**The text in bold black font are the reviewers' comments.**

The text in black font are the authors' answers to the reviewers' comments. In case of RC1, these are the same answers as were published in the interactive discussion of this paper.

The text in blue font are the specific changes made to the final manuscript as a response to the reviewers' comments.

# RC1

**Major comments**

**1) This perhaps an unusual criticism, but I think the authors may have been a little too hard on their system by choosing ESP as a reference forecast. ESP is not really a 'naive' forecast, and accordingly it is rarely used as a benchmark for performance in seasonal prediction systems. As climatology is often the default assumption by many users of forecasts, it is far more typical as a benchmark. Choosing ESP as a benchmark may have somewhat perverse results: for example, it is possible to have extremely accurate forecasts, but in cases where skill is largely due to IHCs these forecasts will not appear to be skilful (or may even be negatively skilful). This may be compounded by the use of ESP forcings that have not been cross-validated (though I may be wrong here - more information please) - i.e. it appears that an ESP hindcast from, say, 1995, could include a rainfall sequence from 1995 (a perfect forecast!) as one member of its forcing ensemble. In a small ensemble, the effect of one perfect rainfall forecast may offer some advantage to ESP forecasts compared to SyS4. I offer the following suggestions to deal with these issues: i) When introducing the ESP reference forecasts (Section 2.1.2) please note that this is an unusually high benchmark, and why. I would also reiterate this when discussing results. ii) If possible, cross-validate the ESP forcing ensembles (if this hasn't already been done) iii) Note more strongly that the ROC results - which are compared to a naive benchmark - offer a more typical assessment of performance compared to CRPSS/MAESS scores calculated against ESP.**

We agree that the ESP is a harder benchmark to beat than, for example, climatology. However the key reason that we chose the ESP as a benchmark in this study is to "identify whether there is any added value in using Sys4 instead of historical meteorological observations for forecasting the streamflow on seasonal timescales over Europe" (as mentioned on P6 L23-25).

In addition we note that the ESP is used as a benchmark in many seasonal forecasting papers, such as: Bazile et al. (2017), Bell et al. (2017), Candogan Yossef et al. (2017), Crochemore et al. (2016), Meißner et al. (2017) and Mendoza et al. (2017); all from this special issue.

To address the specific suggestions:

i) We will mention the superiority of the ESP to, for example, climatology, in Sections 2.2.4 and 4.1.

This was added on P7 L23-24 and P10 L7-9 of the final manuscript.

ii) We thank the reviewer for pointing this out. The ESP hindcast does not contain the 'perfect' year of meteorological observations as one member of its forcing ensemble. The 'perfect' year was removed to avoid increasing the ESP quality artificially (as the 'perfect' meteorological observations are not available to run the ESP in real-time). This will be clarified in section 2.1.2 on P4 L24-25 with the following addition to the sentence "(i.e. the same as the meteorological observations used to produce

the EFAS-WB, excluding the year of meteorological observations corresponding to the year that is being forecasted)".

This was added on P5 L5-6 of the final manuscript.

**2) I would have liked more discussion of the prospects for improving reliability. Sophisticated statistical methods for calibrating ensemble climate forecasts are available to solve these issues. I would like to hear the authors' views on whether it is viable - both technically and logistically - to apply such methods within the EFAS system.**

This is a very interesting point that we will add to the discussion section of this paper.

A paragraph discussing this point has been added on P14 L11-15 of the final manuscript.

**3) I thought the presentation of Figure 4 could have been improved. It's very difficult to see what the cause of the poor reliability is - bias? incorrect ensemble spread? - when all 74 basins are presented in each panel. I suggest selecting two or three example catchments and presenting only these as case studies of possible causes of poor reliability. For comparing all 74 basins, the reliability from the PIT diagrams can be summarised with Renard et al.'s (2010) alpha index. The alpha index can be converted to a skill score, and could be added in Figure 3.**

We agree with the reviewer that it is difficult to see what the cause of the poor reliability is in Figure 4. As suggested, we will replace this figure with boxplots of the alpha index (converted to a skill score), which will be added to Figure 3. We however believe that selecting a few case studies to show the possible causes of poor reliability goes against the aim of this paper, which is to present an overall image of the performance of the EFAS seasonal streamflow hindcasts. In order to include some general information on the causes of poor reliability, we will perform a visual analysis of all the curves displayed in Figure 4 and add a summary of these causes in the paper. This could for example be summarised in a table, containing the percentage of curves in each category (i.e. narrow forecast, large forecast, under-prediction or over-prediction) for each season and lead times one and seven.

We will update the text in the methods section 2.2 where needed.

Boxplots of the alpha index (converted to a skill score) were created for all seasons and lead times and added to Figure 3. The order of the verification scores in Figure 3 was changed in order to better follow the flow of the results.

The causes for poor reliability (bias and spread) were analysed and quantified using scores introduced by Keller and Hense (2011) for all seasons, regions and lead times. These results are presented in Figure 4, as plots of the percentage of hindcasts (ESP and CM-SSF) falling within each reliability category (reliable, too large, too narrow, under- and over-predicting) for all seasons and lead time one and seven months.

The methods (sections 2.2.3 and 2.2.4), results (section 3.1) and discussion (section 4.1) sections were altered accordingly.

**Specific (minor) comments**

**P1 L25-26 "Unlike forecasts at shorter timescales, they currently do not have skill to predict the exact streamflow at a specific location and time." I would argue that exact forecasts are not possible at shorter timescales either, though of course I agree that there is substantially more uncertainty at seasonal timescales.**

We thank the reviewer for this very good point. The wording is perhaps a bit misleading and we therefore propose the following alteration to this sentence: “Unlike forecasts at shorter timescales, which aim to predict individual events, seasonal streamflow forecasts aim at predicting long-term (i.e. weekly to seasonal) averages.”

This was changed on P1 L30-P2 L1 of the final manuscript.

**P2 L16 "Precipitation variability was however soon identified as a major source of error in the ESP forecasts (Pagano and Garen, 2006), as this forecasting method is based on the assumption that past meteorological events are representative of future events, where each historical year has an equal likelihood of occurrence in the forecast year. As a result, the ESP forecasts are skilful as long as the weather experienced in the current year is not extraordinarily extreme compared to all the historical years of meteorological observations available (Day, 1985)." This is a little misleading, and should be recast. ESP forecasts assume that the vast majority of skill comes from IHC, and thus uses uninformative forcings. In catchments/seasons where precipitation forcings are important, this assumption does not hold, and forecasts may be inaccurate. This is distinct from out-of-sample ("extreme") events, which can occur in any system (whatever the dominant source of skill) and are (usually) difficult to predict.**

We thank the reviewer for this other very good point. We suggest to rephrase this sentence to: “In basins where the meteorological forcings drive the predictability, however, the lack of information on the future climate is a limitation of the ESP forecasting method and might result in unskilful ESP forecasts.”

This was changed on P2 L19-20 of the final manuscript.

**P2 L29 "(Wood et al., 2002 and references therein)" This reference is fine, but quite a lot of work has been done in this area since then and it’s probably worthwhile including a few more recent references.**

We thank the reviewer for pointing this out. We will change this reference to “(Maraun et al., 2010 and references therein)”.

This was changed on P2 L30 of the final manuscript and the reference to the paper by Maraun et al. (2010) was added to the references section of this paper.

**P2 L30 There are also two papers from Greuell et al. (in review) for this special issue on a Europe-wide seasonal streamflow forecasting system.**

We thank the reviewer for pointing this out. We will add a reference to Greuell et al.’s paper on “Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations” here and in the discussion (Section 4.1), where relevant.

Reviewer 2 has requested that unpublished papers ‘in review’ should be removed from the final manuscript. As they are still in review, the papers from Greuell et al. (in review) have therefore not been added to this final manuscript.

**P4 L1 "The Lisflood model was calibrated..." I’d like to hear a little more detail (perhaps a sentence or two) on the calibration method and the periods it was calibrated to. Is this calibration cross-validated?**

We will add the following sentences to the paper: “The calibration was performed from 1994-2002 using the Standard Particle Swarm Optimisation 2011 (SPSO-2011) algorithm. The results were

validated using the Nash-Sutcliffe efficiency for the validation period 2003-2012 (see Zajac et al., 2013 and Smith et al., 2016 for more details)".

This was added on P4 L5-12 of the final manuscript.

**P4 L24 "...randomly resampled..." I'm not clear on how this process is randomised. Do you simply mean 'sampled'?**

The 20 years of historical meteorological observations used for the ESP were indeed simply randomly sampled/selected from the full set of years of historical meteorological observations available (i.e. 25 in total, excluding the 'perfect' year). We will clarify this in the paper by changing "resampled" to "sampled".

This was changed on P5 L4 of the final manuscript.

**P5 L24 "...hence excluding model errors from the analysis." I think this statement is too general, and could probably be removed or softened. Hydrological model errors often vary with magnitude, so different (e.g. biased) forcings can result in different hydrological error characteristics. So errors will not necessarily be 'excluded' though I understand what the authors are getting at: the main difference in forecasts and these 'observations' will be due to the forecast forcings.**

We thank the reviewer for this comment and will change this sentence to: "The EFAS-WB streamflow simulations were used as a proxy for observation against which the seasonal streamflow hindcasts were evaluated, hence minimising the impact of model errors on the hindcasts' quality".

This was changed on P6 L7-9 of the final manuscript.

**P6 L5 "...The sharpness should not be looked at in isolation and should be analysed together with the hindcast accuracy." I would say it's more important to check it against reliability, as sharpness can trade off reliability (e.g. a deterministic forecast is perfectly sharp, but unless it is perfect it is overconfident).**

We will remove this sentence from the paper as we are in any case not looking at any scores in isolation in this paper.

This was removed from the final manuscript.

**P6 L15 "...horizontal [vertical]..." this isn't really a very clear description. Forecasts that are too wide will have something like an s-shape, and forecasts that are too narrow will look something like a transposed s. The authors may like to refer to Laio and Tamea 2007, who describe these shapes in detail, for readers unfamiliar with PIT diagrams.**

We thank the reviewer for sharing the reference to this paper. We will remove the following sentence "A hindcast that is too narrow [wide] will have a horizontal [vertical] PIT diagram." and change it adequately using the explanations from Laio and Tamea (2007).

This was changed on P6 L27-30 of the final manuscript and the reference to the paper by Laio and Tamea (2007) was added to the references section of this paper.

**P14 References. A few of the papers that are listed as 'in review' are now published. Please update these.**

The references will be updated accordingly.

The references were updated in the final manuscript.

**Typos/grammar.** The suggestions will all be incorporated.

These were incorporated.

# RC2

**Major comments**

**1) The manuscript provides a very valuable insights on forecasting capabilities through Europe by comparing of the use of weather forecasts and historical conditions. However the contribution of the paper could be improved if the authors could establish a link between hydrological processes, climatic conditions and seasonal predictability and it is strongly recommended to include an analysis on these topics (eg. CM-SSF predictability skills in snow-dominated regions, arid regions, cold regions, etc.) 3.1 Overall skill of the CM-SSF. In this section almost all the comparison between CM-SSF and ESP is qualitative and general. The authors could further contribute to the forecasting literature, by relating the spatial distribution of skill with physical processes and watershed type. (Eg. Do snow dominant basins shows more predictability than rainfall dominated ones?). Table 1. This is a great contribution, downscale the regional to local scale. As commented before, it is recommended to include dominant physical processes, aridity index or other descriptors to better understand forecast skill of in different hydro climatic regimes. P10 L14-16 The analysis done in this paragraph and link to hypothesis across physical processes is desirable and should be expanded across the manuscript.**

We thank the reviewer for this constructive comment. The main aim of the paper is to give an overall overview of the skill of the EFAS seasonal streamflow hindcasts in Europe, compared to the ESP. In the discussion section of this paper (section 4.1), several hypotheses are made linking the added predictability from Sys4 for forecasting higher or lower streamflows than normal and the hydro-climatic conditions over Europe that could affect this predictability (positively or negatively). These will be clarified and extended in places.

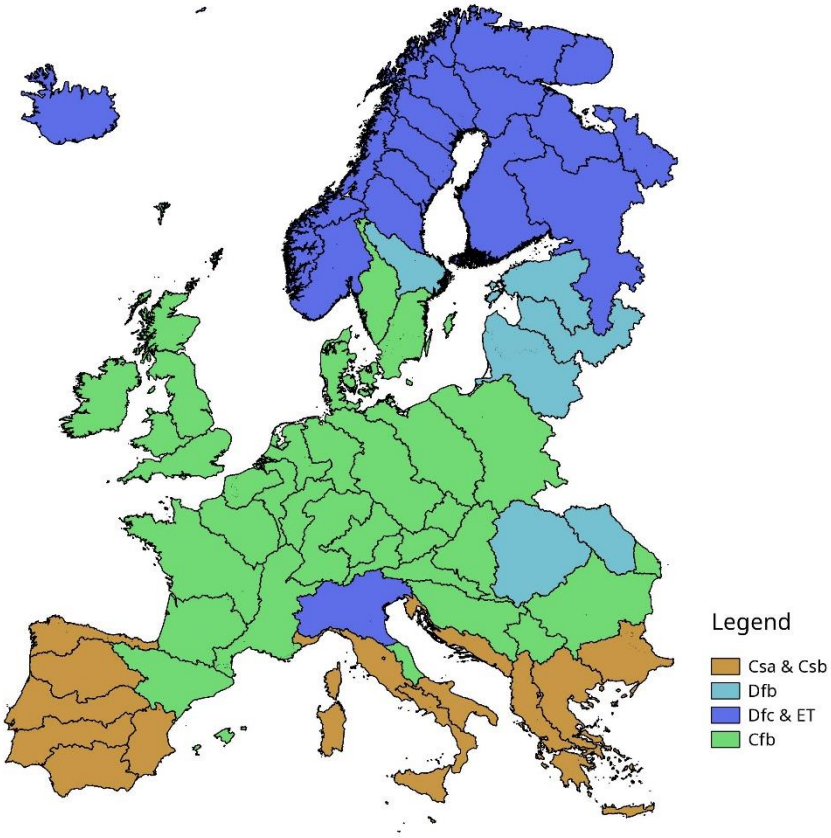
We do agree that further results and discussion along these lines would be a great addition to this paper. In an attempt to explore this, we have split the 74 European regions used for the analysis presented in this paper in 4 climatic zones, according to the well accepted Köppen-Geiger climate classification (see map below). The latter was chosen as there is no commonly accepted hydro-climatic classification over Europe we could use. The 4 climatic zones into which the 74 regions were split are:

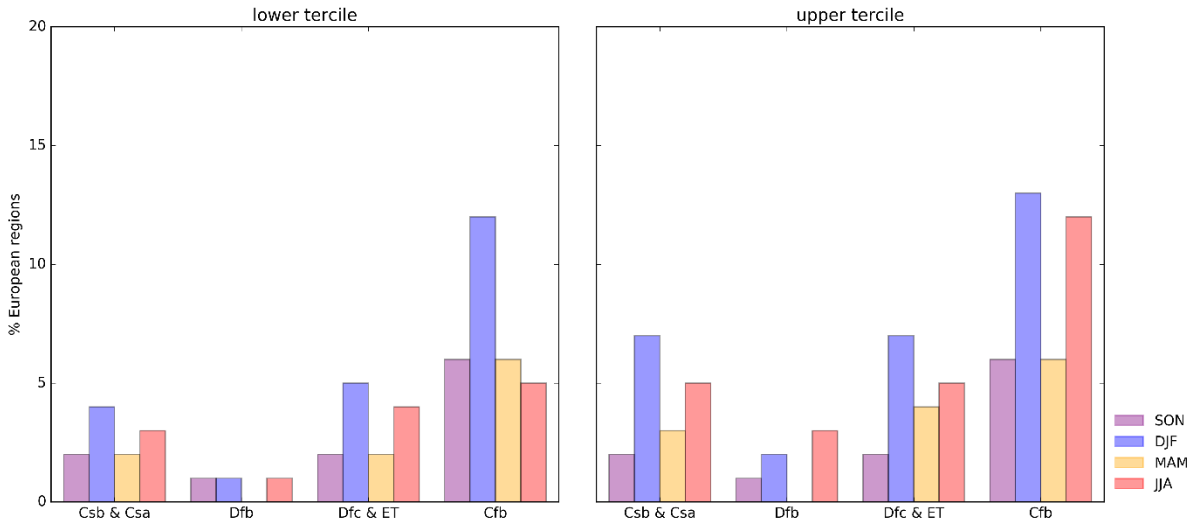
- Csa & Csb: warm temperate, dry and warm to hot summers
- Dfb: snow, fully humid and warm summers
- Dfc & ET: snow, fully humid and cool summers & polar, polar tundra
- Cfb: warm temperate, fully humid and warm summers

The bar chart shown below presents the percentage of European regions within each climate zone (x-axes) for which the CM-SSF is the best system at predicting anomalously low or high streamflows (lower and higher terciles in the left-hand and right-hand plots, respectively; in terms of the ROC). The results are shown for each season (SON, DJF, MAM and JJA; as indicated by the legend).

This analysis shows two main results with regards to the link between the CM-SSF predictability and the climate characteristics of the regions. More specifically, for both terciles and all seasons:

- The largest CM-SSF predictability can be found for regions in the Cfb zone, especially in winter (and summer for the upper tercile).
- The lowest CM-SSF predictability can be found for regions in the Dfb zone, especially in spring (the snowmelt season).




However, these results present several limitations. First, the regions used in the analysis presented in this paper are quite large. As a result, their classification in hydro-climatic zones is likely to be biased as the climate and hydrological (to a large extent) characteristics might vary greatly within a single region. Moreover, the Köppen-Geiger classification does not really capture the hydrological characteristics (i.e. infiltration, importance of groundwater reservoir, etc) of the land surface and is

more representative of the climate. Finally, the number of regions within each climate zone is highly variable and could have influenced the results shown in the bar chart.

Discussing more extensively the link between the CM-SSF predictability and the hydro-climatic characteristics of the regions from the results presented in this paper, as recommended by the editor, would require further analysis. However, as shown by the short analysis performed and described above, significant results would require a longer and more in depth analysis (i.e. by looking at the predictability for smaller river basins, by deriving basin hydro-meteorological indices and linking those with the CM-SSF predictability in those smaller basins), which we believe goes beyond the scope of this paper (i.e. to give an overall overview of the quality of the EFAS seasonal streamflow hindcasts in Europe). Therefore, we suggest to exclude any further analysis from the final manuscript and to discuss this point as an opportunity for future work in the final manuscript.

The hypotheses were clarified and extended on P10 L19 and P10 L34 – P11 L3, and the wider point was discussed on P14 L21-30 of the final manuscript.

**2) As the study region is very large, a quantitative comparison through the seasons and regions is a hard task. It is necessary to include some numbers in the result and discussion sections, perhaps there is a link between different hydro climatic regions, ungauged regions, etc. and seasonal forecast skill. P9 L21 "The CM-SSF is more skilful in many at predicting anomalously low and high streamflows than ESP in certain season and regions". It is very difficult to qualitatively judge the skill of the predictions, but at least in terms of geographic regions, the authors could add quantitative indicators, e. g. 60% of the analyzed regions shows better performance in CM-SSF predictions rather than ESP forecast. P10 L8 Same comment as before.**

We believe that this will be a great addition to the paper and will accompany the qualitative sentences in the abstract, results, discussion and conclusion sections of the paper with quantitative indicators such as the example given by the reviewer.

Following comments from the other reviewer of this paper, more quantitative results were added about the hindcasts' reliability (section 3.1). Quantitative indicators were additionally added to the abstract, discussion (P10 L18, L29 and P11 L6) and conclusion sections of the final manuscript.

**Specific (minor) comments**

**Abstract. The abstract needs to clarify the scientific questions and give a more quantitative assessment for the forecasting comparison (not only lead time). It is strongly suggested to include a brief methodology, materials, datasets and methods. The name of the hydrological model, should be included in the abstract.**

We will add research questions, mention the attributes of the hindcasts covered by the verification scores used for the analysis and the model name to the abstract.

This information was added to the abstract for the final manuscript.

**1. Introduction, general comments. The scientific question is clear, the literature review is good but focused on European regions. It would be very nice to include studies in other regions. (eg. Mendoza et al. 2014 in the Andes and Seibert et al. 2017 in southern Africa).**

In order to limit the length of the already extensive literature review and because this paper looks at the quality of the EFAS seasonal streamflow hindcasts over Europe (hence for discussion purposes), we have decided to focus the literature review on European studies. There are however references to key seasonal hydrological forecasting papers outside of Europe in the introduction of this paper.

Please see P2 L31-33 of the final manuscript for an introduction to climate-model-based seasonal streamflow forecasting experiments outside of Europe.

**P1 L28 Are these large-scale climatic patterns the only predictors for seasonal climatic conditions in Europe?**

The large-scale climate patterns mentioned here are indeed only a subset of the patterns that are currently studied and used for predicting the climate on seasonal timescales globally. We have decided to constrain our list to a few patterns that recurrently appeared in literature about seasonal hydro-meteorological forecasting.

**P2 L12 How much is high to measure forecast quality? The response time for IHC is one of the terms that modulate the predictability of flows but also the storage capacity of watersheds in different hydrological processes (eg. Glacier, sub-surface process).**

We agree with the reviewer that qualifying forecast quality of "high" is qualitative. We will improve this sentence by stating that the exact forecast quality depends on the time of year, the type and location of the basin examined.

This was added on P2 L17-18 of the final manuscript.

The storage capacity of watersheds is indeed another modulator of the flow predictability. This is however implied in the sentence "The quality of the ESP forecasts can be high in basins where the IHC dominate the surface hydrological cycle for several months" on P2 L16-17 of the final manuscript. Indeed, in basins where the storage capacity is high, the IHC are expected to have a longer impact on the output streamflow.

**P3 L15 Are there any efforts or initiatives to improve communication and outreach?**

This is a very interesting question which we have touched on slightly in the discussion section of the paper (P13 L4-9 of the final manuscript). We will however mention another international initiative, HEPEX, which has for more than a decade engaged in communication and outreach of the use of ensemble hydro-meteorological prediction for decision-making in water-related applications.

This was added on P13 L9-12 of the final manuscript.

**P3 L27 I strongly suggest to improve this paragraph by including information on spatial distribution, time step for modelling, quantity and quality of the data, etc.**

Information on spatial distribution, time step for modelling and quantity of the data is already given in sections 2.1.1 to 2.2. We will however add information on the quality of the Lisflood simulations to section 2.1.1.

This was added on P4 L9-12 of the final manuscript.

We will additionally add a sentence here clarifying that information about the data used in this paper is given below.

This was added on P3 L29-30 of the final manuscript.

**P3 L29 Please include a citation for the Lisflood model.**

The two main publications about the Lisflood model are mentioned on P4 L4 of the final manuscript and can be found in the references of this paper.

**P4 L1 Please explain what hydrological processes were calibrated.**

A list of the Lisflood hydrological processes that were calibrated will be given in section 2.1.1.

This was added on P4 L6-9 of the final manuscript.

**P4 L3-14 This paragraph should be moved to section 2.1. More detail should be given to EFAS-WB (i.e., references, hydrological processes reproduced, model uncertainties, etc.) Please include forecast quality indices whenever EFAS-WB is considered the best estimate of hydrological state.**

We believe that the current structure of section 2 is adequate and gives a nice flow to the paper. We have however added a sentence on P3 L29-30 to clarify that information about the data used in this paper is given in the following sections. More details about the hydrological processes represented within Lisflood have been given on P4 L6-9 of the final manuscript and will be referred to here (P4 L15). Information about the Lisflood simulation quality (in validation) was added on P4 L9-12.

**P5 L15 Did you assess of sub-monthly predictability of streamflow? Are the time scales considered enough? Is that time step enough for decision makers?**

The sub-monthly predictability of streamflow was not assessed for this paper. Monthly streamflow aggregations were chosen here in order to give an overview of the degradation of skill over the full seven months of lead time of the forecast. Monthly flow aggregations are valuable to decision-makers for many applications of the water sector. Indeed, in most papers cited on P3 L14-16 of the final manuscript, the authors have looked at monthly flow aggregations, with a few authors looking at three-monthly aggregations.

The use of monthly streamflow aggregations in this paper is justified on P5 L32 – P6 L3 of the final manuscript.

**P5 L22 Please explain if the performance measures of Crochemore et al. (2016) are the ones described in the next numerals.**

Here, we will clarify exactly which verification scores used in this paper were also used in Crochemore et al. (2016) and mention that these verification scores are described below.

This was changed on P6 L5-7 of the final manuscript.

**Sections 2.2.1 to 2.2.5 should be addressed in a Figure as a resume scheme for forecast skill.**

We believe that sections 2.2.1 to 2.2.5 are well explained and we do not see the added value of adding an additional figure to this paper to summarise their content.

**P8 L31 Did the authors find specific regions where the predictability of extreme years was better/worse? If that was the case, can you please provide an explanation?**

This is answered in more detail in what follows of section 3.2 and is discussed in section 4.1.

**P9 L1 For most seasons (and regions?)**

Thank you for pointing this out. It will be added here.

This was added on P9 L24 of the final manuscript.

**P9 L11 In my opinion, Meißner et al. (in review) should not be cited if there is not an accessible reference. I suggest including a DOI, or delete the reference.**

This paper is now published and the reference and corresponding in-text citations will be updated accordingly.

This was changed within the final manuscript.

**P11 L23 Neumann et al. (submitted to J. Hydrometeorol.), same comment than Meißner et al. (in review).**

This paper being currently in review, it was removed from the references.

**P13 L3-5 "The impact of this evaluation strategy in this paper should be minimal,.." But how does it impact low flows or drought predictability?**

Please note that this evaluation strategy was not used for calculating the ROC, as mentioned on P13 L27 of the final manuscript. For the latter, the full CM-SSF ensemble was used, hence not impacting low flow predictability assessed in this paper.

**P13 L11-14 Statistical or probabilistic approaches (eg. Han and Coulibaly, 2017; Mendoza et al. 2017), should be discussed. Future work could include a different comparison, merging climate forecast with other predictors.**

The comparison of the CM-SSF to statistical seasonal streamflow forecasting approaches will be mentioned in the discussion section of this paper.

A paragraph discussing this was added on P14 L16-20 of the final manuscript.

# Skilful seasonal forecasts of streamflow over Europe?

Louise Arnal[1,2], Hannah L. Cloke[1,3,4,5], Elisabeth Stephens[1], Fredrik Wetterhall[2], Christel Prudhomme[2,6,7], Jessica Neumann[1], Blazej Krzeminski[2] and Florian Pappenberger[2]

[1]Department of Geography and Environmental Science, University of Reading, RG6 6AB, United Kingdom
[2]European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG6 9AX, United Kingdom
[3]Department of Meteorology, University of Reading, RG6 6BB, United Kingdom
[4]Department of Earth Sciences, Uppsala University, Uppsala, SE-752 36, Sweden
[5]Centre of Natural Hazards and Disaster Science, CNDS, Uppsala, SE-752 36, Sweden
[6]Department of Geography, Loughborough University, Loughborough, LE11 3TU, United Kingdom
[7]NERC Centre for Ecology & Hydrology, Wallingford, OX10 8BB, United Kingdom

*Correspondence to*: Louise Arnal (l.l.s.arnal@pgr.reading.ac.uk; louise.arnal@ecmwf.int)

**Abstract.** This paper considers whether there is any added value in using seasonal climate forecasts instead of historical meteorological observations for forecasting streamflow on seasonal timescales over Europe. ~~This paper presents a~~A Europe-wide analysis of the skill of the newly operational EFAS (European Flood Awareness System) seasonal streamflow forecasts (produced by forcing the Lisflood model with the ECMWF System 4 seasonal climate forecasts), benchmarked against the Ensemble Streamflow Prediction (ESP) forecasting approach (produced by forcing the Lisflood model with historical meteorological observations), is undertaken. The results suggest that, on average, the System 4 seasonal climate forecasts improve the streamflow predictability over historical meteorological observations for the first month of lead time only (in terms of hindcast accuracy, sharpness and overall performance). However, the predictability varies in space and time and is greater in winter and autumn. Parts of Europe additionally exhibit a longer predictability, up to seven months of lead time, for certain months within a season. In terms of hindcast reliability, the EFAS seasonal streamflow hindcasts are on average less skilful than the ESP for all lead times. The results also highlight the potential usefulness of the EFAS seasonal streamflow forecasts for decision-making (measured in terms of the hindcast discrimination for the lower and upper terciles of the simulated streamflow). Although the ESP is the most potentially useful forecasting approach in Europe, the EFAS seasonal streamflow forecasts appear more potentially useful than the ESP in some regions and for certain seasons, especially in winter for almost 40% ~~most~~ of Europe. Patterns in the EFAS seasonal streamflow hindcasts skill are however not mirrored in the System 4 seasonal climate hindcasts, hinting the need for a better understanding of the link between hydrological and meteorological variables on seasonal timescales, with the aim to improve climate-model based seasonal streamflow forecasting.

## 1 Introduction

Seasonal streamflow forecasts predict the likelihood of a difference from normal conditions in the following months. Unlike forecasts at shorter timescales, which aim to predict individual events, seasonal streamflow forecasts aim at predicting long-term (i.e. weekly to seasonal) averages~~they currently do not have skill to predict the exact streamflow at a specific location and time~~. The predictability in seasonal streamflow forecasts is driven by two components of the Earth system, the initial hydrological conditions (IHC; i.e. of snowpack, soil moisture, streamflow and reservoir levels, etc.) and large-scale climate patterns, such as the El Niño-Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), the Pacific-North American (PNA) pattern and the Indian Ocean Dipole (IOD) (Yuan et al., 2015b).

The first seasonal streamflow forecasting method, based on a regression technique developed around 1910-11 in the United States, harnessed the predictability from accurate IHC of snowpack to derive streamflow for the following summer (Church, 1935). This statistical method recognised antecedent hydrological conditions and land surface memory~~,~~ as key drivers of streamflow generation for the following months.

Alongside the physical understanding of streamflow generation processes came technical developments, such as the creation of the first hydrological models and the acquisition of longer observed meteorological time series, which led to the creation of the first operational model-based seasonal streamflow forecasting system. This system, called Extended Streamflow Prediction (ESP; i.e. note that ESP nowadays stands for Ensemble Streamflow Prediction, although it refers to the same forecasting method), was developed by the United States National Weather Service (NWS) in the 1970s (Twedt et al., 1977; Day, 1985). The ESP forecasts are produced by forcing a hydrological model, initialised with the current IHC, with the observed historical meteorological time series available. The output is an ensemble streamflow forecast (where each year of historical data is a streamflow trace) for the following season(s) (Twedt et al., 1977; Day, 1985). The quality of the ESP forecasts can be high in basins where the IHC dominate the surface hydrological cycle for several months (the exact forecast quality depending on the time of year, the type and location of the basin; Wood and Lettenmaier, 2008).

In basins where the meteorological forcings drive the predictability, however, the lack of information on the future climate is a limitation of the ESP forecasting method and might result in unskilful ESP forecasts.~~Precipitation variability was however soon identified as a major source of error in the ESP forecasts (Pagano and Garen, 2006), as this forecasting method is based on the assumption that past meteorological events are representative of future events, where each historical year has an equal likelihood of occurrence in the forecast year. As a result, the ESP forecasts are skilful as long as the weather experienced in the current year is not extraordinarily extreme compared to all the historical years of meteorological observations available (Day, 1985).~~ This drawback led to the investigation of the use of seasonal climate forecasts, in place of the historical meteorological inputs, to feed hydrological models and extend the predictability of hydrological variables on seasonal timescales (Pagano and Garen, 2006). This investigation was made possible by technical and scientific advances. Scientifically, seasonal climate forecasts were improved greatly by the understanding of ocean-atmosphere-land interactions and the identification of large-scale climate patterns as drivers of the hydro-meteorological predictability (Goddard et al.,

2001; Troccoli, 2010). This was technically implementable with the increase of computing resources, making it possible to run dynamical coupled ocean-atmosphere-land general circulation models on the global scale at high spatial and temporal resolutions (Doblas-Reyes et al., 2013). An additional technical challenge, the coarse spatial resolution of seasonal climate forecasts compared to the finer resolution of hydrological models, had to be addressed. To tackle this issue, many authors have explored different ways of downscaling climate variables for hydrological applications (Maraun et al., 2010~~Wood et al., 2002~~ and references therein).

While climate-model-based seasonal streamflow forecasting experiments are more common outside of Europe, for example for the United States (Wood et al., 2002; 2005; Mo and Lettenmaier, 2014), Australia (Bennett et al., 2016), Africa (Yuan et al., 2013), they remain limited in Europe, with a few examples in France (Céron et al., 2010; Singla et al., 2012; Crochemore et al., 2016), in Central Europe (Demirel et al., 2015; Meißner et al., ~~in review~~2017), in the United Kingdom (Bell et al., 2017; Prudhomme et al., ~~in review~~2017) and at the global scale (Yuan et al., 2015a; Candogan Yossef et al., 2017). This is because, although the quality of seasonal climate forecasts has increased over the past decades, there remains limited skill in seasonal climate forecasts for the extra-tropics, particularly for the variables of interest for hydrology, notably precipitation and temperature (Arribas et al., 2010; Doblas-Reyes et al., 2013).

In Europe, the NAO is one of the strongest predictability sources of seasonal climate forecasts; it is associated with changes in the surface westerlies over the North Atlantic and Europe, and hence with changes in temperature and precipitation patterns over Europe (Hurrell, 1995; Hurrell and Van Loon, 1997). It was shown to affect streamflow predictability, especially during winter (Dettinger and Diaz, 2000; Bierkens and van Beek, 2009; Steirou et al, 2017), additionally to the IHC and the land surface memory. It was ~~additionally~~furthermore shown to be an indicator of flood damage and occurrence in parts of Europe (Guimarães Nobre et al., 2017).

As the quality and usefulness of seasonal streamflow forecasts increases, their usability for decision-making has lagged~~s~~ behind. Translating the quality of a forecast into an added value for decision-making and incorporating new forecasting products into established decision-making chains are not easy tasks. This has been explored for many water-related applications ~~of the water sector~~, such as navigation (Meißner et al., ~~in review~~2017), reservoir management (Viel et al., 2016; Turner et al., ~~in review~~2017), drought-risk management (Sheffield et al., 2013; Yuan et al., 2013; Crochemore et al., 2017), irrigation (Chiew et al., 2003; Li et al., ~~in review~~2017), water resources management (Schepen et al., 2016) and hydropower (Hamlet et al., 2002); but seasonal streamflow forecasts~~it has not been~~ have yet to be adopted by the flood preparedness community.

The European Flood Awareness System (EFAS) is at the forefront of seasonal streamflow forecasting, with one of the first operational pan-European seasonal hydrological forecasting systems. The aim of this paper is to bridge the current gap in pan-European climate-model-based seasonal streamflow forecasting studies. Firstly, the setup of the newly operational EFAS climate-based seasonal streamflow forecasting system is presented. A Europe-wide analysis of the skill of this forecasting system compared to the ESP forecasting approach is then presented, in order to identify whether there is any added value in using seasonal climate forecasts instead of historical meteorological observations for forecasting streamflow

on seasonal timescales over Europe. Subsequently, the potential usefulness of the EFAS seasonal streamflow forecasts for decision-making is assessed.

## 2 Data and methods

### 2.1 EFAS hydrological simulation and seasonal hindcasts

The data used in this paper include a streamflow simulation and two seasonal streamflow hindcasts (Fig. 1). Further information on these datasets is given below.

#### 2.1.1 Hydrological modelling and streamflow simulation

The Lisflood model was used to produce all the simulation and hindcasts used in this paper. Lisflood is a GIS-based hydrological rainfall-runoff-routing distributed model written in the PCRaster Dynamic Modelling Language, which enables it to use spatially distributed maps (i.e. both static and dynamic) as input (De Roo et al., 2000; Van Der Knijff et al., 2010). The Lisflood model was calibrated to produce pan-European parameter maps. The calibration was performed from 1994-2002 using the Standard Particle Swarm Optimisation 2011 (SPSO-2011) algorithm. The calibration was carried out for parameters controlling: snowmelt, infiltration, preferential bypass flow through the soil matrix, percolation to the lower groundwater zone, percolation to deeper groundwater zones, residence times in the soil and subsurface reservoirs, river routing and reservoir operations for a few basins. The results were validated with the Nash-Sutcliffe efficiency (NSE) for the validation period 2003-2012. In validation, Lisflood was shown to have explanatory power (i.e. NSE $\geq$ 0) for 90% of the basins. Basins with large discrepancies between the observed and simulated flow statistics were situated mainly on the Iberian Peninsula and on the Baltic coasts (see Zajac et al., 2013 and Smith et al., 2016 for further details).

The Lisflood model is run operationally in EFAS, with the simulation domain covering Europe at a 5 x 5 km resolution. A reference simulation, called the EFAS water balance (EFAS-WB), is available on a daily time step starting from February 1990. Lisflood simulates the hydrological processes within a basin (most of which are mentioned above), starting from the previous day IHC (e.g. snow cover, storage in the upper and lower zones, soil moisture, initial streamflow, reservoir filling) and forced with the most recent observed meteorological fields (i.e. of precipitation, potential evapotranspiration and temperature; provided by the EFAS meteorological data collection centres). The observed meteorological fields are daily maps of spatially interpolated point measurements of precipitation (from more than 6000 stations) and temperature (from more than 4000 stations) at the surface level. These same data are used to produce interpolated potential evapotranspiration maps from the Penman–Monteith method (Alfieri et al., 2014). All meteorological variables are interpolated on a 5 x 5 km grid using an inverse distance weighting scheme and the temperature is first corrected using the elevation (Smith et al., 2016).

The EFAS-WB is the best estimate of the hydrological state at a given time and for a given grid point in EFAS and is thus used as initial conditions from which the seasonal hydrological forecasts are started.

### 2.1.2 Ensemble seasonal streamflow hindcasts

In this paper, two types of ensemble seasonal streamflow hindcasts are used: the Ensemble Streamflow Prediction (ESP) hindcast (hereafter referred to as ESP) and the System 4-driven seasonal streamflow hindcast [hereafter referred to as CM-SSF (climate-model-based seasonal streamflow forecast), following the notation from Yuan et al. (2015b)].

They are both initialised from the EFAS-WB, on the first day of each month, to produce a new ensemble streamflow forecast up to a lead time of seven months (215 days), with a daily time step. Both hindcasts are generated from February 1990 for the same European domain as the EFAS-WB, at the same 5 x 5 km resolution. The unique difference between the ESP and the CM-SSF is the meteorological forcing used to drive the hydrological model, described below.

The ESP is produced by driving the Lisflood model with 20 (the number of years of data available at the time the hindcast was produced) randomly ~~re~~sampled years of historical meteorological observations (i.e. the same as the meteorological observations used to produce the EFAS-WB, excluding the year of meteorological observations corresponding to the year that is being forecasted). A new 20-member ESP is thus generated at the beginning of each month and for the next seven months.

The CM-SSF is produced by driving the Lisflood model with the ECMWF System 4 seasonal climate hindcast (Sys4; i.e. of precipitation, evaporation and temperature). Sys4 has a spatial horizontal resolution of about 0.7 degrees (approximately 70 km). It is re-gridded to the Lisflood spatial resolution using an inverse distance weighting scheme and the temperature is first corrected using the elevation. Sys4 is made of 15 ensemble members, extended to 51 every three months (Molteni et al., 2011). From 2011 onwards the Sys4 forecasts were run in real time and all contained 51 ensemble members. A new 15 to 51-member CM-SSF is hence produced at the beginning of each month and for the next seven months. Operationally, the CM-SSF forecasts are currently used in EFAS to generate a seasonal streamflow outlook for Europe at the beginning of every month.

## 2.2 Hindcast evaluation strategy

For this study, monthly region specific discharge averages of the hindcasts (CM-SSF and ESP) and EFAS-WB were used. The specific discharge is the discharge per unit area of an upstream basin. For this paper, the gridded daily specific discharge was calculated by dividing the gridded daily discharge output maps (of the hindcasts and the EFAS-WB) by the Lisflood gridded upstream area static map. Subsequently, the gridded daily specific discharge maps were used to calculate daily region averaged specific discharges (for each region in Fig. 2) by summing up the daily specific discharge values of each grid cell within a region, divided by the number of grid cells in that region. Finally, monthly specific discharge region averages were calculated for each calendar month.

The regions displayed in Fig. 2 were created by merging several basins together (basins used operationally in EFAS for the shorter timescales forecasts), while respecting hydro-climatic boundaries. They were chosen for the analysis presented in this

paper for two main reasons. Firstly, they are the regions used operationally to display the EFAS seasonal streamflow outlook. Secondly, they were created in order to capture large-scale variability in the weather.

The analysis of the hindcasts was performed on monthly specific discharge (hereafter referred to as streamflow) region averages for hindcast starting dates spanning February 1990 to November 2016 (included; approximately 27 years of data), with one to seven months of lead time. In this paper, one month of lead time refers to the first month of the forecast (e.g. the January 2017 streamflow for a forecast made on the 1st ~~first~~ of January 2017). Two months of lead time is the second month of the forecast (e.g. the February 2017 streamflow for a forecast made on the 1st ~~first~~ of January 2017), etc. Monthly averages were selected for the analysis presented in this paper as it is a valuable aggregation time step for decision-makers for many water-related applications [as shown in the literature for applications such as, for example, navigation (Meißner et al., 2017), reservoir management (Viel et al., 2016; Turner et al., 2017), drought-risk management (Yuan et al., 2013), irrigation (Chiew et al., 2003; Li et al., 2017) and hydropower (Hamlet et al., 2002)].

Several verification scores were selected in order to assess the hindcasts' quality. These verification scores were chosen to cover a wide range of hindcast attributes (i.e. accuracy, sharpness, reliability, overall performance and discrimination). All ~~Most~~ of these verification scores, except for the verification score selected to look at hindcast discrimination, are the same as chosen in Crochemore et al. (2016), and are described below~~, with an additional verification score selected to look at hindcast discrimination~~. The EFAS-WB streamflow simulations were used as a proxy for observation against which the seasonal streamflow hindcasts were evaluated, hence minimising the impact of model errors on the hindcasts' quality~~excluding model errors from the analysis~~.

### 2.2.1 Hindcast accuracy

Both hindcasts (CM-SSF and ESP) were assessed in terms of their accuracy; the magnitude of the errors between the hindcast ensemble mean and the 'truth' (i.e. the EFAS-WB). For this purpose, the mean absolute error (MAE) was calculated for each region, target month (i.e. the month that is being forecast) and lead time (i.e. one to seven months). The lower the MAE, the more accurate the hindcast.

### 2.2.2 Hindcast sharpness

Both hindcasts were also assessed in terms of their sharpness; an attribute of the hindcast only, which is a measure of the spread of the ensemble members of a hindcast. In this paper, the 90% interquantile range (IQR; ~~) (~~i.e. the difference between the 95$^{th}$ and the 5$^{th}$ percentiles of the hindcast distribution) was calculated for each region, target month and lead time. The lower the IQR, the sharper the hindcast. ~~The sharpness should not be looked at in isolation and should be analysed together with the hindcast accuracy.~~

### 2.2.3 Hindcast reliability

Both hindcasts were additionally assessed in terms of their reliability; the statistical consistency between the hindcast probabilities and the observed frequencies. For this purpose, the probability integral transform (PIT) diagram was calculated for each region, target month and lead time (Gneiting et al., 2007). The PIT diagram is the cumulative distribution of the PIT values as a function of the PIT values. The PIT values measure where the 'truth' (i.e. EFAS-WB) falls relative to the percentiles of the hindcast distribution. For a perfectly reliable hindcast, the 'truth' should fall uniformly in each percentile of the hindcast distribution, giving a PIT diagram that falls exactly on the 1 to 1 diagonal. A hindcast that systematically under- [over-] predicts~~estimates~~ the 'truth' will have a PIT diagram below [above] the diagonal~~, outside of the ±0.1 tolerance bands from the 1 to 1 diagonal~~. A hindcast that is too narrow (i.e. underdispersive; hindcast distribution smaller than the distribution of the observations) [large~~wide~~ (i.e. overdispersive; hindcast distribution greater than the distribution of the observations)] will have a transposed S-shaped [S-shaped]~~horizontal [vertical]~~ PIT diagram (Laio and Tamea, 2007).

In order to compare the reliability across all regions, target months and lead times, the area between the PIT diagram and the 1 to 1 diagonal was computed for all PIT diagrams (Renard et al., 2010). The smaller this area, the more reliable the hindcast.

Furthermore, to disentangle the causes for poor reliability, the spread and bias of the hindcasts were calculated for all PIT diagrams, using two measures first introduced by Keller and Hense (2011): ß-score and ß-bias, respectively. By definition, a perfectly reliable hindcast (with regards to its spread) will have a β-score of zero (to which a tolerance interval of ±0.09 was added), whereas a hindcast that is too narrow [large] will have a negative [positive] β-score (outside of the tolerance interval). A perfectly reliable hindcast (with regards to its bias) will have a β-bias of zero (to which a tolerance interval of ±0.09 was added), whereas a hindcast that systematically under- [over-] predicts the 'truth' will have a negative [positive] β-bias (outside of the tolerance interval).

~~The PIT diagram presents the hindcast reliability differently from a reliability diagram, in the sense that the observed frequency is not plotted on a PIT diagram.~~

### 2.2.4 Hindcast overall performance

The hindcasts were furthermore assessed in terms of their overall performance from the continuous rank probability score (CRPS), calculated for each region, target month and lead time (Hersbach, 2000). The CRPS is a measure of the difference between the hindcast and the observed (i.e. EFAS-WB) cumulative distribution functions. The lower the CRPS, the better the overall performance of the hindcast.

In this paper, the skill of the CM-SSF is benchmarked with respect to the ESP in order to identify whether there is any added value in using Sys4 instead of historical meteorological observations for forecasting the streamflow on seasonal timescales over Europe. To this end, skill scores were calculated for the MAE, IQR, PIT diagram area and CRPS, using the following equation:

$$Skill\ score = 1 - \frac{score_{CM-SSF}}{score_{ESP}} \quad (1)$$

Skill scores were calculated for each region, target month and lead time and will be referred to as: MAESS, IQRSS, PITSS and CRPSS, respectively. Skill scores larger [smaller] than zero indicate more [less] skill in the CM-SSF compared to the ESP. A skill score of zero means that the CM-SSF is as skilful as the ESP. Note that as the ESP is not a 'naive' forecast, using it as a benchmark might lead to lower skill than benchmarking the CM-SSF against, for example, climatology.

#### 2.2.5 Hindcast potential usefulness

For decision-making, the ability of a seasonal forecasting system to predict the right category of an event (e.g. above or below normal conditions) months ahead is of great importance (Gobena and Gan, 2010). In this paper, the potential usefulness of the CM-SSF and the ESP to forecast lower and higher than normal streamflow conditions within their hindcasts is assessed.

To do so, the relative operating characteristic (ROC) score, a measure of hindcast discrimination (Mason and Graham, 1999), was calculated. The thresholds selected to calculate the ROC are the lower and upper terciles of the EFAS-WB climatology for each season. They were calculated for the simulation period (February 1990 to May 2017), by grouping together EFAS-WB monthly streamflows for each month falling in a season (SON: September-October-November, DJF: December-January-February, MAM: March-April-May and JJA: June-July-August). For each season and each region a lower and upper tercile streamflow value was obtained, subsequently used as thresholds against which to calculate the probability of detection (POD) and the false alarm rate (FAR; ) (with 0.1 probability bins) for both hindcasts, for each region, season and lead time. Finally, the area under the ROC curve, i.e. the ROC score, was calculated for both hindcasts, for each region, season and lead time. The ROC score ranges from 0 to 1, with a perfect score of 1. A hindcast with a ROC score $\leq 0.5$ is unskilful, i.e. less good than the long term average climatology which has a ROC of 0.5, and therefore not useful.

Because the ROC score was calculated from a low number of events [i.e. approximately 27 years × 3 months in each season × 1/3 (lower or upper tercile) = 27 simulated events], the hindcasts were judged skilful and useful when their ROC score $\geq$ 0.6 instead of 0.5. Moreover, the CM-SSF was categorised as more useful than the ESP when the CM-SSF's ROC score was at least 10% larger than the ESP's ROC score.

## 3 Results

### 3.1 Overall skill of the CM-SSF

In the first part of the results, the skill of the CM-SSF (benchmarked with respect to the ESP) is presented, in terms of the accuracy (MAESS), sharpness (IQRSS), reliability (PITSS diagrams) and overall performance (CRPSS) in the hindcast

datasets. This will benchmark the added value of using Sys4 against the use of historical meteorological observations for forecasting the streamflow on seasonal timescales over Europe.

As shown by the MAESS boxplots (Fig. 3), the CM-SSF appears on average more accurate than the ESP for the first month of lead time only, for all seasons excluding spring (MAM). Beyond one month of lead time, the CM-SSF becomes on average as or less accurate than the ESP. There are however noticeable differences between the different seasons. The CM-SSF shows the largest improvements in the average accuracy compared to the ESP in winter (DJF) and for the first month of lead time. For ~~higher~~ longer lead times (i.e. two to seven months), the accuracy of the CM-SSF is on average quite similar to that of the ESP in autumn (SON) and winter, and on average lower in spring and summer (JJA). ~~The boxplots for the autumn and winter are smaller than for the spring and summer, which hints a smaller variability in the MAESS amongst regions and target months in autumn and winter compared to the spring and summer.~~ The boxplots for the CRPSS look very similar to the MAESS boxplots, the main difference being the lower average scores for two to seven months of lead time in autumn and winter (Fig. 3).

The boxplots of the IQRSS show that the CM-SSF predictions are on average as sharp as those of the ESP for the first month of lead time (slightly sharper in autumn; Fig. 3). For two to seven months of lead time, in autumn and winter, the CM-SSF predictions are on average sharper than those of the ESP, whereas in spring and summer, the CM-SSF predictions are on average slightly less sharp than the ESP predictions. ~~As for the MAESS, the boxplots of the IQRSS for the autumn and winter are slightly smaller than for the spring and summer, hinting a smaller variability in the IQRSS amongst regions and target months in autumn and winter than in spring and summer~~

As shown by the boxplots of the PITSS (Fig. 3), the CM-SSF predictions are less reliable than the ESP prediction for all seasons and months of lead time. For the first month of lead time and all seasons, 10-20% of the ESP hindcasts and less than 5% of the CM-SSF hindcasts are reliable (Fig. 4). 40-60% of the ESP hindcasts are not reliable for the first month of lead time and all seasons due to the ensemble spread. Approximately half of these hindcasts are too large, while the other half (slightly more in autumn and winter) is too narrow. 50-80% of the ESP hindcasts furthermore under-predict the simulated streamflow for the first month of lead time and all seasons. The percentage of reliable [unreliable] ESP hindcasts increases [decreases] with lead time, as the effect of the IHC fades away. 70-90% of the CM-SSF hindcasts are too narrow for the first month of lead time and all seasons. With increasing lead time, the percentage of CM-SSF hindcasts that are too narrow [large] decreases [increases], especially in spring. 40-50% of the CM-SSF hindcasts over-predict the simulated streamflow in spring and summer for the first month of lead time (and increasingly over-predict with longer lead times). In autumn and winter, about 70% of the CM-SSF hindcasts under-predict the simulated streamflow for the first month of lead time (and increasingly under-predict with longer lead times).

For all verification scores, the boxplots for autumn and winter are slightly smaller than for spring and summer, hinting a smaller variability in the verification scores amongst regions and target months in autumn and winter than in spring and summer. Furthermore~~Overall~~, the presence of the boxplots above the zero line (i.e. no skill line) for all lead times suggests that the CM-SSF is more skilful than the ESP for some regions and target months, beyond the first month of lead time.

~~As shown in the ESP PIT diagrams (Fig. 4), lines are concentrated around the diagonal, within the tolerance bands, for all seasons and both lead times (i.e. one and seven months; this was also observed for two to six months of lead time, not shown). This signifies that the ESP is mostly reliable, with the exception of a few cases where it is under predictive for the first month of lead time in winter and summer. This is expected, as, by design, the ESP reverts back to climatology at increasing lead times (i.e. when the effect of the IHC fades away).~~

~~The CM-SSF appears generally less reliable than the ESP, especially at longer lead times (i.e. increasing spread of lines around the diagonal; Fig. 4). The CM-SSF is on average most [least] reliable for both lead times in autumn [spring], shown by the smaller [larger] spread of lines around the diagonal. For the autumn and winter, for both lead times, most lines are situated below the diagonal and outside of the tolerance bands, signifying that the CM-SSF mostly under predicts the simulated streamflow within the hindcast period. For the winter, a few horizontal or near-horizontal lines can be observed for both lead times, meaning that the CM-SSF predictions are sometimes too narrow. For the spring and summer, most lines are situated above the diagonal and outside of the tolerance bands, suggesting that the CM-SSF mostly over predicts the simulated streamflow.~~

### 3.2 Potential usefulness of the CM-SSF

In the second part of the results, the potential usefulness of the CM-SSF compared to the ESP is described~~,~~ for decision-making. Here, potential usefulness is defined as the ability of the forecasting systems to predict lower or higher streamflows than normal, as measured with the ROC score.

Generally, either of the two forecasting systems (CM-SSF or ESP) is capable of~~to~~ predicting skilfully whether the streamflow will be anomalously low or high in the coming months (Fig. 5). However, for a few seasons and regions, none of the two forecasting systems is skilful at predicting lower and/or higher streamflows than normal. This is especially noticeable in winter.

For most seasons and regions, the ESP is more skilful than the CM-SSF at predicting lower and higher streamflows than normal. However, in winter for most regions and during other seasons for several regions, the CM-SSF appears more skilful than the ESP. Regions where the CM-SSF best predicts lower and higher streamflows than normal at most lead times are summarised in Table 1 for all four seasons and the lower and upper terciles of the simulated streamflow.

## 4 Discussion

### 4.1 Does seasonal climate information improve the predictability of seasonal streamflow forecasts over Europe?

On average over Europe and across all seasons, the CM-SSF is skilful (in terms of hindcast accuracy, sharpness and overall performance, using the ESP as a benchmark), for the first month of lead time only. This means that, on average, Sys4 improves the predictability over historical meteorological information for pan-European seasonal streamflow forecasting for the first month of lead time only. At longer lead times, historical meteorological information becomes as good as or better

than Sys4 for seasonal streamflow forecasting over Europe. Crochemore et al. (2016) and Meißner et al. (~~in review~~2017) similarly found positive skill in the seasonal streamflow forecast (Sys4 forced hydrological model compared to an ESP) for the first month of lead time, after which the skill faded away, for basins in France and Central Europe respectively. Additionally, ~~O~~on average over Europe and across all seasons, the CM-SSF is less reliable than the ESP for all lead times. This is due to a combination of too narrow and biased CM-SSF hindcasts, where the bias depends on the season that is being forecasted. As mentioned in the methods section of this paper, the ESP is not a 'naive' benchmark, which might partially explain the limited predictability gained from Sys4.

T~~However, t~~he predictability varies per season and the CM-SSF predictions are on average sharper ~~(increasingly at increasing lead times)~~ than and as accurate as the ESP predictions in autumn and winter beyond the first month of lead time (and increasingly sharper with longer lead times). The CM-SSF however~~however~~ tends to systematically under-predict the autumn and winter simulated streamflow (and increasingly under-predicts with longer lead times). In spring and summer, the CM-SSF predictions are on average less sharp and less accurate than the ESP predictions, and they tend to systematically over-predict the simulated streamflow (and increasingly over-predicts with longer lead times). ~~By design, the ESP is almost perfectly reliable for all seasons, regions and lead-times, with the exception of a few cases where it is under predictive for the first month of lead time in winter and summer, due to the IHC. Contrastingly, the CM-SSF tends to systematically under- [over-] predict the autumn and winter [spring and summer] simulated streamflow.~~

The added predictability gained from Sys4 was shown to lead to skilful CM-SSF predictions of lower and higher streamflows than normal for specific seasons and regions. The CM-SSF is more skilful at predicting anomalously low and high streamflows than the ESP in certain seasons and regions, and noticeably in winter in almost 40% ~~many parts~~ of the Europe~~an regions, mostly clustered in rainfall-dominated areas of Western and Central Europe~~. Several authors have discussed the higher winter predictability over (parts of) Europe, with examples in basins in France (Crochemore et al., 2016), Central Europe (Steirou et al., 2017), the UK (Bell et al., 2017) and the Iberian Peninsula (Lorenzo-Lacruz et al., 2011). Bierkens and van Beek (2009) additionally showed that there was a higher winter predictability in Scandinavia, the Iberian Peninsula and around the Black Sea. Our results are mostly consistent with these findings, except for Scandinavia, where the ESP is ~~mostly~~ more skilful than the CM-SSF in winter. Bierkens and van Beek (2009) produced the seasonal streamflow forecast analysed in their paper by forcing a hydrological model with resampled years of historical meteorological information based on their winter NAO index. However, Sys4 has difficulties in forecasting the NAO over Europe (Kim et al., 2012), which could have led to these inconsistent results with the ones presented by Bierkens and van Beek (2009).

In spring, the CM-SSF is more skilful than the ESP at predicting lower and higher streamflows than normal beyond one month of lead time in approximately 15% of the European regions, and mostly in regions of Western Europe. This could be due to a persistence of the skill from the previous winter through the land surface memory (i.e. groundwater-driven streamflow or snowmelt-driven streamflow), as highlighted by Bierkens and van Beek (2009) for Europe, Singla et al. (2012) for parts of France, Lorenzo-Lacruz et al. (2011) for the Iberian Peninsula and Meißner et al. (~~in review~~2017) for the

Rhine. Moreover, it could be that most of the gained predictability occurs in March, a transition month between the more predictable winter (as mentioned above) and spring, as discussed by Steirou et al. (2017). The ESP is overall more skilful than the CM-SSF at predicting the spring streamflow in snow-dominated regions (e.g. most of Fennoscandia and parts of Central and Eastern Europe). This hints the importance of the IHC (i.e. of snowpack) and the land surface memory for forecasting the spring streamflow in snow-dominated regions in Europe.

The added predictability from Sys4 for forecasting lower and higher streamflows than normal is limited in summer and autumn for most regions. The CM-SSF is more skilful at predicting anomalously low and high streamflows than the ESP in about 10-20% of the European regions during those seasons. Other studies have found similar patterns for (parts of) Europe; these include, less skill in summer than in winter overall for basins in France (Crochemore et al., 2016), less skill for the low flow season (July to October) for basins in Central Europe (Meißner et al., ~~in review~~2017), negative correlations in summer and autumn seasonal streamflow forecasts in Central Europe as the influence of the winter NAO fades away (Steirou et al., 2017), and less skill overall in summer than in winter in Europe (Bierkens and van Beek, 2009). The lower CM-SSF skill for predicting lower and higher streamflows than normal in summer could additionally be due to the convective storms in summer over Europe, which are hard to predict, and to the fact that it is the dry season in most of Europe, where rivers are groundwater fed. Therefore, in this season, the quality of the IHC controls the streamflow predictability.

While the CM-SSF is most skilful (in terms of hindcast accuracy, sharpness and overall performance, using the ESP as a benchmark) in autumn and winter and most potentially useful in winter ~~and spring~~, this does not appear to correlate with high performance in the Sys4 precipitation and temperature hindcasts [as seen on the maps of correlation for Sys4 precipitation and temperature for all four seasons (SON, DJF, MAM and JJA) and with two months of lead time (as identified in this paper); available at https://meteoswiss.shinyapps.io/skill_metrics/]. Over Europe, the Sys4 precipitation and temperature hindcasts are the most skilful in summer and the least skilful in autumn and winter. Moreover, the regions of high CM-SSF skill for predicting lower and upper streamflows than normal do not clearly correspond to regions of high performance in the Sys4 precipitation and temperature hindcasts. These differences could be partially induced by the different benchmark used to evaluate the skill of the CM-SSF (i.e. the ESP) compared to the one used to look at the performance of the Sys4 precipitation and temperature hindcasts (i.e. ERA Interim). However, these results clearly indicate that looking at the performance of the Sys4 precipitation and temperature hindcasts only does not give a good indication of the skill and potential usefulness of the seasonal streamflow hindcasts over Europe ~~[as shown by Neumann et al. (submitted to J. Hydrometeorol.) for the 2013/14 Thames River floods]~~, and that marginal performance in seasonal climate forecasts can translate through to more predictable seasonal streamflow forecasts, and vice versa. The added predictability in the CM-SSF could be due to the combined predictability in the precipitation and temperature hindcasts, as well as a lag in the predictability from the land surface memory.

In most regions and for most seasons, at least one of the two forecasting systems (CM-SSF or ESP) is able to predict lower or higher streamflows than normal. However, in winter, the number of regions and lead times for which none of the

forecasting systems are skilful increases. This could be because in winter, many regions experience weather-driven high streamflows and the performance of Sys4 is limited at this time of year (as mentioned above). In those regions, the seasonal streamflow forecasts could be improved either by improving the IHC, through for example data assimilation, or by improving the seasonal climate forecasts.

Overall, the ESP appears very skilful at forecasting lower or higher streamflows than normal, showing the importance of IHC and the land surface memory for seasonal streamflow forecasting (Wood and Lettenmaier, 2008; Bierkens and van Beek, 2009; Yuan et al., 2015b).

### 4.2 What is the potential usefulness and usability of the EFAS seasonal streamflow forecasts for flood preparedness?

What appears like little added skill does not necessarily mean no skill for the forecast users and can in fact be a large added value for decision-making (Viel et al., 2016). The ability of a seasonal streamflow forecasting system to predict the right category of an event months ahead is valuable for many water-related applications ~~in the water sector~~ (e.g. navigation, reservoir management, drought-risk management, irrigation, water resources management, hydropower and flood preparedness). From the results presented in this paper, it appears that either of the two forecasting systems (CM-SSF or ESP) are capable of predicting lower or higher streamflows than normal months in advance, thanks to the predictability gained from the IHC, the land surface memory and the seasonal climate hindcast in some regions and for certain seasons.

However, as highlighted by White et al. (2017), there is currently a gap between usefulness and usability of seasonal information. What is a useful scientific finding does not automatically translate into usable information which will fit into any user's decision-making chain (Soares and Dessai, 2016). While several authors have already investigated the usability of seasonal streamflow forecasts for applications such as navigation (Meißner et al., ~~in review~~2017), reservoir management (Viel et al., 2016; Turner et al., ~~in review~~2017), drought-risk management (Sheffield et al., 2013; Yuan et al., 2013; Crochemore et al., 2017), irrigation (Chiew et al., 2003; Li et al., ~~in review~~2017), water resources management (Schepen et al., 2016) and hydropower (Hamlet et al., 2002), its application to flood preparedness is still left mostly unexplored. ~~One exception being Neumann et al. (submitted to J. Hydrometeorol.) who look at the use of the CM-SSF to predict the 2013/14 Thames basin floods.~~ This is partially due to the complex nature of flood generating mechanisms, still poorly studied on seasonal timescales beyond snowmelt-driven spring floods, as well as the fact that seasonal forecasts reflect the likelihood of abnormal seasonal streamflow totals, but without much skilful information on the exact timing, location and the severity of the impact of individual flood events within that season. Coughlan de Perez et al. (2017) looked at the usefulness of seasonal rainfall forecasts for flood preparedness in Africa and highlighted the complexities behind using these forecasts as a proxy for floodiness [for discussion on floodiness see Stephens et al. (2015)]. Furthermore, decision-makers in the navigation, reservoir management, drought-risk management, irrigation, water resources management and hydropower sectors are familiar with working on long timescales (i.e. several weeks to months ahead). In contrast, the flood preparedness community is currently mostly used to working on timescales of hours to a couple of days.

The Red Cross Red Crescent Climate Centre has recently designed a new approach that harnesses the usefulness of seasonal climate information for decision-making for disaster management. This approach, called 'Ready-Set-Go!', is made of three stages. The 'Ready' stage is based on seasonal forecasts, where they are used as monitoring information to drive contingency planning (e.g. volunteer training). The 'Set' stage is triggered by sub-seasonal forecasts, used as early-warning information to alert volunteers. Finally, the 'Go!' stage is based on short-range forecasts and consists in the evacuation of people and the distribution of aid (White et al., 2017). Using a similar approach, seasonal streamflow forecasts could complement existing forecasts at shorter timescales and provide monitoring and early-warning information for flood preparedness.

Such an approach however requires the use of consistent forecasts from short to seasonal timescales. In this context, moving to seamless forecasting is becoming vital ~~(Wetterhall and Di Giuseppe, in review)~~.

Soares and Dessai (2016) also identified the accessibility to the information, enhanced by collaborations and ongoing relationships between users and producers, as a key enabler of the usability of seasonal information. International projects, such as the Horizon 2020 IMPREX (IMproving PRedictions and management of hydrological EXtremes) project (van den Hurk et al., 2016), alongside promoting scientific progress on hydrological extremes forecasting from short to seasonal timescales over Europe, gather together forecasters and decision-makers and can effectively demonstrate the added value of the integration of seasonal information in decision-making chains. The Hydrologic Ensemble Prediction EXperiment (HEPEX) is another international initiative that brings together researchers and practitioners in the field of ensemble prediction for water-related applications. It is an ideal environment for collaboration~~s~~ and fosters communication and outreach on~~f~~ topics such as the usefulness and usability of seasonal information for decision-making.

### 4.3 Aspects for future work

In this paper, terciles of the simulated streamflow are used. However, and because the application of the EFAS seasonal streamflow forecasts is of particular relevance for flood preparedness, the evaluation of the hindcasts for lower and higher streamflow extremes (for example the $5^{th}$ and the $95^{th}$ percentiles respectively) would be more relevant and might give very different results. This was not done in this paper as the time period covered by the seasonal streamflow hindcasts (i.e. approximately 27 years) was not long enough for statistically reliable results for lower and higher streamflow extremes. The limited hindcast length is a common problem in seasonal predictability studies. Increasing the hindcast length back in time could lead to more stable Sys4 hindcasts and hence to more stable and potentially skilful seasonal streamflow hindcasts (Shi et al., 2015).

Furthermore, in this paper, the hindcasts were analysed against simulated streamflow, used as a proxy for observed streamflow. This is necessary because it enables an analysis of the quality of the hindcasts over the entire computation domain, rather than at non-evenly spaced stations over the same domain (Alfieri et al., 2014). Further work could however include carrying out a similar analysis for selected river stations in Europe, in order to account for model errors in the hindcast evaluation.

The calculation of the verification scores (excluding the ROC) was made by randomly selecting 15 ensemble members from the 51 ensemble members of the CM-SSF hindcasts, for starting dates for which the ensemble varies between 15 and 51 members (i.e. hindcasts made on the 1st of January, March, April, June, July, September, October and December; this is due to the split between 15 and 51 ensemble members in the Sys4 hindcasts, as described in Sect. 2.1.2 of this paper). In order to investigate the potential impact of this evaluation strategy on the results presented in this paper, the CRPSS was calculated for 15 and 51 ensemble members of the CM-SSF hindcasts for starting dates for which 51 ensemble members are available for the full hindcast period (i.e. hindcasts made on the 1st of February, May, August and November). This is displayed in Fig. 6 for all hindcast starting dates, lead times (i.e. one to seven months) and regions combined. Overall, it is apparent that the impact of this evaluation strategy on the results presented in this paper should be minimal, as all points align themselves approximately with the 1 to 1 diagonal.

The next version of the ECMWF seasonal climate forecast, SEAS5, was~~is due to be~~ released in November 2017. Future work could include forcing the Lisflood model with SEAS5 and comparing the obtained seasonal streamflow hindcasts to the CM-SSF presented in this paper. This should indicate whether developments to the seasonal climate forecast translate through to better pan-European seasonal streamflow forecasts, which is of particular interest for regions and seasons when neither the ESP nor the CM-SSF are currently skilful.

The operational EFAS medium-range streamflow forecasts are currently post-processed as a means to improve their reliability (Smith et al., 2016 and references therein). Results from this paper have shown that the CM-SSF is mostly unreliable (with regards to the EFAS-WB) and could hence benefit from post-processing of the seasonal climate forecast. However, post-processing techniques used for the EFAS medium-range streamflow forecasts might not be suitable for the CM-SSF, as the seasonal climate forecast used for the latter should be post-processed in terms of its seasonal anomalies rather than for errors in the timing, volume and magnitude of specific events. This is currently being considered for operational implementation within EFAS and is an active area of discussion within the EFAS user community.

For the analysis presented in this paper, the CM-SSF was benchmarked against the ESP. Several other techniques exist for seasonal streamflow forecasting, such as statistical methods using predictors ranging from climate indices to antecedent observed precipitation and crop production metrics, to mention a few (e.g. Mendoza et al., 2017; Slater et al., 2017). Further analysis could include benchmarking the CM-SSF against one or multiple statistical methods, to assess the relative benefits of various seasonal streamflow forecasting techniques.

In this paper, the ability of both systems (~~i.e. ESP and~~ CM-SSF and ESP) to forecast lower and higher streamflows than normal was explored, with several hypotheses made to link the streamflow~~is~~ predictability to regions' hydro-climatic processes. This includes the higher potential usefulness of the ESP in forecasting~~predictability~~ the~~in~~ spring streamflow in snow-dominated regions and the ~~in~~ summer streamflow in regions where rivers are groundwater-fed. In these regions and for these seasons, ~~both due to~~ the IHC and the land surface memory drive the predictability. ~~It furthermore includes a high~~The CM-SSF provides an added potential usefulness in~~predictability in~~ winter in the rainfall-dominated regions of Central and Western Europe, where the skill appears to persist through to spring due to the land surface memory (i.e.~~via~~ groundwater-

driven streamflow and snowmelt-drive streamflow)~~processes~~. While further exploration of these hypotheses is outside of the scope of this paper, future work is required to ~~should aim to~~ disentangle the links between the added predictability from Sys4 and the basins' hydro-~~-~~climatic characteristics.~~,~~ ~~f~~For example understanding the predictability ~~and potential skill~~ in snow-dominated basins, arid regions and temperate groundwater-fed basins.

~~This paper assesses the added predictability from using seasonal climate forecasts, additionally to the IHC.~~In this context, ~~A~~additional work to further disentangle and quantify the contribution of both predictability sources (seasonal climate forecasts versus IHC) to seasonal streamflow forecasting quality over Europe could be carried out by using the EPB (end point blending) method (Arnal et al., 2017).

## 5 Conclusions

In this paper, the newly operational EFAS seasonal streamflow forecasting system [producing the CM-SSF forecasts by forcing the Lisflood model with the ECMWF System 4 seasonal climate forecasts (Sys4)] was presented and benchmarked ~~(in terms of hindcast accuracy, sharpness, reliability and overall performance)~~ against the ESP forecasting approach (ESP forecasts produced by forcing the Lisflood model with historical meteorological observations) for the hindcast period 1990 to 2017. On average, Sys4 improves the predictability over historical meteorological information for pan-European seasonal streamflow forecasting for the first month of lead time only (in terms of hindcast accuracy, sharpness and overall performance). However, the predictability varies per season and the CM-SSF is more skilful on average at predicting autumn and winter streamflows than spring and summer streamflows. Additionally, parts of Europe exhibit a longer predictability, up to seven months of lead time, for certain months within a season. In terms of hindcast reliability, the CM-SSF is on average less skilful than the ESP for all lead times, due to a combination of too narrow and biased CM-SSF hindcasts, where the bias depends on the season that is being forecasted.

Subsequently, the potential usefulness of the two forecasting systems (CM-SSF and ESP) was assessed by analysing their skill in predicting lower and higher streamflows than normal. Overall, at least one of the two forecasting systems is capable of predicting those events months in advance. The ESP appears the most skilful on average, showing the importance of IHC and the land surface memory for seasonal streamflow forecasting. Nevertheless, for certain regions and seasons the CM-SSF is the most skilful at predicting anomalously low or high streamflows beyond one month of lead time, noticeably in winter for almost 40% of the European regions~~, beyond one month of lead time~~. This potential usefulness could be harnessed by using seasonal streamflow forecasts as complementary information to existing forecasts at shorter timescales, to provide monitoring and early-warning information for flood preparedness.

Overall, ~~P~~patterns in skill in the CM-SSF are however not mirrored in the Sys4 precipitation and temperature hindcasts. This hints that using seasonal climate forecast performance as a proxy for seasonal streamflow forecasting skill is not adequate and that more work is needed to understand the link between meteorological and hydrological variables on seasonal timescales over Europe.

**Acknowledgments and data availability**

L. Arnal, H. L. Cloke and J. Neumann gratefully acknowledge financial support from the Horizon 2020 IMPREX project (Grant Agreement 641811) (project IMPREX: www.imprex.eu). L. Arnal's time was additionally partly funded by a University of Reading PhD Scholarship. F. Wetterhall, C. Prudhomme and B. Krzeminski's work was supported by the EFAS computational centre in support to the Copernicus Emergency Management Service/Early Warning Systems (Flood) (contract No198702 from JRC-IES). E. Stephens is thankful for support from the Natural Environment Research Council and Department for International Development (Grant number NE/P000525/1) under the Science for Humanitarian Emergencies and Resilience (SHEAR) research programme. The data from the European Flood Awareness System are available to researchers upon request (subject to licensing conditions). Please visit www.efas.eu for more details.

**References**

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, J. Hydrol., 517, 913-922, doi:10.1016/j.jhydrol.2014.06.035, 2014.

Arnal, L., Wood, A. W., Stephens, E., Cloke, H. L., and Pappenberger F.: An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity, J. Hydrometeorol., 18, 1715-1729, doi:10.1175/JHM-D-16-0259.1, 2017.

Arribas, A., Glover, M., Maidens, A., Peterson, K., Gordon, M., MacLachlan, C., Graham, R., Fereday, D., Camp, J., Scaife, A. A., Xavier, P., McLean, P., and Colman, A.: The GloSea4 Ensemble Prediction System for Seasonal Forecasting, Mon. Weather. Rev., 139, 1891-1910, doi:10.1175/2010MWR3615.1, 2010.

Bell, V. A., Davies, H. N., Kay, A. L., Brookshaw, A., and Scaife, A. A.: A national-scale seasonal hydrological forecast system: development and evaluation over Britain, Hydrol. Earth Syst. Sc., 21, 4681-4691, doi: 10.5194/hess-21-4681-2017, 2017.

Bennett, J. C., Wang, J, Q., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resour. Res., 52, 8238-8259, doi:10.1002/2016WR019193, 2016.

Bierkens, M. F. and van Beek, L. P.: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, J. Hydrometeorol., 10, 953-968, doi:10.1175/2009JHM1034.1, 2009.

Candogan Yossef, N., van Beek, R., Weerts, A., Winsemius, H., and Bierkens, M. F.: Skill of a global forecasting system in seasonal ensemble streamflow prediction, Hydrol. Earth Syst. Sc., 21, 4103-4114, doi:10.5194/hess-21-4103-2017, 2017.

Céron, J.-P., Tanguy, G., Franchistéguy, L., Martin, E., Regimbeau, F., and Vidal, J.-P.: Hydrological seasonal forecast over France: feasibility and prospects, Atmos. Sci. Lett., 11, 78-82, doi:10.1002/asl.256, 2010.

Chiew, F. H., Zhou, S. L., and McMahon, T. A.: Use of Seasonal Streamflow Forecasts in Water Resources Management, J. Hydrol., 270, 135–144, doi:10.1016/S0022-1694(02)00292-5, 2003.

Church, J. E. and Merrill, M. C. (Eds.): Principles of snow surveying as applied to forecasting stream flow, Vol. 51, Journal of Agricultural Research, Washington, D. C., 1935.

Coughlan de Perez, E., Stephens, E., Bischiniotis, K., van Aalst, M., van den Hurk, B., Mason, S., Nissan, H., and Pappenberger, F.: Should seasonal rainfall forecasts be used for flood preparedness? Hydrol. Earth Syst. Sc., 21, 4517-4524, doi:10.5194/hess-21-4517-2017, 2017.

Crochemore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrol. Earth Syst. Sc., 20, 3601-3618, doi:10.5194/hess-2016-78, 2016.

Crochemore, L., Ramos, M.-H., Pappenberger, F., and Perrin, C.: Seasonal streamflow forecasting by conditioning climatology with precipitation indices. Hydrol. Earth Syst. Sc., 21, 1573-1591, doi:10.5194/hess-21-1573-2017, 2017.

Day, G. N.: Extended streamflow forecasting using NWSRFS, J. Water Res. Plan. Man., 111, 157-170, doi:10.1061/(ASCE)0733-9496(1985)111:2(157), 1985.

De Roo, A. P., Wesseling, C. G., and Van Deursen, W. P.: Physically based river basin modelling within a GIS: the LISFLOOD model, Hydrol. Process., 14, 1981-1992, doi:10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F, 2000.

Demirel, M. C., Booij, M., and Hoekstra, A.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, Hydrol. Earth Syst. Sc., 19, 275-291, doi:10.5194/hess-19-275-2015, 2015.

Dettinger, M. D. and Diaz, H. F.: Global characteristics of stream flow seasonality and variability, J. Hydrometeorol., 1, 289-310, doi:10.1175/1525-7541(2000)001<0289:GCOSFS>2.0.CO;2, 2000.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R.: Seasonal climate predictability and forecasting: status and prospects, WIRES Clim. Change., 4, 245-268, doi:10.1002/wcc.217, 2013.

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

Gobena, A. K. and Gan, T. Y.: Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system, J. Hydrol., 385, 336-352, doi:10.1016/j.jhydrol.2010.03.002, 2010.

Goddard, L., Mason, S. J., Zebiak, S. E., Ropelewski, C. F., Basher, R., and Cane, M. A.: Current approaches to seasonal to interannual climate predictions, Int. J. Climatol., 21, 1111-1152, doi:10.1002/joc.636, 2001.

Guimarães Nobre, G., Jongman, B., Aerts, J., and Ward, P. J.: The role of climate variability in extreme floods in Europe, Environ. Res. Lett., 12, 084012, doi:10.1088/1748-9326/aa7c22, 2017.

Hamlet, A. F., Huppert, D., and Lettenmaier, D. P.: Economic Value of Long-Lead Streamflow Forecasts for Columbia River Hydropower, J. Water Res. Plan. Man., 128, 91-101, doi:10.1061/(ASCE)0733-9496(2002)128:2(91), 2002.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559-570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hurrell, J. W.: Decadal trends in the North Atlantic oscillation: Regional temperatures and precipitation, Science, 269, 676-679, doi:10.1126/science.269.5224.676, 1995.

Hurrell, J. W. and Van Loon, H.: Decadal Variations in Climate Associated with the North Atlantic Oscillation, in: Climatic Change at High Elevation Sites, Diaz, H. F., M. Beniston and R. S. Bradley (Eds.), Springer, Dordrecht, 69-94, doi:10.1007/978-94-015-8905-5_4, 1997.

Keller, J. D. and Hense, A.: A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms, Meteorol. Z., 20, 107-117, doi:10.1127/0941-2948/2011/0217.

Kim, H.-M., Webster, P. J., and Curry, J. A.: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, Clim. Dynam., 39, 2957–2973, doi: 10.1007/s00382-012-1364-6, 2012.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sc., 11, 1267–1277, doi:10.5194/hess-11-1267-2007, 2007.

Li, Y., Giuliani, M., and Castelletti, A.: A coupled human-natural system to assess the operational value of weather and climate services for irrigated agriculture, Hydrol. Earth Syst. Sc., 21, 4693-4709, doi: 10.5194/hess-21-4693-2017, 2017.

Lorenzo-Lacruz, J., Vicente-Serrano, S. M., López-Moreno, J. I., González-Hidalgo, J. C., and Morán-Tejeda, E.: The response of Iberian rivers to the North Atlantic Oscillation, Hydrol. Earth Syst. Sc., 15, 2581-2597, doi:10.5194/hess-15-2581-2011, 2011.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I.: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, Reviews of Geophysics, 48, Rg3003, doi:10.1029/2009rg000314, 2010.

Mason, S. J. and Graham, N. E.: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, Weather Forecast., 14, 713-725, doi:10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2, 1999.

Meißner, D., Klein, B., and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in Central Europe, Hydrol. Earth Syst. Sc., 21, 6401-6423, doi:10.5194/hess-21-6401-2017, 2017~~in review~~.

Mendoza, P. A., Wood, A. W., Clark, E., Rothwell, E., Clark, M. P., Nijssen, B., Brekke, L. D., and Arnold, J. R.: An intercomparison of approaches for improving operational seasonal streamflow forecasts, Hydrol. Earth Syst. Sc., 21, 3915-3935, doi:10.5194/hess-21-3915-2017, 2017.

Mo, K. C. and Lettenmaier, D. P.: Hydrologic Prediction over the Conterminous United States Using the National Multi-Model Ensemble, J. Hydrometeorol., 15, 1457-1472, doi:10.1175/JHM-D-13-0197.1, 2014.

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), ECMWF Tech. Memorandum 656, 1-49, 2011.

~~Neumann, J. L., Arnal, L., Magnusson, L., and Cloke, H.: The 2013/14 Thames basin floods: Do improved meteorological forecasts lead to more skilful hydrological forecasts at seasonal timescales?, submitted to J. Hydrometeorol.~~

Pagano, T. C. and Garen, D. C.: Integration of climate information and forecasts into western US water supply forecasts, Climate variations, climate change, and water resources engineering, 86-103, 2006.

Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., Jackson, C., Svensson, C., Parry, S., Bachiller-Jareno, N., Davies, H. N., Davis, R., ~~Harrigan, S.,~~ Mackay, J., Mackenzie, A., Rudd, A. C., Smith, K., Bloomfield, J., Ward, R., and Jenkins, A.: Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales, Hydrolog. Sci. J., 62, 2753-2768, doi: 10.1080/02626667.2017.1395032, 2017~~in review~~.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, doi:10.1029/2009WR008328, 2010.

Schepen, A., Zhao, T., Wang, Q. J., Zhou, S., and Feikema, P.: Optimising seasonal streamflow forecast lead time for operational decision making in Australia, 20, 4117–4128, doi:10.5194/hess-20-4117-2016, 2016.

Sheffield, J., Wood, E. F., Chaney, N., Guan, K., Sadri, S., Yuan, X., Olang, L., Amani, A., Ali, A., Demuth, S., and Ogallo, L.: A Drought Monitoring and Forecasting System for Sub-Sahara African Water Resources and Food Security, B. Am. Meteorol. Soc., 95, 861-882, doi:10.1175/BAMS-D-12-00124.1, 2013.

Shi, W., Schaller, N., MacLeod, D., Palmer, T. N., and Weisheimer, A.: Impact of hindcast length on estimates of seasonal climate predictability, Geophys. Res. Lett., 42, 1554–1559, doi: 10.1002/2014GL062829, 2015.

Singla, S., Céron, J. P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., and Vidal, J. P.: Predictability of soil moisture and river flows over France for the spring season, Hydrol. Earth Syst. Sc., 16, 201-216, doi:10.5194/hess-16-201-2012, 2012.

Slater, L. J., Villarini, G., Bradley, A. A., and Vecchi, G. A.: A dynamical statistical framework for seasonal streamflow forecasting in an agricultural watershed, Clim. Dynam., 1-17, doi:10.1007/s00382-017-3794-7, 2017.

Smith, P., Pappenberger, F., Wetterhall, F., Thielen, J., Krzeminski, B., Salamon, P., Muraro, D., Kalas, M., and Baugh, C.: On the operational implementation of the European Flood Awareness System (EFAS), ECMWF Tech. Memorandum 778, 1-34, 2016.

Soares, M. B. and Dessai, S.: Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe, Climatic Change, 137, 89–103, doi:10.1007/s10584-016-1671-8, 2016.

Steirou, E., Gerlitz, L., Apel, H., and Merz, B.: Links between large-scale circulation patterns and streamflow in Central Europe: A review, J. Hydrol., 549, 484-500, doi:10.1016/j.jhydrol.2017.04.003, 2017.

Stephens, E., Day, J. J., Pappenberger, F., and Cloke, H.: Precipitation and floodiness, Geophys. Res. Lett., 42, 10,316–10,323,doi: 10.1002/2015GL066779, 2015.

Svensson, C., Brookshaw, A., Scaife, A. A., Bell, V. A., Mackay, J. D., Jackson, C. R., Hannaford, J., Davies, H. N., Arribas, A., and Stanley, S.: Long-range forecasts of UK winter hydrology, Environ. Res. Lett., 10, 064006, doi:10.1088/1748-9326/10/6/064006, 2015.

Troccoli, A.: Seasonal climate forecasting, Meteorol. Appl., 17, 251-268, doi:10.1002/met.184, 2010.

Turner, S. W., Bennett, J. C., Robertson, D. E. and Galelli, S.: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations~~Value of seasonal streamflow forecasts in emergency response reservoir management~~, Hydrol. Earth Syst. Sc., 21, 4841-4859, doi:10.5194/hess-21-4841-2017, 2017~~in review~~.

Twedt, T. M., Schaake, J. C., and Peck, E. L.: National Weather Service extended streamflow prediction [USA], Proceedings Western Snow Conference, 1977.

van den Hurk, B. J. J. M., Bouwer, L. M., Buontempo, C., Döscher, R., Ercin, E., Hananel, C., Hunink, J., Kjellström, E., Klein, B., Manez, M., Pappenberger, F., Pouget, L., Ramos, M.-H., Ward, P. J., Weerts, A., and Wijngaard, J.: Improving predictions and management of hydrological extremes through climate services: www.imprex.eu, Climate Services, 1, 6-11, doi:10.1016/j.cliser.2016.01.001, 2016.

Van Der Knijff, J. M., Younis, J., and De Roo, A. P.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, Int. J. Geogr. Inf. Sci., 24, 189-212, doi:10.1080/13658810802549154, 2010.

Viel, C., Beaulant, A.-L., Soubeyroux, J.-M., and Céron, J.-P.: How seasonal forecast could help a decision maker: an example of climate service for water resource management, Adv. Sci. Res., 13, 51-55, doi:10.5194/asr-13-51-2016, 2016.

~~Wetterhall, F. and Di Giuseppe, F.: The benefit of seamless forecasts for hydrological predictions over Europe, Hydrol. Earth Syst. Sc., in review.~~

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., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., and Zebiak, S. E.: Potential applications of subseasonal-to-seasonal (S2S) predictions, Meteorol. Appl., 24, 315-325, doi:10.1002/met.1654, 2017.

Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States, J. Geophys. Res.-Atmos., 110, doi:10.1029/2004JD004508, 2005.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35, L14401, doi:10.1029/2008GL034648, 2008.

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P.: Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res.-Atmos., 107, 4429, doi:10.1029/2001JD000659, 2002.

Yuan, X., Roundy, J. K., Wood, E. F., and Sheffield, J.: Seasonal forecasting of global hydrologic extremes: system development and evaluation over GEWEX basins, B. Am. Meteorol. Soc., 96, 1895-1912, doi:10.1175/BAMS-D-14-00003.1, 2015a.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, Wiley Interdisciplinary Reviews: Water, 2, 523-536, doi:10.1002/wat2.1088, 2015b.

Yuan, X., Wood, E. F., Chaney, N. W., Sheffield, J., Kam, J., Liang, M., and Guan, K.: Probabilistic Seasonal Forecasting of African Drought by Dynamical Models, J. Hydrometeorol., 14, 1706-1720, doi:10.1175/JHM-D-13-054.1, 2013.

Zajac, Z., Zambrano-Bigiarini, M., Salamon, P., Burek, P., Gentile, A., and Bianchi, A.: Calibration of the lisflood hydrological model for europe - calibration round 2013, Joint Research Centre, European Commission, 2013.

Forecast skill metrics: https://meteoswiss.shinyapps.io/skill_metrics/, last access: 3 October 2017.

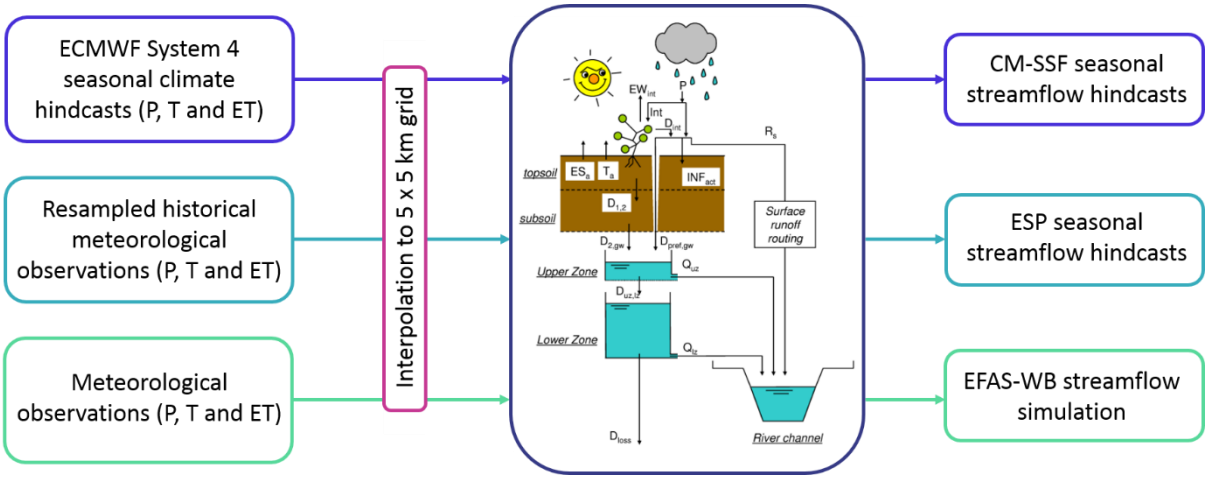
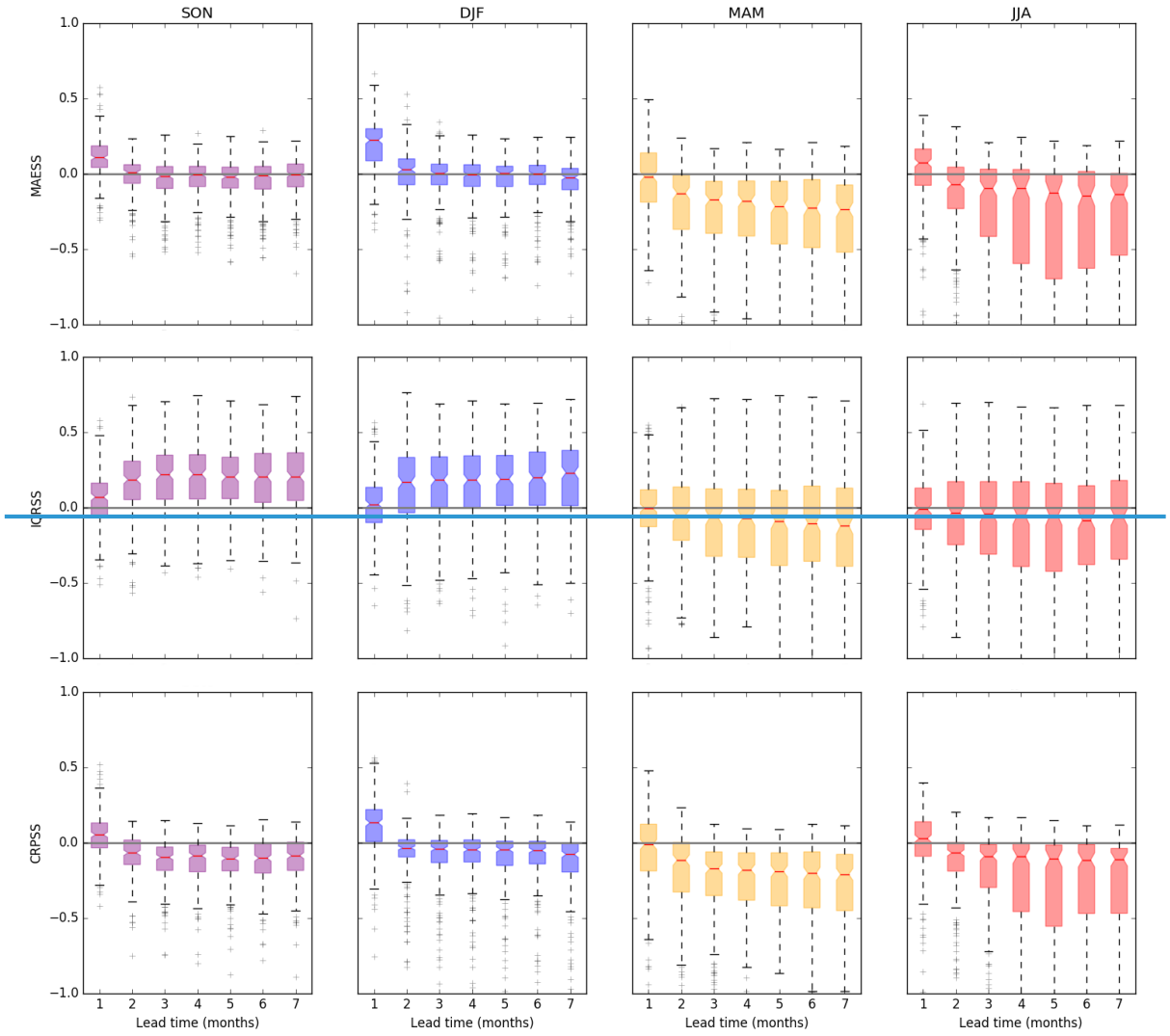



**Figure 1: Schematic of the EFAS-WB streamflow simulation and of the CM-SSF and ESP seasonal streamflow hindcasts generation.**


**Figure 22: Map of the 74 European regions (dark blue outlines) selected for the analysis of the CM-SSF and the ESP.**


SON
DJF
MAM
JJA
MAESS
IGRSS
CRPSS
1.0
0.5
0.0
−0.5
−1.0
1 2 3 4 5 6 7
Lead time (months)

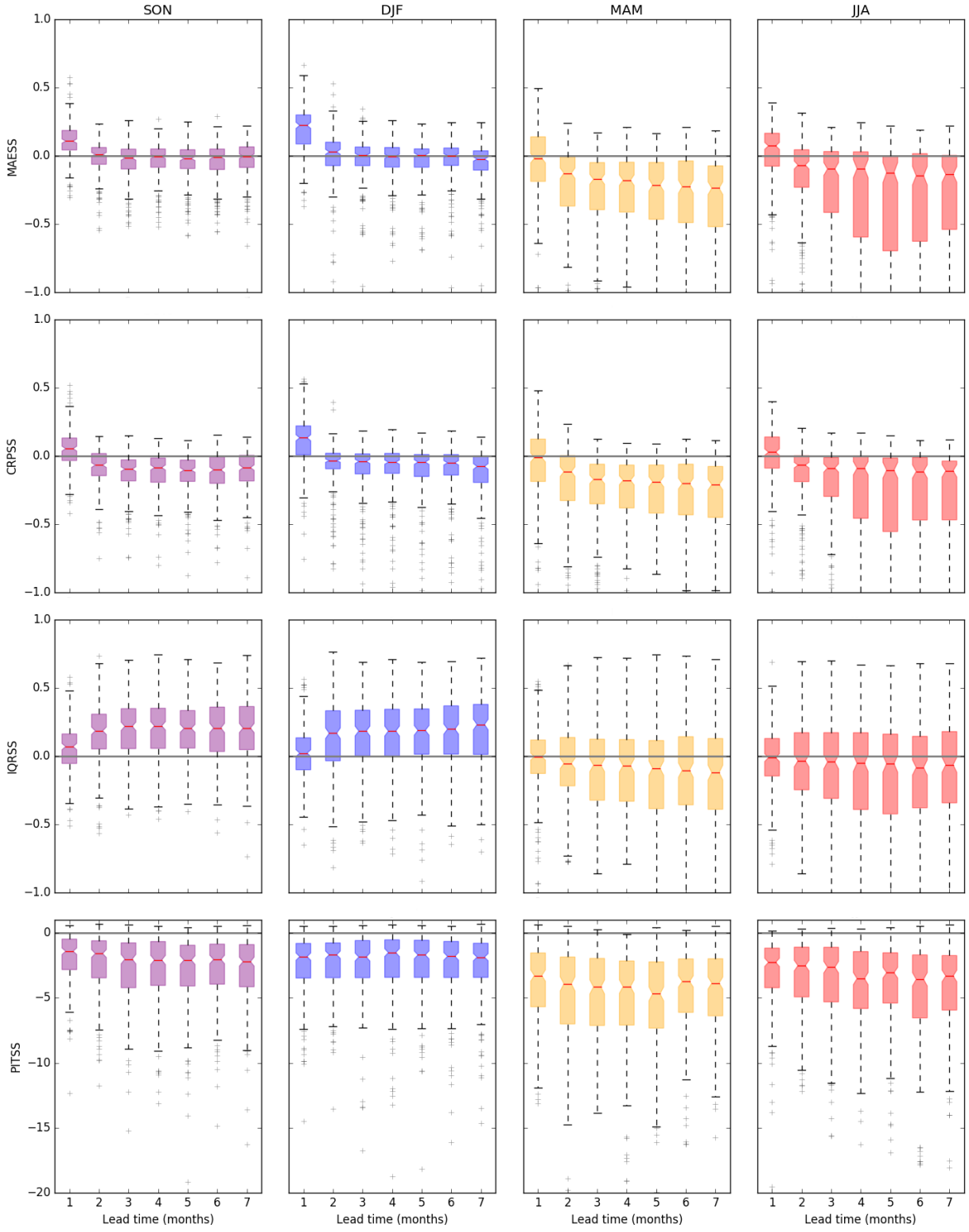

SON
DJF
MAM
JJA
MAESS
CRPSS
IQRSS
PITSS
1.0
0.5
0.0
−0.5
−1.0
0
−5
−10
−15
−20
1 2 3 4 5 6 7
Lead time (months)

**Figure 3: Boxplots of the MAESS ~~(top row)~~, CRPSS, IQRSS ~~(middle row)~~ and ~~CRPSS (bottom row)~~PITSS (from the top to the bottom row) for all four seasons (SON, DJF, MAM and JJA from the left-most to the right-most column) as a function of lead time (i.e. ~~i.e.~~ one to seven months). The boxplots contain the scores for all target months falling in a given season and all 74 European regions. For all scores, values larger [smaller] than zero indicate that the CM-SSF is more [less] skilful than the ESP (benchmark). Where the skill is zero, the CM-SSF is as skilful as the ESP for the hindcast period. Note that the PITSS plots have a different y-axis scale.**

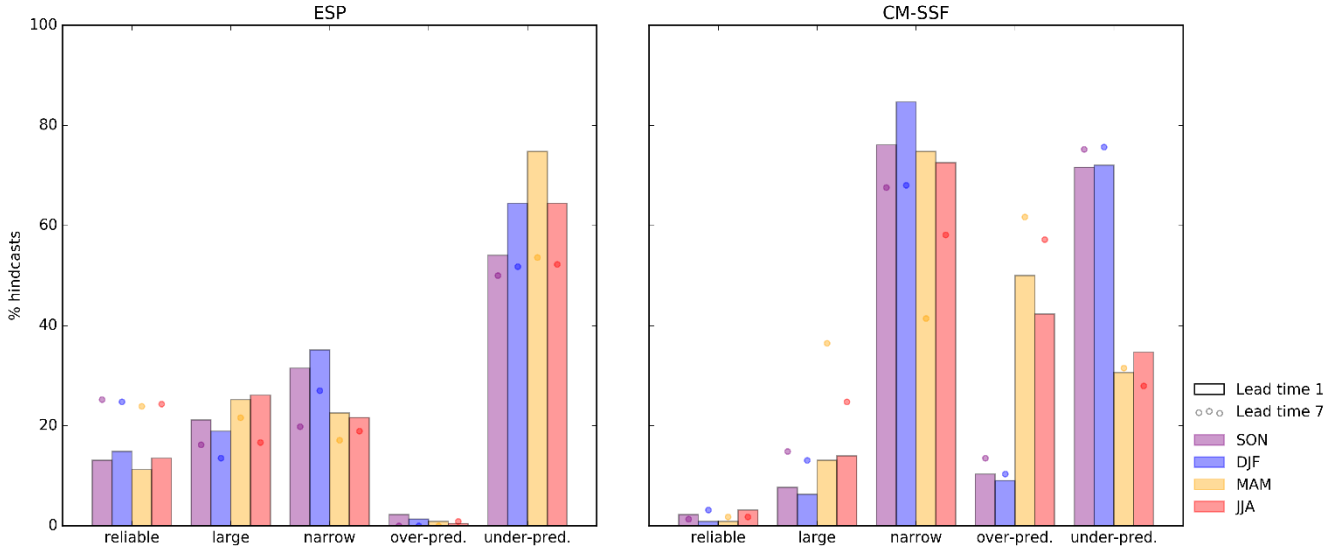



**Figure 4: Plots of the percentage of the ESP (left-hand side) and the CM-SSF (right-hand side) hindcasts falling in each reliability category (reliable - in terms of both spread and bias, too large, too narrow, over-predicting and under-predicting) for all four seasons (SON, DJF, MAM and JJA from the left-most to the right-most bars in each reliability category). The results are shown as bar charts for the first month of lead time and as circles for the seventh month of lead time. These lead times were selected for display to highlight the evolution of reliability between the first and last month of the hindcast. The percentages were calculated from hindcasts for all target months falling in a given season and all 74 European regions.**

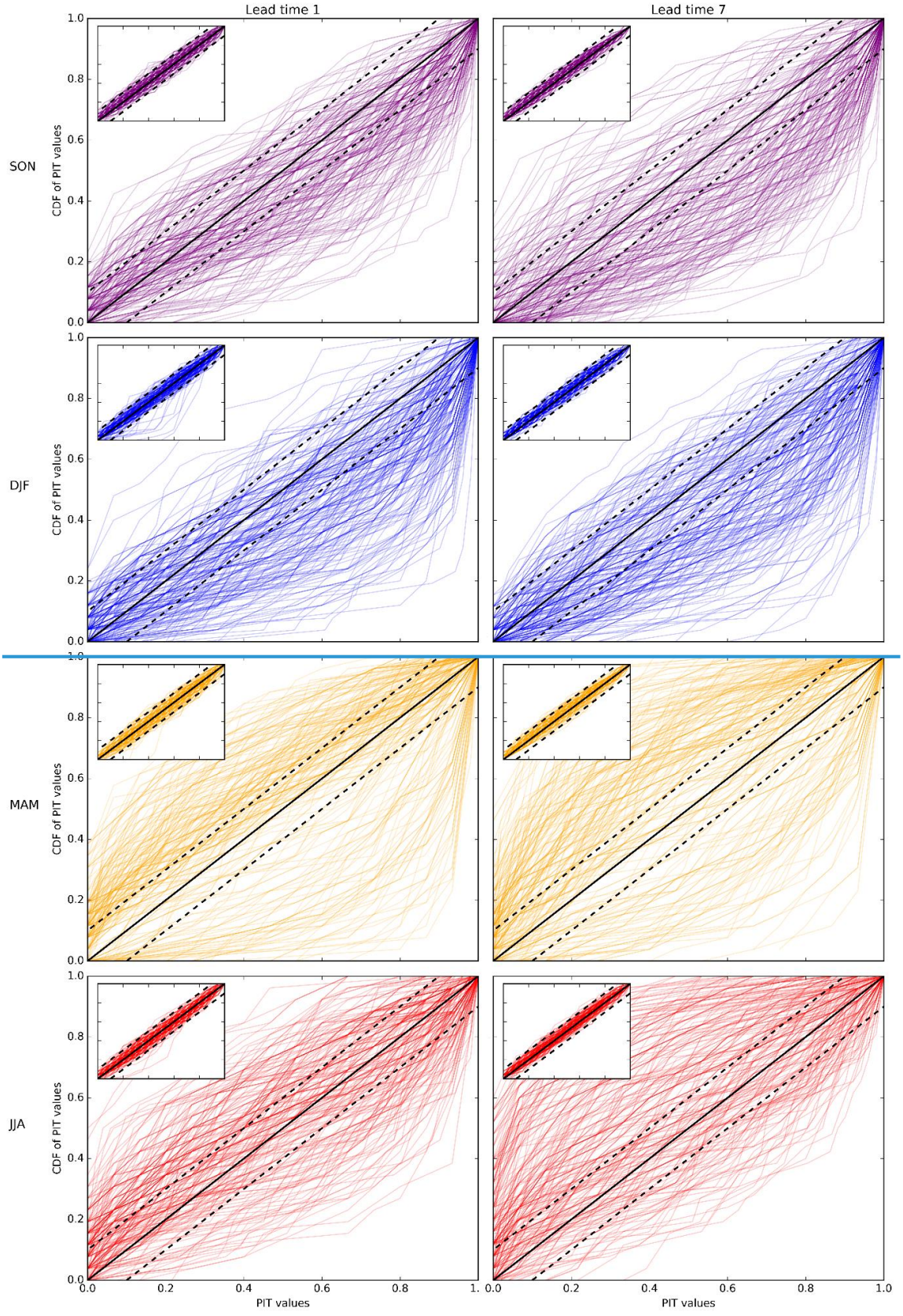

Lead time 1
Lead time 7
SON
DJF
MAM
JJA
CDF of PIT values
PIT values
0.0
0.2
0.4
0.6
0.8
1.0

**Figure 4: PIT diagrams for the CM-SSF (main plots) and the ESP (inplots) for all four seasons (SON, DJF, MAM and JJA from the top to the bottom row) and for one and seven months of lead time (left and right columns, respectively). These lead times were selected for display to highlight the evolution of reliability between the first and last month of the hindcast. The lines in the plots correspond to all the target months falling in a season and all 74 regions. The dotted diagonal lines are the ±0.1 tolerance bands from the 1 to 1 diagonal. Lines above [below] the diagonal and outside of the tolerance bands signify an over-prediction [under-prediction] of the simulated streamflow by the hindcast.**

**Table 1: Regions where the CM-SSF is more skilful than the ESP at predicting anomalously low (lower tercile; first column) or high (upper tercile; second column) streamflows for all four seasons (SON, DJF, MAM and JJA from the top to the bottom row). This is a summary of the information displayed in Fig. 5.**

| | Lower tercile | Upper tercile |
|---|---|---|
| **SON** | - Few regions in Fennoscandia<br>- Po River basin (northern Italy)<br>- Elbe River basin (south of Denmark)<br>- Upstream of the Rhine River basin<br>- Upstream of the Danube River basin<br>- Duero River basin (Iberian Peninsula) | - Few regions in Fennoscandia<br>- Iceland<br>- Parts of the Danube River basin<br>- Segura River basin (Iberian Peninsula) |
| **DJF** | Many regions except:<br>- in most of Fennoscandia North of the Baltic Sea,<br>- parts of Central Europe. | Same as lower tercile. |
| **MAM** | - Few regions on the Iberian Peninsula<br>- Few regions in the western part of Central Europe | Same as lower tercile. |
| **JJA** | - Few regions in the United Kingdom (UK)<br>- Ireland<br>- North-western edge of the Iberian Peninsula<br>- Regions in Fennoscandia around the Baltic Sea<br>- Regions south of the North Sea | - Northern part of the UK<br>- Ireland<br>- North-western edge of the Iberian Peninsula<br>- Regions in Fennoscandia around the Baltic Sea<br>- Around the Elbe River basin<br>- Upstream of the Danube River basin<br>- Along the Adriatic Sea in Italy |

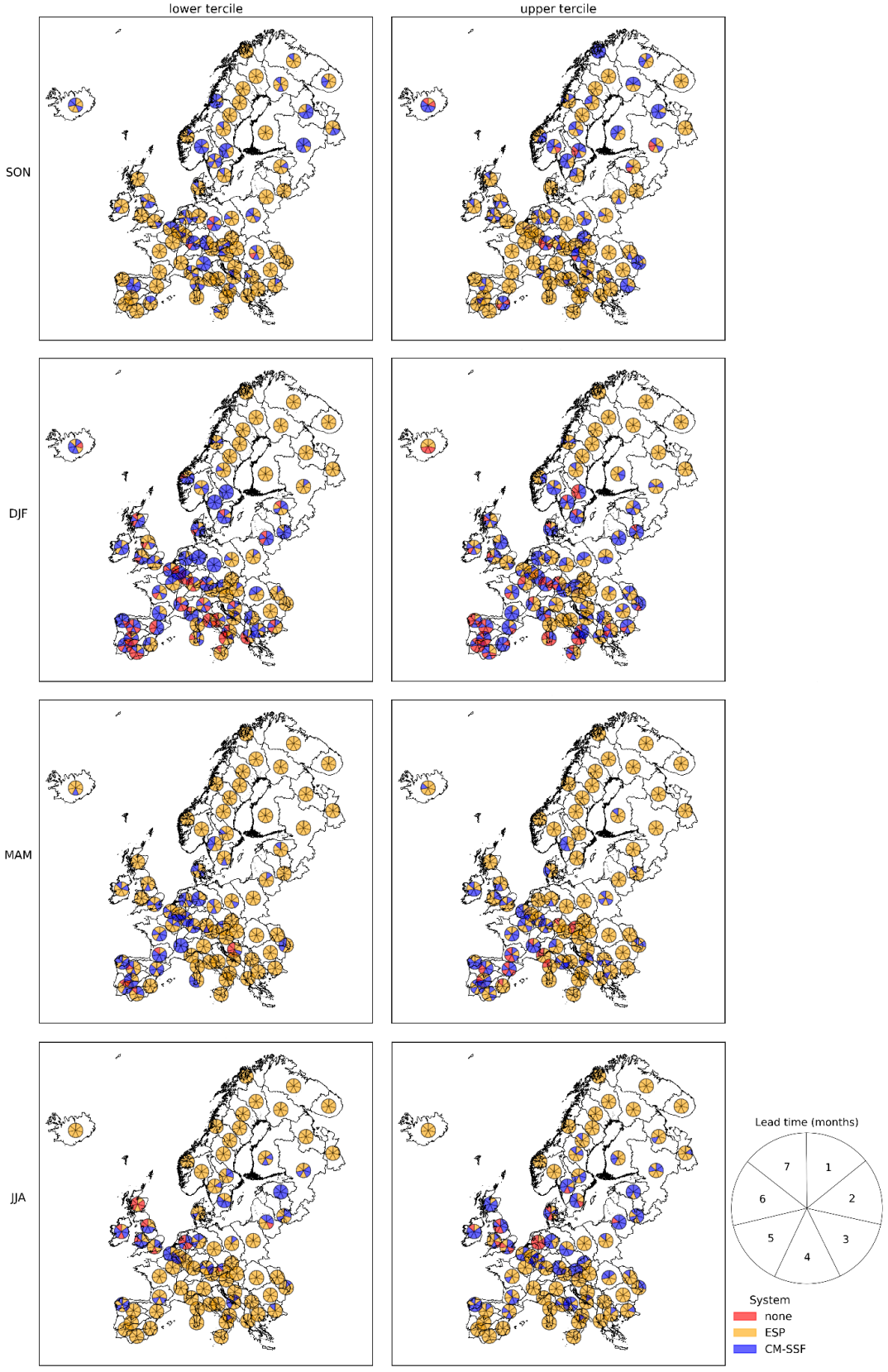

lower tercile
upper tercile
SON
DJF
MAM
JJA
Lead time (months)
1
2
3
4
5
6
7
System
none
ESP
CM-SSF

**Figure 5: Maps of the best system (as measured with the ROC score) for all four seasons (SON, DJF, MAM and JJA from the top to the bottom row) and the lower and upper simulated streamflow seasonal terciles (left-most and right-most columns respectively) in each region. The pie charts display the best system for each lead time (i.e. one to seven months), as shown on the example pie chart on the bottom right of this figure. There are three possible cases: 1) neither the ESP nor the CM-SSF is skilful (red colours), 2) the ESP is skilful and better than the CM-SSF (yellow colours), and 3) the CM-SSF is skilful and better than the ESP (blue colours).**

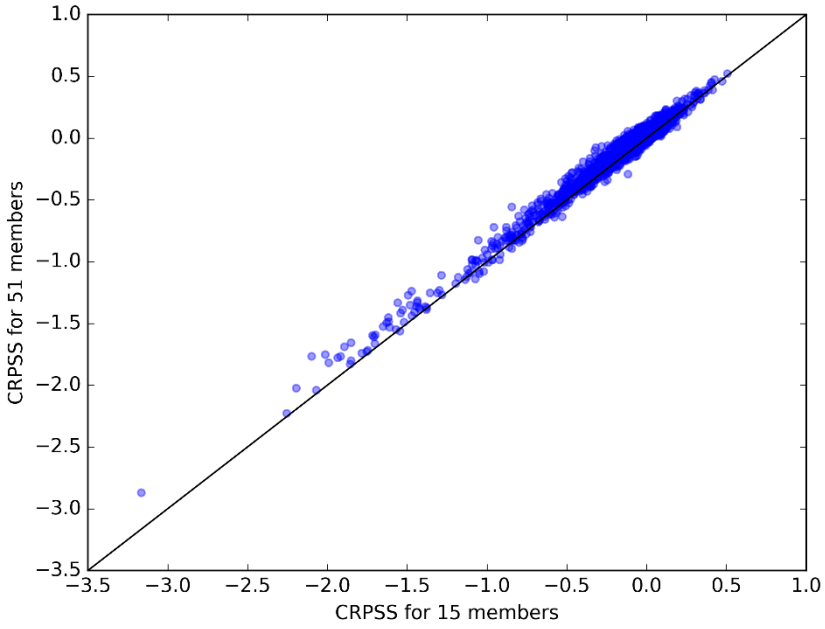



**Figure 6: CRPSS calculated for the CM-SSF against the ESP (benchmark) for hindcasts made on the 1st of February, May, August and November, all lead times (i.e. one to seven months) and all 74 European regions. The x-axis [y-axis] contains the CRPSS calculated from 15 [all 51] ensemble members of the CM-SSF.**