias sensitive to the forecast and can be considered as a measure of potential usefulness because it is conditioned by the observations (i.e., given that a drought occurred, what was the corresponding forecast?). The area under the ROC curve can be calculated and ranges between 0 and 1. Higher numbers indicate a better forecast.

The reliability diagram, which is conditioned on the forecasts, is a good complementary score to the ROC because it assesses the average agreement between the forecast values and the observed values. In a reliability diagram the forecast probability is plotted against the observed relative frequency (Nurmi, 2003). A perfect score is associated with the $1:1$ line, the climatology score (i.e. no resolution) corresponds to the mean observed frequency (i.e. observed relative frequency of $y = 0.159$ for SPI < -1).

The accuracy of the probability forecasts is assessed using the Brier Score (Brier, 1950) :

$$BS_f = \sum_{k=1}^{r}\sum_{j=1}^{m}(p_f(j,k) - I_o(j,k))^2 \tag{1}$$

where p_t is the probability that was forecast, I_o the observation of the event (1 or 0 if it does happen or not), r the number of classes (here 2) and m is the number of forecasting instances. A skill score can be derived by comparing the Brier score to climatology.

$$BSS = 1 - BS_f/BS_c \tag{2}$$

The Brier Skill Score ranges from $-$inf to 1. The higher the score the more skilful is the forecast and any negative values indicate that the climatological forecast outperforms the probabilistic forecast. The scores above are complemented by the correlation of the ensemble mean and the Root Mean Square Error of the ensemble mean as those are frequently used in the evaluation of seasonal forecasts.

Several scores exist which deal with the contingency table and where the forecasted and observed solutions are Booleans. In this paper, we have used 5 of them. The Probability Of Detection (POD, perfect $=$ 1) is the ratio of the total number of observed events that have been forecasted.

$$POD = \frac{hits}{hits + misses} \tag{3}$$

The False Alarm Rate (FAR, perfect $=$ 0) is the fraction of the forecasted events which actually did not occur.

$$FAR = \frac{false\,alarms}{hits + false\,alarms} \tag{4}$$

The extreme dependency score (EDS, see equation 5) is an informative assessment of skill in deterministic forecasts of rare events that can converge to different values for different

forecasting systems and furthermore it does not explicitly depend upon the bias of the forecasting system. (Ferro and Stephenson, 2011).

$$EDS = \frac{2log(\frac{hits+misses}{total})}{log(\frac{hits}{total})} - 1 \tag{5}$$

The percent correct (PC, perfect = 1) is the ratio of good forecasting events in relation to the total number of events.

$$PC = \frac{hits + correct\,negative}{total} \tag{6}$$

Finally, the Gilbert score balances POD and PC cases (Jolliffe and Stephenson, 2003; Hogan et al., 2010) and measures the fraction of observed and/or forecasted events that were correctly predicted, and adjusted for hits associated with random chance.

$$GSS = \frac{hits - hits_{random}}{hits + misses + false\,alarms - hits_{random}} \tag{7}$$

3 Results

3.1 Evaluation of the SPI calculation

The sensitive part of the SPI calculation is the fitting of a theoretical distribution to the empirical distribution. In this study, the Gamma distribution is fitted to the probability density function of monthly precipitation. It is therefore necessary to set a threshold at which minimum cumulative precipitation can be considered as significant.

Different thresholds were tested (0, 1, 5, 10 and 20 mm, not shown) and it was decided that only monthly precipitations larger than 10 mm are considered significant. This threshold allows the retention of a large number of events and the discarding of events or regions with non significant monthly accumulated rainfall. As outlined in the methodology, fitting a Gamma distribution to precipitation data relies on an adequate sample size (adequate with

respect to the variability of the data). The Gamma distribution was fitted to the distribution if a grid point possesses at least 66 % of values significantly larger than 0 (i.e. larger than 10 mm). That ensures a minimum number of events to fit the distribution. These thresholds allow for the removal of arid areas, where the fitting of the Gamma distribution resulted in biased values due to the low spread and low sampling of the time series.

The performance of the fitting procedure and of the underlying assumptions can be analysed by investigating the resulting SPI-1 distribution. This was done by calculating the integral of the differences between the fitted Gamma distribution and the empirical distribution. Zero values are considered as perfect values (no bias of the SPI-1 calculated), whereas positive or negative values indicate bias and therefore question the validity of the fitting procedure. In Fig. 1 the bias of the Gamma distribution over the entire globe is shown. It can be seen that the Gamma distribution is well adapted for most of Europe (see also Stagge et al. (2015)).

Nevertheless, the low precipitation amounts over the southern part of Spain can create some bias in the fitting. This is especially true during the summer season and therefore the assumptions for fitting the Gamma distribution are not valid for the entire year. This analysis shows that it will be necessary to adapt the method in particular over dry areas, for example, by focusing the study only during the rainy seasons.

3.2 Validation during the hindcast period

This evaluation is based on the hindcast period (see Table 1) of ENS and SEAS. It allows a long-term evaluation using the same version of the model. The correlation and root mean square error of the ensemble means are displayed in Fig. 2. The mean correlation (0.32) and the mean RMSE (1.02) for ENS is better than that for SEAS (0.05 and 1.45 respectively, not shown). Neither the correlation nor the RMSE are significantly different from zero suggesting that a mean monthly forecast has no skill. In addition, the spatial variability is low, meaning that there is no significant spatial difference in the ability of the model to predict the SPI-1, on average. Note that the correlation of the SPI-1 is comparable to the anomaly correlation coefficient (ACC) that removes the seasonal cycle. Indeed, the SPI-1 is

the anomaly of a monthly precipitation in relation to the climatology of that specific month. So this correlation coefficient is much more robust but also less likely significant.

The SPI-1 values of individual ensemble members and observations were analysed in bins to assess whether these results are also valid for extreme events. Here, the individual ability (for each member independently) was assessed by decomposing the SPI-1 forecasted and observed over Europe during the hindcast period in 10 classes (from SPI-1 lower than -2, to SPI-1 larger than +2, at intervals of 0.5). The frequency in each bin naturally follows the Gamma distribution which generates a large number of cases centred around 0. This distribution was normalized by computing the ratio between the empirical distribution and the theoretical distribution. The result is shown in Fig. 3. The figure shows that the more a drought is forecasted, the more it is observed (red bars). In addition it has to be noted that the distribution is highly non-symmetric. This indicates that the forecasts of extreme dry events are more accurate than the forecasts of extreme wet events. This result could be explained by the usually large spatial and temporal scales of drought events that are better predictable by a global model even one month ahead.

3.3 Validation during the forecast period

The analysis of the forecast period from November 2012 to November 2013 largely confirms earlier findings in this paper of the forecasts over a significantly longer temporal period, but allows for a more detailed investigation of the distributions due to the larger ensemble number (see Table 1).

Figure 4a compares the behaviour of the ENS members during observed extreme wet and dry events. In both cases, the normal distributions of the ranked ensemble members are quite similar. The only difference is the shift towards negative values of forecasted SPI when a drought is observed (red line) in comparison with when wet events are observed (blue line). Nevertheless, the SD (indicated by the barlines) highlights that there is no significant difference (significance level of 0.9) between the two events. It is interesting to observe that the value of the ensemble mean increases with the increase of the observed SPI-1 (black line in Fig. 4b), whereas the spread of the ensemble (defined as the SD) shows

little sensitivity (yellow line in Fig. 4b). It can be concluded that only the ensemble mean displays a significant difference between wet and dry anomalies, whilst there is no such relation in the SD. In SEAS, the same trends are observed but the difference between the two conditional distributions is reduced (Fig. 4c and d). This indicates that ENS has a stronger resolution than SEAS, and therefore a greater ability to discriminate events with different frequency distributions.

These results are confirmed by analysing the ROC curve. Over the European continent, the ROC curves show an improvement in relation to the no skill curve (1 : 1 in Fig. 5). The ROC area is slightly better for ENS than for SEAS ($+0.4$ and $+0.2$ for SPI-1 < -1 and SPI-1 < -1.5, respectively).

Both ENS and SEAS present a positive but low reliability for detection of SPI-1 < -1 (Fig. 6). Indeed, the observed relative frequencies increase with the increase in the forecast probabilities. The distribution of cases per percentage (not shown) indicates more events with a large percentage of members associated with a drought in ENS rather than SEAS. This result indicates the better consistency between the members in ENS to forecast an extreme rainfall deficit than in a case of SEAS. Using ENS, several events are forecasted with more than 93 % of members associated with a drought forecasting, whereas using SEAS, the maximum is 81 %. The ENS and SEAS systems are better than climatology, achieving values of 0.14 and 0.12 respectively. But, here the difference between ENS and SEAS is not significant.

3.4 Sensitivity to drought scales

All analysis so far has been performed on a scale of 1 ° by 1 °, however the sensitivity to different resolutions needs to be analysed, because the impacts of large-scale droughts will be stronger. Figure 5 shows SPI-1 values smoothed to 3 and 5 square degrees, using a simple upscaling method based on the average of the values. The resolution of about 1 square degree has been kept to compare the impact of the resolution in the native grid. The results show a slight improvement of the ROC area with a coarser resolution (broken and dotted lines in Fig. 5). The smoothed signal favours the large-scale signatures that are

better represented in models than small-scale structures of droughts. The effect of spatial upscaling can also be seen in the ROC results as a little positive impact of SEAS for the largest forecast probabilities (Fig. 6d). However as mentioned previously, the number of events in these cases is low. The effect has been quantified using the BSS (see equation 2), which goes up to 0.17 and 0.14 respectively for the 5-degree smoothed signal.

3.5 Spatial and seasonal variabilities

3.5.1 Spatial variability

The analysis so far has ignored the spatial and seasonal scale. Figure 7 shows the ROC anomaly for the forecast period, which is the ROC area for each grid cell in relation to the average (0.67 for ENS). The anomaly is preferred to the raw value to highlight regions where the ROC is improved or reduced. A maximum variability of 20 % can be observed. For the hindcast period (not shown), this variability is much lower at $\sim 6\,\%$. There is a difference in spatial patterns between the two periods, suggesting that the spatial patterns are not significant and are mainly driven by the extreme cases encountered during the period.

3.5.2 Seasonal variability

A seasonal decomposition is used to highlight the temporal variabilities. ROC scores and curves where independently calculated for the autumn (September to November), winter (December to February), spring (March to May) and summer (June to August) seasons and are displayed in Fig. 8 (for SPI-1 < -1).

The four ROC areas are very similar, and the four distributions are identical for ENS, meaning that the skill to forecast droughts is identical throughout the year. In contrast, SEAS shows some differences between the seasons, with a small improvement in the forecast during the autumn season. Identical interpretations can be derived for the SPI-1 < -1.5 and are therefore not shown.

3.6 Index performance

Figure 9 shows the POD (see equation 3) and the FAR (see equation 4) for ENS and SEAS. POD indicates that, on average, one in three drought events over Europe is correctly forecasted one month in advance. This is significantly better than the climatology (16 %) and better than the deterministic forecast (around 25 %, green line in Fig. 9).

The importance of the drought duration has also been tested. The scores were calculated independently for a drought onset (first SPI-1 lower than thresholds), persistence (consecutive SPI-1 lower than the threshold), or end of the drought (first SPI-1 above the threshold). First, the duration of a large majority of SPI-1 lower than -1 (more than 80 %) is one month (isolated values, dry spell). The scores display a slight increase of the score for the persistent droughts (condition unchanged), for the median the POD score increases from 0.33 to 0.36. But the difference is not significant according to the t-test.

The highest POD is achieved by using the 13 percentile (7th member of the ranked ensemble distribution), and the product using the Q13 and Q23 (noted SpD). The mean of the ensemble (last point on the right of each panel), which is used widely, is not the best method for detecting droughts.

The POD values of the wettest members of the ranked distribution (noted Q77 and Q88 in Fig. 9) give the worst results of all methods, meaning that there is a low consistency between the extreme dry and wet members. The FAR displays a low variability between the methods, but every single one is better than the deterministic solution (red lines). It is also worth noting that, using the ENS, the driest members are associated with a decrease of FAR in relation to the dry members. This can be explained by the previous scores, which show a larger consistency between the members. However, it could also be due to a technical effect, since the number of events selected is constant, these scores could be dependent on each other.

The highest EDS is achieved for the driest members (Q13 and Q23, Fig. 10), whereas the wettest members (Q77 and Q88) have the lowest scores. The score of the ensemble mean is better than that of the median. Even if the POD and FAR differences are partially

statistically significant, the improvement of the EDS for the driest members is significant for all differences larger than 0.04.

ENS and SEAS are reliable (see Fig. 6) and hence a potentiel method for drought forecasting could be simply based on the percentage of ensembles predicting a drought. In total, 10 different percentage thresholds were selected. Figure 11 shows that the percentage correct (PC, see equation 6) is increasing with the increase in the percentage used for both models (black points in Fig. 11a and c) which is in agreement with the positive reliability. This means that with more members forecasting a drought, the chance to observe one is increased. However, with an increasing threshold, the number of misses also increases (provided by the POD value, red points in Fig. 11a and c). For example, if the threshold to determine a drought is defined with the 10 % of members associated with a drought forecasting, around 80 % of droughts that occurred were correctly detected (red points), but more than 50 % of those forecasted are associated with false alarms. Contrarily, if the threshold of detection is defined with a percentage larger than 70 %, the percentage correct is about 85 %, but the POD is close to 3 %. Based on this result, the user can tune the percentage depending on an acceptable false alarm ration and misses.

The maximum Gilbert score (Fig. 11b and d, see equation 7) is achieved for a threshold of 30 % for ENS and 40 % for SEAS. In that case, 40 % of droughts observed are forecasted and 75 % of forecasts are hits. The number of missed events becomes too high with a larger percentage threshold, whereas for lower percentage thresholds the errors are associated with false alarms.

3.7 Assessing the uncertainties of the forecasts

Several previous studies (He et al., 2009; Palmer, 2000; Georgakakos et al., 2004; Doblas-Reyes et al., 2009) have shown that probabilistic simulations can provide additional information to assess the uncertainties of the simulation.

The idea here is to estimate the quality of the forecast, based on a specific behaviour of the simulation. So the characteristics of the ensemble in the four different cases of the contingency table have been analysed. This table has been built using the threshold of SPI-

$1 < -1$ to detect a drought and the forecast method is based on the median of the members. The mean SPI-1 of the 51 ranked members for the four cases is illustrated in Fig. 12. During correct negative events (i.e. events without droughts forecasted nor observed), where more than 70 % of the events are located, a normal distribution is observed, with a mean slightly larger than 0. During the missed cases, the median is very close to 0 and the distribution of the ranked members is very close to the ensemble mean.

In addition, the spread of the members is displayed (barb lines) and shows the increase of the spread for extreme members. The fact that the two distributions become undistinguishable means that the response of the model is no different to a normal distribution and it is not significant to find a specific behaviour of the model to assess the missed events.

Finally, the distributions of the members during hits and false alarms are compared. In that case, there is no significant difference. The average and the distribution of the mean SPI-1 of the ensemble are quite similar. These results are in agreement with Table 3, which quantifies the ensemble spread for each case in the contingency table. Based on these results, it appears impossible to evaluate the uncertainties of the ensemble simulation associated with a Boolean decision.

4 Discussion

Most drought studies use SPI with three- to six- months or even longer accumulation periods for drought monitoring and characterization. To forecast droughts over such long-term periods a very accurate and reliable atmospheric model is required. Since it is well known that the current reliability of precipitation forecasts decreases drastically after the first month, the benefit of using a lead time of two months or more is, however, not obvious (Dutra et al., 2013).

This paper, therefore, looks as a first step at the possibilities to provide a reliable one month forecast over the European continent. This information, in combination with monitoring data such as satellite or in-situ measurements that provide an accurate characterization of ongoing drought conditions (e.g. during the last two months), can provide the best

estimate of near future conditions. However, also such a combination of monitoring and forecasting data will not allow looking more than one month ahead and an amalgamation of both information types would bias the testing of the forecast skill, which is the intention of this paper. Several meteorological services or agencies, such as the Bureau of Meteorology in Australia or the United States National Drought Mitigation Center, provide relevant monitoring data as well as a one month outlook. For the case of Europe, the European Drought Observatory (EDO) at the Joint Research Centre of the European Commission provides relevant monitoring data, but up to now lacks the forecast beyond 7 days.

A one month forecast with a good reliability is considered to be a very valuable product for decision makers as it provides information on the probability of occurrence of a dry spell (in case of ongoing normal conditions) and of the probable persistence or end of a drought (in case of an ongoing precipitation deficit). Before providing such information, it is however necessary to assess the quality of the forecasts, which was the first aim of this study. The second objective was to define the most robust (Boolean) method to activate alert levels for the end users of the forecast information. Both steps are essential in an operational early warning environment.

5 Conclusions

This study provides the first assessment of the predictability of meteorological droughts over Europe and of the ability to issue an early warning of such droughts with a one month lead time. The analysis is based on the one month forecast of the SPI-1 from the precipitation outputs provided by two ECMWF ensemble systems. In a first step the ability to forecast SPI-1 from the ensemble outputs was tested, showing that

- The reliability of the ensemble is better than the climatology,
- The spatial variability of the scores can reach up to 20 % over Europe and the seasonal variability is not significant,
- Ensemble models are better at forecasting large-scale droughts.

In a second step the ability to provide a robust Boolean index for drought forecasting was analyzed. The best method is defined by using a threshold of 30 % of ensemble members associated with a drought. In that case, slightly more than 40 % of the droughts observed are forecasted correctly one month ahead, with only 25 % of false alarms. This is significantly better than using the climatology (16 %) or the deterministic models (around 25 %). Finally, this study has shown that there is no possibility to provide uncertainties associated with the boolean index.

By providing the first assessment of meteorological drought forecasting in Europe, this work will be particularly useful by as a benchmark comparison for future studies using, for example, statistical weather prediction methods based on atmospheric predictors, which are better represented in the seasonal models. As a follow-up of the analysis presented in this paper work, we will assess the advantages of predicting droughts by analysing specific Weather Types that are related to the occurrence and persistence of droughts in Europe (Kingston et al., 2015). It could further be useful to investigate the use of moving windows of 10 day cumulative precipitation to detail the temporal behaviour of the forecasted SPI-1. As the forecast skills are better for short lead times, an SPI-1 lower than -1, explained by a strong decrease in precipitation at the beginning of the period, should be more reliable.

References

Arribas, A., Glover, M., Maidens, A., Peterson, K., Gordon, M., MacLachlan, C., Graham, R., Fereday, D., Camp, J., Scaife, A., Xavier, P., McLean, P., Colman, A., and Cusack, S.: The GloSea4 ensemble prediction system for seasonal forecasting, Mon. Weather Rev., 139, 1891–1910, 2011.

Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S., and DeWitt, D. G.: Skill of real-time seasonal ENSO model predictions during 2002-11: is our capability increasing?, B. Am. Meteorol. Soc., 93, 631–651, 2012.

Below, R., Grover-Kopec, E., and Dilley, M.: Documenting drought-related disasters a global reassessment, J. Environ. Develop., 16, 328–344, 2007.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, 1950.

Buizza, R., Houtekamer, P., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M.: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems, Mon. Weather Rev., 133, 1076–1097, 2005.

Cacciamani, C., Morgillo, A., Marchesi, S., and Pavan, V.: Monitoring and forecasting drought on a regional scale: Emilia-Romagna Region, in: Methods and Tools for Drought Analysis and Management, Springer, 29–48, 2007.

Doblas-Reyes, F., Weisheimer, A., Déqué, M., Keenlyside, N., McVean, M., Murphy, J., Rogel, P., Smith, D., and Palmer, T.: Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts, Q. J. Roy. Meteor. Soc., 135, 1538–1559, 2009.

Dutra, E., Di Giuseppe, F., Wetterhall, F., and Pappenberger, F.: Seasonal forecasts of droughts in African basins using the Standardized Precipitation Index, Hydrol. Earth Syst. Sci., 17, 2359–2373, doi:10.5194/hess-17-2359-2013, 2013.

Dutra, E., Pozzi, W., Wetterhall, F., Di Giuseppe, F., Magnusson, L., Naumann, G., Barbosa, P., Vogt, J., and Pappenberger, F.: Global meteorological drought – Part 2: Seasonal forecasts, Hydrol. Earth Syst. Sci., 18, 2669–2678, doi:10.5194/hess-18-2669-2014, 2014.

Edossa, D. C., Babel, M. S., and Gupta, A. D.: Drought analysis in the Awash river basin, Ethiopia, Water Resour. Manag., 24, 1441–1460, 2010.

Eshel, G., Cane, M. A., and Farrell, B. F.: Forecasting eastern Mediterranean droughts, Mon. Weather Rev., 128, 3618–3630, 2000.

Ferro, C. A. and Stephenson, D. B.: Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events, Weather Forecast., 26, 699–713, 2011.

Fraser, E. D., Simelton, E., Termansen, M., Gosling, S. N., and South, A.: “Vulnerability hotspots”: integrating socio-economic and hydrological models to identify where cereal production may decline in the future due to climate change induced drought, Agr. Forest Meteorol., 170, 195–205, 2013.

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298, 222–241, 2004.

Guttman, N. B.: Accepting the Standardized Precipitation Index: a calculation algorithm, J. Am. Water Resour. As., 35, 311–322, 1999.

Guy Merlin, G. and Kamga, F. M.: Computation of the Standardized Precipitation Index (SPI) and its use to assess drought occurrences in Cameroon over recent decades, J. Appl. Meteorol. Clim., 53, 2310–2324, 2014.

Hamill, T. M., Brennan, M. J., Brown, B., DeMaria, M., Rappaport, E. N., and Toth, Z.: NOAA's future ensemble-based hurricane forecast products, B. Am. Meteorol. Soc., 93, 209–220, 2012.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res.-Atmos., 113, D20119, doi:10.1029/2008JD010201, 2008.

He, Y., Wetterhall, F., Cloke, H., Pappenberger, F., Wilson, M., Freer, J., and McGregor, G.: Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions, Meteorol. Appl., 16, 91–101, 2009.

Hogan, R. J., Ferro, C. A., Jolliffe, I. T., and Stephenson, D. B.: Equitability revisited: why the "equitable threat score" is not equitable, Weather Forecast., 25, 710–726, 2010.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification, A Practtioners Guide in Atmospheric Science, John Wiley and Sons, Chichester, UK, 240 pp., 2003.

Khan, S., Gabriel, H., and Rana, T.: Standard precipitation index to track drought and assess impact of rainfall on watertables in irrigation areas, Irrig. Drain. Systems, 22, 159–177, 2008.

Kim, T.-W. and Valdés, J. B.: Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks, J. Hydrol. Eng., 8, 319–328, 2003.

Kingston, D. G., Stagge, J. H., Tallaksen, L. M., and Hannah, D. M.: European-scale drought: understanding connections between atmospheric circulation and meteorological drought indices. Journal of Climate, 28(2), 505-516, 2015.

Lavaysse, C., Carrera, M., Bélair, S., Gagnon, N., Frenette, R., Charron, M., and Yau, M.: Impact of surface parameter uncertainties within the Canadian Regional Ensemble Prediction System, Mon. Weather Rev., 141, 1506–1526, 2013.

Lloyd-Hughes, B. and Saunders, M. A.: A drought climatology for Europe, Int. J. Climatol., 22, 1571–1592, 2002.

Mason, I.: A model for assessment of weather forecasts, Aust. Meteorol. Mag., 30, 291–303, 1982.

McKee, T. B., Doesken, N. J., and Kleist, J.: The relationship of drought frequency and duration to time scales, in: Proceedings of the 8th Conference on Applied Climatology, Anaheim, CA, USA, Am. Meteorol. Soc., 179–184, 1993.

Mishra, A. and Desai, V.: Drought forecasting using stochastic models, Stoch. Env. Res. Risk A., 19, 326–339, 2005.

Mishra, A. and Desai, V.: Drought forecasting using feed-forward recursive neural network, Ecol. Model., 198, 127–138, 2006.

Mishra, A., Desai, V., and Singh, V.: Drought forecasting using a hybrid stochastic and neural network model, J. Hydrol. Eng., 12, 626–638, 2007.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Q. J. Roy. Meteor. Soc., 122, 73–119, 1996.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), European Centre for Medium-Range Weather Forecasts, Reading, UK, 2011.

Nurmi, P.: Recommendations on the verification of local weather forecasts, ECMWF Tech. Memo. 430, 18 pp., 2003.

Palmer, T. N.: Predicting uncertainty in forecasts of weather and climate, Rep. Prog. Phys., 63, 71, doi:10.1088/0034-4885/63/2/201, 2000.

Pereira, S. C., Carvalho, A. C., Ferreira, J., Nunes, J. P., Keizer, J. J., and Rocha, A.: Simulation of a persistent medium-term precipitation event over the western Iberian Peninsula, Hydrol. Earth Syst. Sci., 17, 3741–3758, doi:10.5194/hess-17-3741-2013, 2013.

Richardson, D., Bidlot, J., Ferranti, L., Haiden, T., Hewson, T., Janousek, M., Prates, F., and Vitart, F.: Evaluation of ECMWF forecasts, including 2012–2013 upgrades, Tech. rep., ECMWF Technical Memo, Reading, UK, 2013.

Singleton, A.: Forecasting drought in Europe with the Standardized Precipitation Index, Tech. rep., JRC Sctintific and Technical Reports, Italy, 2012.

Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F., Stahl, K.: Candidate distributions for climatological drought indices (SPI and SPEI). International Journal of Climatology, 2015.

Stockdale, T., Anderson, D., Alves, J., and Balmaseda, M.: Global seasonal rainfall forecasts using a coupled ocean–atmosphere model, Nature, 392, 370–373, 1998.

Sunyer, M. A., Sørup, H. J. D., Christensen, O. B., Madsen, H., Rosbjerg, D., Mikkelsen, P. S., and Arnbjerg-Nielsen, K.: On the importance of observational data properties when assessing regional climate model performance of extreme precipitation, Hydrol. Earth Syst. Sci., 17, 4323–4337, doi:10.5194/hess-17-4323-2013, 2013.

Van den Besselaar, E., Haylock, M., Van der Schrier, G., and Klein Tank, A.: A European daily high-resolution observational gridded data set of sea level pressure, J. Geophys. Res.-Atmos., 116, D11110, doi:10.1029/2010JD015468, 2011.

Vicente-Serrano, S. M.: Differences in spatial patterns of drought on different time scales: an analysis of the Iberian Peninsula, Water Resour. Manag., 20, 37–60, 2006.

Vitart, F.: Monthly forecasting at ECMWF, Mon. Weather Rev., 132, 2761–2779, 2004.

Vitart, F.: Evolution of ECMWF sub-seasonal forecast skill scores, Q. J. Roy. Meteor. Soc., Part B, 114, 1889–1899, 2014.

Vitart, F., Buizza, R., Alonso Balmaseda, M., Balsamo, G., Bidlot, J.-R., Bonet, A., Fuentes, M., Hofstadler, A., Molteni, F., and Palmer, T. N.: The new VAREPS-monthly forecasting system: a first step towards seamless prediction, Q. J. Roy. Meteor. Soc., 134, 1789–1799, 2008.

Weigel, A. P., Baggenstos, D., Liniger, M. A., Vitart, F., and Appenzeller, C.: Probabilistic verification of monthly temperature forecasts, Mon. Weather Rev., 136, 5162–5182, 2008.

Weisheimer, A. and Palmer, T.: On the reliability of seasonal climate forecasts, J. R. Soc. Interface, 11, doi:10.1098/rsif.2013.1162, 2014.

Weisheimer, A., Corti, S., Palmer, T., and Vitart, F.: Addressing model error through atmospheric stochastic physical parametrizations: impact on the coupled ECMWF seasonal forecasting system, Philos. T. R. Soc. A, 372, doi:10.1098/rsta.2013.0290, 2014.

Wilhite, D. A. and Glantz, M. H.: Understanding: the drought phenomenon: the role of definitions, Water Int., 10, 111–120, 1985.

WMO: Standardized Precipitation Index, User Guide, Tech. Rep. 1090, Geneva, Switzerland, 2012.

Table 1. ENS and SEAS configurations for the hindcast and the forecast periods.

Periods	Evaluation Period	ENS	SEAS
Hindcasts	Nov 1992 to Oct 2012	5 members	15/51 members
Forecasts	1 Nov 2012 to 31 Oct 2013	51 members	51 members

Table 2. List of the 10 methods used to provide a Boolean index for drought forecasting using an ensemble system.

Name	Definition
13 percentile (Q13)	member located at the 13 % of the CDF
23 percentile (Q23)	member located at the 23 % of the CDF
Median (MED)	member located at the 50 % of the CDF
77 percentile (Q77)	member located at the 77 % of the CDF
88 percentile (Q88)	member located at the 88 % of the CDF
Large spread (SpL)	sum of the extreme members (Q13 + Q88)
Low spread (Spl)	sum of the members (Q23 + Q78)
Dry spread (SpD)	sum of the dry members (Q13 + Q23)
Flood spread (SpF)	sum of the wet members (Q77 + Q88)
Mean	ensemble mean

Table 3. Contingency table (in percentage) obtained using the median of ENS to forecast a drought. The definition of the drought observed is an SPI-1 lower than −1 and a drought forecasted is when the ensemble median is lower than the 16th percentile. The second values of each case indicate the ensemble spread and its SD is given in brackets.

		drought	observed
		yes	no
drought	yes	4.4 %/2.31 (0.4)	10.7 %/2.37 (0.4)
forecasted	no	10.4 %/1.99 (0.4)	74.5 %/1.88 (0.3)

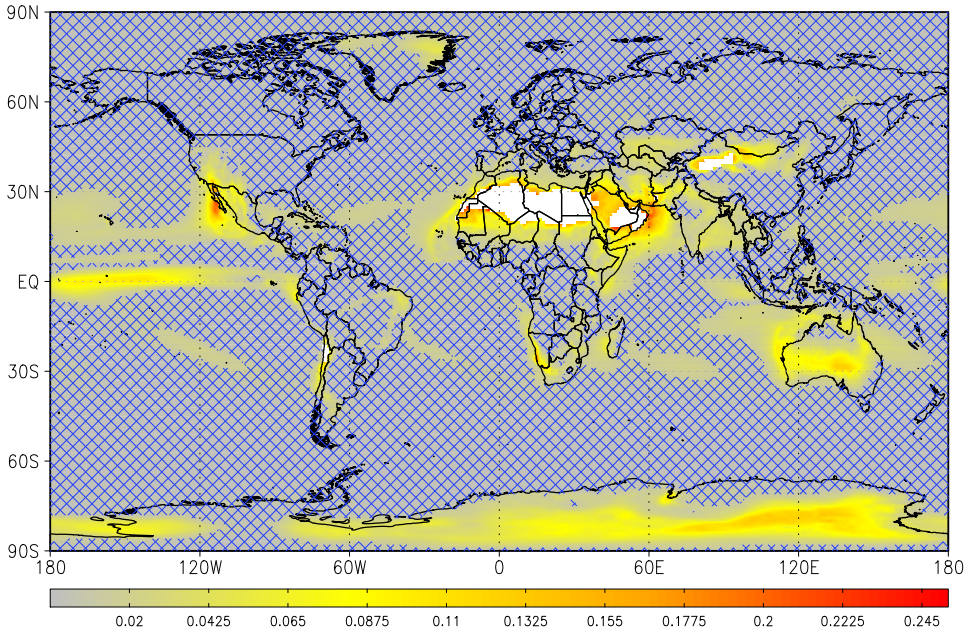
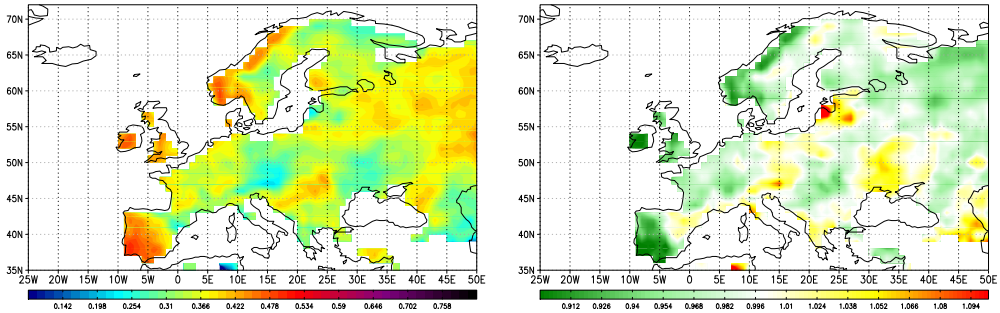


Figure 1. Bias of the SPI-1 calculated between the fitted Gamma distribution and the observed monthly cumulative precipitation (see text for more details). Regions in white are considered as too dry to fit this distribution. Regions where the bias becomes significantly different to 0 (non-hatched areas) could generate bias in the SPI calculation.



(a) Correlation (b) RMSE

Figure 2. (a) Correlation of the forecasted (using the mean of the ensemble) and observed SPI-1 during the hindcast period (from November 1992 to November 2012). **(b)** Same but for the RMSE.

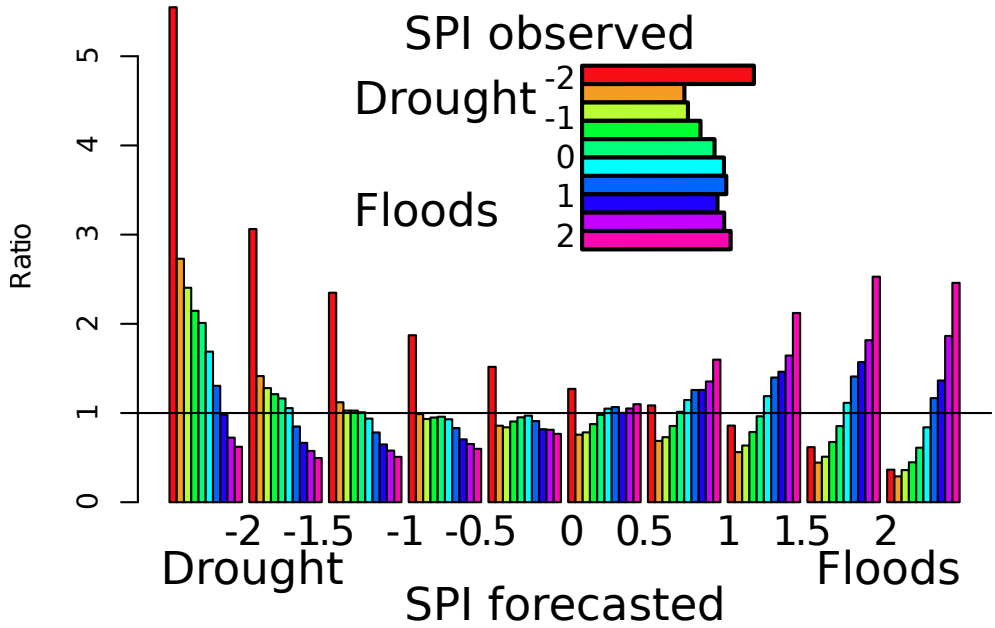
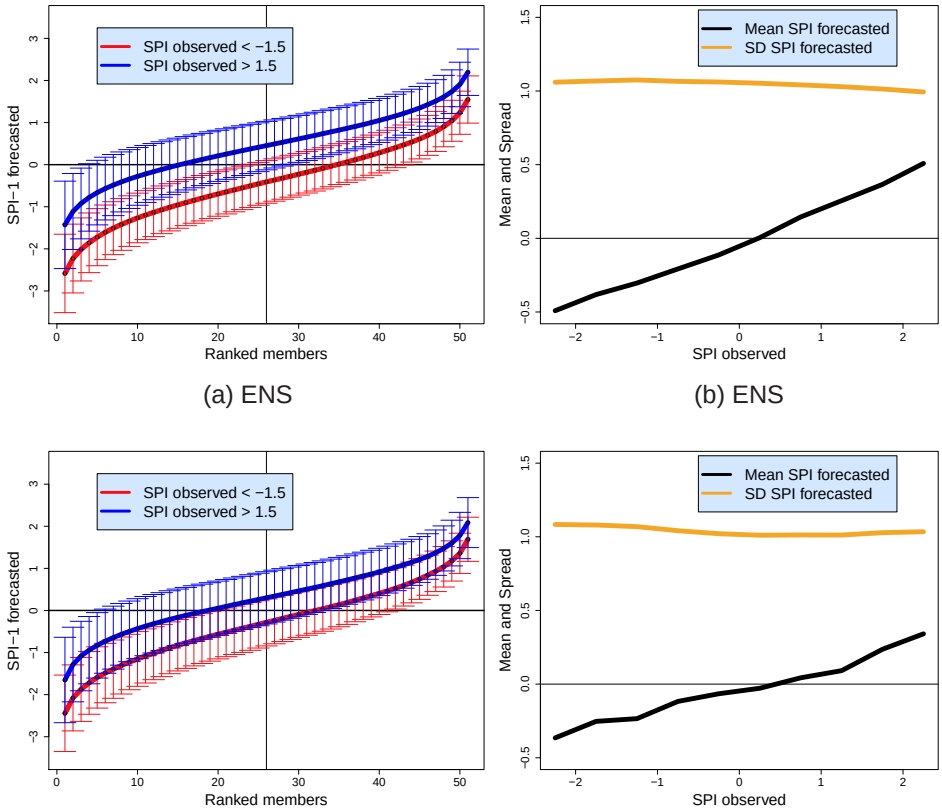


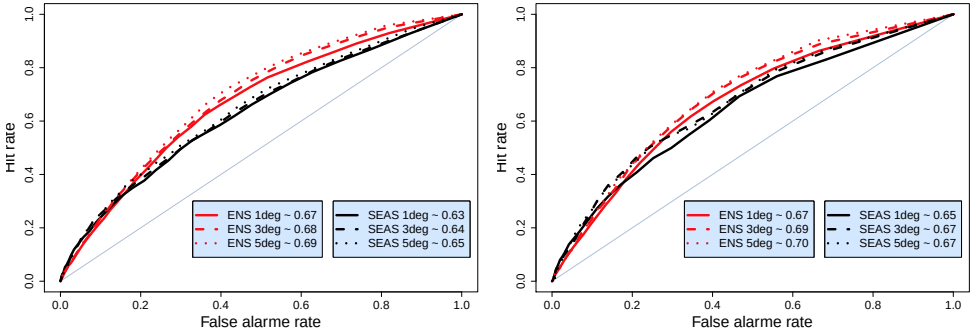
Figure 3. Ratio of events following the forecasted (x axis) and observed SPI-1 (color bars) over Europe using the hindcast period in relation to the theoretical distribution. Results are standardized by the theoretical normal distribution of events.



(a) ENS (b) ENS

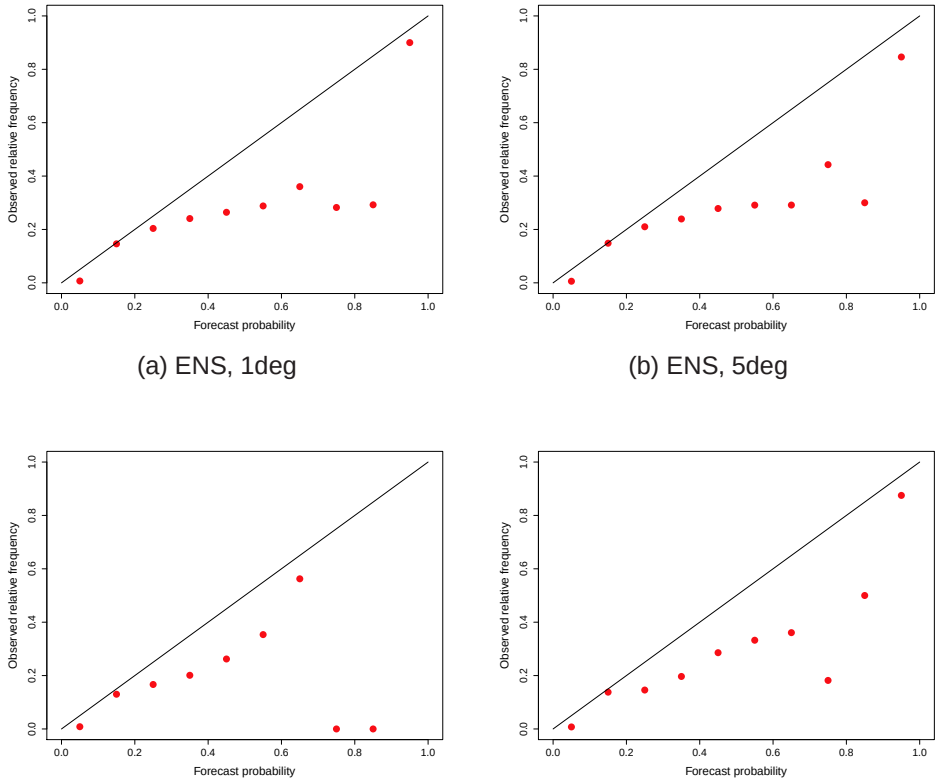
(c) SEAS (d) SEAS

Figure 4. (a) Mean SPI-1 forecasted of ranked members using ENS during observed drought or floods (SPI-1 < -1.5 and SPI-1 > 1.5 respectively). **(b)** Ensemble mean and SD of the SPI-1 forecasted using ENS following the associated observed SPI-1. **(c)** and **(d)** are the same panels as **(a)** and **(b)** but using SEAS.



(a) SPI-1 < -1 (b) SPI-1 < -1.5

Figure 5. ROC curve using ENS and SEAS (red and black lines respectively) for the period from November 2012 to November 2013 over Europe to detect a drought defined as an SPI lower than -1 **(a)** or lower than -1.5 **(b)**. The ROC area values for the different spatial resolutions are indicated.



(a) ENS, 1deg

(b) ENS, 5deg

(c) SEAS, 1deg

(d) SEAS, 5deg

Figure 6. Reliability diagrams for drought detection defined as a SPI-1 lower than -1 using ENS (top panels) and SEAS (bottom panels) in the period from November 2012 to November 2013. The spatial resolution is one square degree (left panels) and 5 square degrees (right panels).

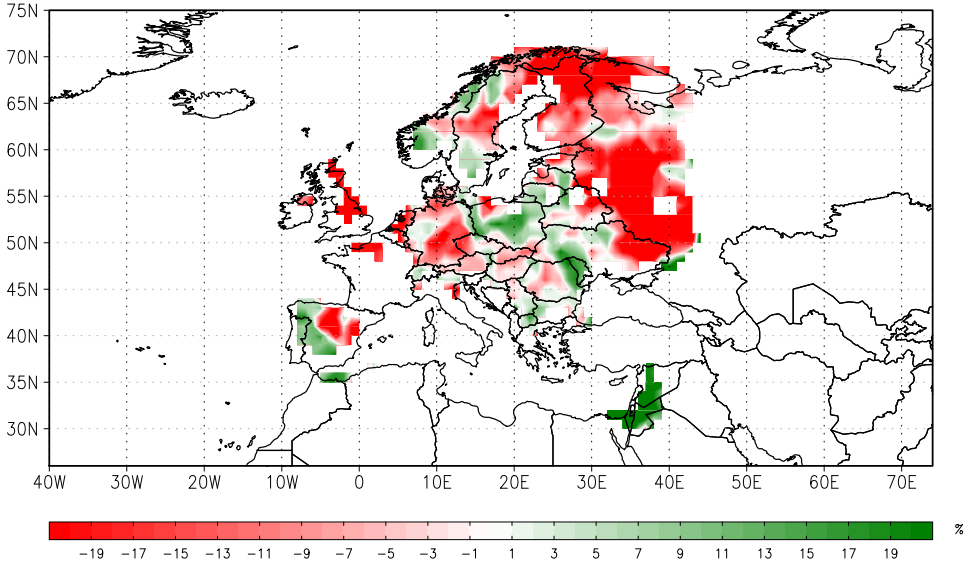
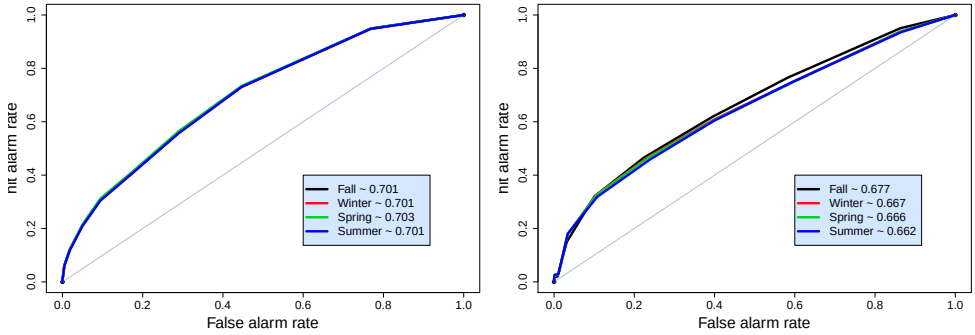
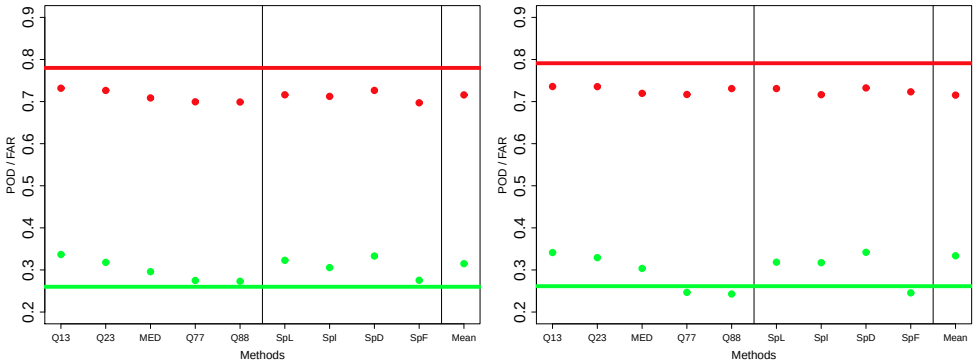


Figure 7. ROC anomaly (in %) in relation to the mean value of the ROC over the domain (equal to 0.67) for the period from November 2012 to November 2013 with drought defined as an SPI-1 < -1.



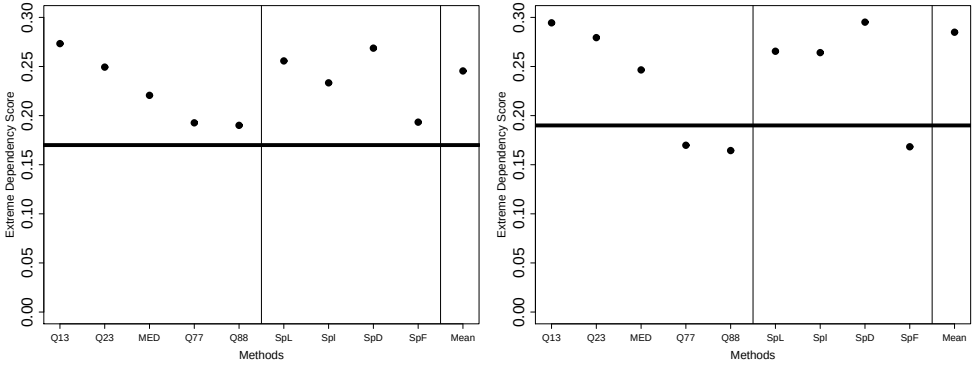
(a) ENS (b) SEAS

Figure 8. Seasonal decomposition of the ROC curves for drought forecasting (with the 5 square degree smoothing) using ENS **(a)** and SEAS **(b)** over Europe for the period from November 2012 to November 2013 with drought defined as an SPI-1 < -1



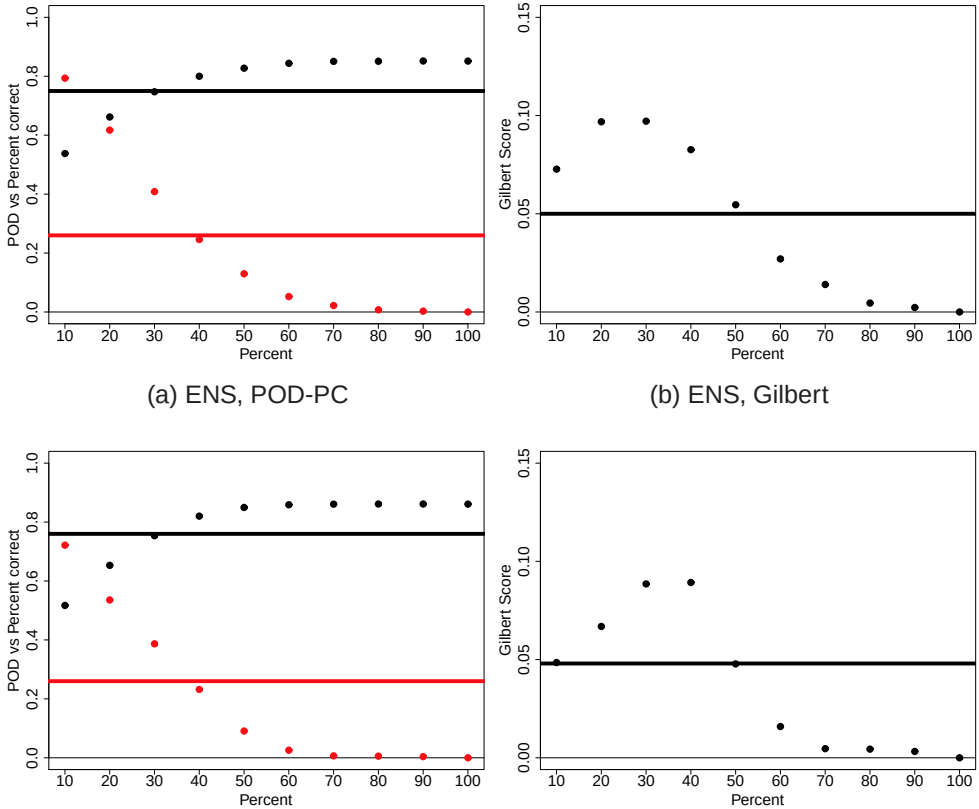
(a) ENS (b) SEAS

Figure 9. Probability of detection (POD, in green, perfect = 1) and False alarm ratio (FAR, in red, perfect = 0) for different methods used to detect drought (x axis), using ENS **(a)** and SEAS **(b)**. Lines indicate the scores of the deterministic model (unperturbed member of the ensemble).



(a) ENS (b) SEAS

Figure 10. Extreme Dependency Score (EDS) for the 10 methods used to forecast a drought (x axis, see Table 1 for more details) using the ENS **(a)** and SEAS **(b)** ensemble system. Black lines indicate the score of the unperturbed member.



(a) ENS, POD-PC (b) ENS, Gilbert

(c) SEAS, POD-PC (d) SEAS, Gilbert

Figure 11. (a) POD (red) and percentage correct (black) using different percentage of members to forecast a drought event using ENS. **(b)** Gilbert score (see text for more details) following the percentage used to forecast a drought using ENS. Lines indicate the score of the deterministic model (unperturbed member). **(c)** and **(d)** are the same panels as **(a)** and **(b)** using SEAS.

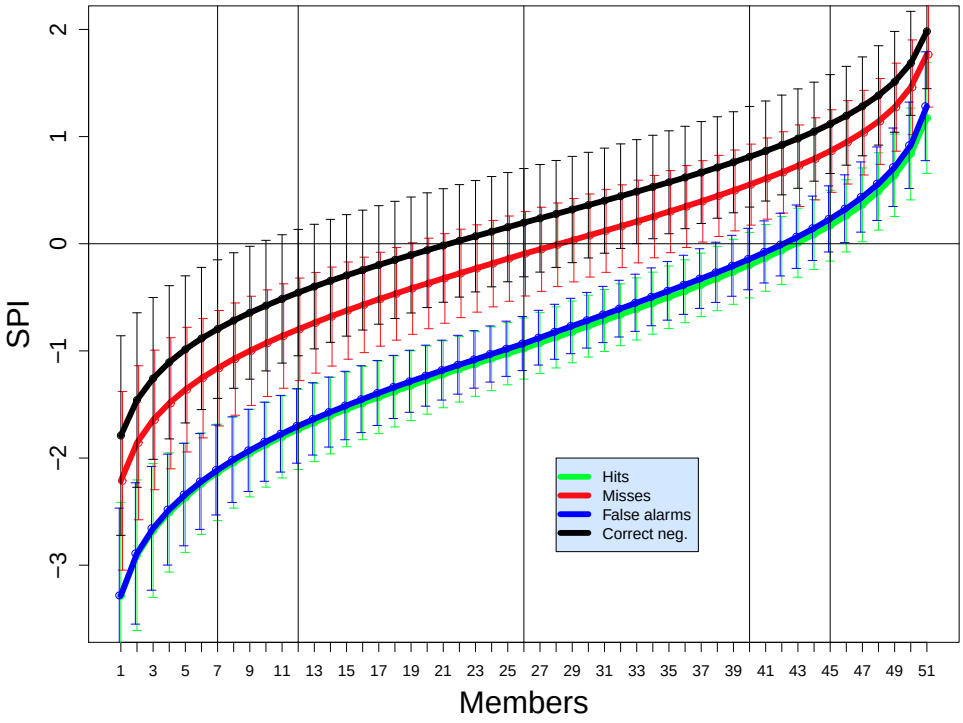


Figure 12. Mean SPI-1 and SD of the ranked members following the four conditions in the contingency table (see Table 2 and text for more details): hits (green), false alarm (red), misses (blue) and correct negative (black line), using ENS. Vertical lines indicate the spread of the members used for the Boolean drought detection methods.