**Summary**

**This is a thorough and well-conceived investigation of the performance of seasonal streamflow forecasts generated by the EFAS system across Europe. The methods and verification metrics are robust and support the authors' conclusions. The manuscript is logically structured, concise and well-written, and in general I found it a very interesting read. The work sits squarely within the subject area of the special issue. I recommend that it be published after the authors consider a few minor issues for revision.**

We thank the reviewer for their very positive review and the constructive comments. Below are our responses to these comments.

**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.**

Overall, we agree with the fact that the ESP is a very good benchmark and therefore a harder one to beat than, for example, climatology. One of the main reasons for choosing the ESP as a benchmark was 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).

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.

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)”.

**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.

**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.

**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.”

**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.”

**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)”.

**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.

**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)”.

**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”.

**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”.

**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.

**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).

**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.

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

**References**

Bazile, R., Boucher, M.-A., Perreault, L., and Leconte, R.: Verification of ECMWF System4 for seasonal hydrological forecasting in a northern climate, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-387, in review, 2017.

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. Sci., 21, 4681-4691, https://doi.org/10.5194/hess-21-4681-2017, 2017.

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

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

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.

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. Sci. Discuss., https://doi.org/10.5194/hess-2017-293, in review, 2017.

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. Sci., 21, 3915-3935, https://doi.org/10.5194/hess-21-3915-2017, 2017.