Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2017-610-RC1, 2017
© Author(s) 2017. This work is distributed under
the Creative Commons Attribution 4.0 License.

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Skilful seasonal forecasts of streamflow over Europe?" *by* Louise Arnal et al.

Anonymous Referee #1

Received and published: 25 October 2017

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.

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 skillful (or may even be negatively skillful). 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 unsually 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.

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.

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 5.

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.

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.

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.

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.

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?

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

P6 L5 "...The sharpness should not be looked at in isolation and 5 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).

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.

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

Typos/grammar

P2 L3 "...hydrological conditions and land surface memory, as key drivers..." Delete comma

P8 L30 "...capable to predict..." should be "...capable of predicting..."

References

Greuell et al. 2017 'Seasonal streamflow forecasts for Europe' I & II, HESS special issue on Sub-seasonal to seasonal hydrological forecasting

Laio F, Tamea S. 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrology and Earth System Sciences 11: 1267-1277. DOI: 10.5194/hess-11-1267-2007.

Renard B, Kavetski D, Kuczera G, Thyer M, Franks SW. 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research 46: W05521. DOI: 10.1029/2009wr008328.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-610, 2017.