Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-604-RC2, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Seasonal streamflow forecasts for Europe – II. Explanation of the skill" *by* Wouter Greuell et al.

Anonymous Referee #2

Received and published: 15 January 2017

This is Part II of a paper describing a new dynamical ensemble seasonal streamflow forecasting system for Europe, which uses meteorological forcing from a coupled prediction system in the VIC hydrological model.

Dynamical continental scale ensemble prediction systems are at the cutting edge of seasonal streamflow forecasting. The general aim of explaining the sources of predictability in different regions is interesting and worthwhile, and traditional ESP and reverse ESP methods are well-established techniques to do this. The paper is generally clearly written and I acknowledge the considerable effort the authors have put into producing this paper. In short, I think the forecasting system is interesting, as is the aim of investigating of sources of skill, and that it deserves ultimately to be published. However, in my view the metrics/methods used to assess prediction performance are too rudimentary, to the point where it is difficult to understand how the system per-

forms. I also had some reservations about their attribution of skill to climate change, the revESP method used here, the description of ESP. Accordingly, I believe the paper requires major revisions before it can be published.

General Comments

Some of my objections relate to both parts I and II of this paper (as part II relies heavily on part I), so the authors may wish to address them in both (or either) papers. Specifically, my major objections are:

1) The authors essentially rely on correlation between forecasts and observations as the major metric of performance. In my view they should not, but should use RPSS instead. The argument (made in the Part I paper) that correlations are 'easier to understand' than RPSS simply doesn't hold water in my opinion: skill scores that describe performance in relation to climatology (like RPSS) make it much easier to understand the value of the forecasting system (even against the 'pseudo observations' used in this paper) than correlations.

In addition, I do not agree with the authors' contention that the RPSS and correlations "are similar to a high degree". The theoretical differences between skill scores and correlations have been documented by Murphy (1988), who concluded: "...use of the correlation coefficient (or its square) may lead to substantial overestimation of forecasting performance" and that "...it is more appropriate to interpret the square of the correlation coefficient as a measure of potential skill than as a measure of actual skill" These differences appear to manifest in practice for WUSHP. As far as I can see, the only evidence the authors present to demonstrate that correlations and RPSS are similar for WUSHP is Figure 8 in the Part I paper (it shows forecasts for May at lead 2). I could not follow the method used to calculate the statistical significance of the RPSS values (please supply more details), but on the face of it the drop in significant performance from 76% of all grid cells (correlation) to 47% of all grid cells (RPSS) reported by the authors is substantial (i.e., 29% of cells appear to have changed from

being designated as 'skillful' to not skillful). The heat maps in Figure 8 of the Part I paper also show substantial divergences between correlation and RPSS. For example, RPSS values of less than zero skill are shown over much of Poland/Belarus/Ukraine, but this region exhibits high (and significant) correlations. Similar divergences between RPSS and correlations happen over Ireland, southern Spain, much of nothern Africa, eastern Germany, Greece and the Balkans, and substantial tracts of Italy, Romania and western Russia.

Given these differences, and the theoretical preferability of RPSS, I think the authors should replace correlations with RPSS as the major metric for skill throughout the paper (though note the comment #9 in 'Other Comments' about reference forecasts), and change their interpretations/conclusions accordingly. I also recommend that the authors use the word 'skill' in the sense more commonly (though admittedly not universally) used in the forecasting literature - i.e., skill is performance with respect to a reference forecast - rather than as a more general synonym for 'accuracy'.

2) The authors present an ensemble forecasting system without any explicit analysis of reliability in either the Part I or Part II paper. Reliability is a crucial property of any ensemble forecasting system (e.g., Mason and Stephenson 2008; Raftery 2016; among many others). The authors should quantify and discuss the reliability of their forecasting system, using established diagnostics of reliability (I particularly recommend the probability integral transform - see, e.g., Gneiting and Katzfuss (2014) - but attributes/reliability diagrams (Hsu & Murphy 1986) are also suitable for binary forecasts). The authors may choose to address this issue in the Part I paper, but it must be addressed somewhere.

3) No mention is made of cross-validation. Part I alludes to the need to calibrate the VIC model, and quantile mapping is applied. In addition, ESP experiments sample from different years (as they should - see comments #5 and #6, below). All these need to be robustly cross-validated to ensure forecast performance is not overstated (e.g., using leave-x-year out cross-validation). Please describe the cross-validation methods

employed.

4) In a number of instances the authors ascribe (or do not ascribe) fractions of skill to climate change by examining trends in data. There are a couple of issues here. First, the methods for detrending/tests of statistical significance of trends are not explained - so I do not know what is being detrended or how (in an Appendix is fine). Second, to do this analysis the authors assume that trends in data are somehow causally related to forecast skill (implicit in Figure 2 f-i, and their discussion of these figs). It is not clear to me why this should be so in all the cases presented - particularly for climate variables like temperature, about which the authors state "Skill in summer temperature is related to climate change occurring in both the observations and the hindcasts...". Because S4 forecasts are initialised by assimilating observations, it's reasonable to expect that the hindcasts have trends in them (induced by the initialisation) that are similar to observations. So it seems (I think) the authors are implying that the thermodynamical/dynamical responses of S4 to initial conditions may not reflect thermodynamical/dynamical changes induced by climate change. Again, I do not know why this would be (after all, climate models are used routinely for future projections, so they are assumed to work similarly in future as now). Please first explain how trends could impact predictability before drawing any conclusions on how climate change driven trends in data influence skill.

5) I am not very familiar with the VIC model, but I would assume that some of the internal states could be highly correlated/anticorrelated (as in most hydrological models). Using states averaged from different periods - as done here for the revESP method - could destroy these correlations, and could lead to unrealistically poor simulations. What I think the authors should have done is generated an ensemble of states by forcing VIC with resamples observations from a number of years (as in the original revESP) and then forced each of these with the ensemble of climate forecasts. This would lead to more ensemble members (15* number of years sampled), but would allow much more satisfacory diagnosis of the contribution of meteorological forcing to overall skill.

This means the revESP experiments would have more ensemble members than the other experiments, but I think the benefits of this approach outweigh potential artefacts arising from different ensemble sizes.

6) The method for the ESP is described as follows: "We did not select atmospheric forcings from observations (e.g. WFDEI), which is the strategy employed in most published ESP experiments. By selecting the forcing from the S4 hindcasts, the ESP experiments remain as close as possible to the Full Hindcasts." I assume the authors have used some way of sampling from different years, but the way this is written makes it possible to interpres this as if the ESP forecasts could simply be analogous to hindcasts. (If the authors have used 'ESP' in the latter way, 1) they should not call it ESP and 2) they cannot claim to isolate the source of skill, which is their aim.) The authors need to clarify what they have done here. If, as I've assumed, they have sampled from hindcasts, this is essentially a new method (which is a good thing). In this case the authors need to spell out their method clearly, including sampling strategies for ensemble members, etc.. They may also like to give it a new name - e.g. HESP for 'hindcast ESP', or similar.

7) As the authors are introducing a new system, it needs to be put in the context of existing operational and experimental prediction systems. The introduction does not really do this at present. For example, the authors could note that statistical systems can much more easily be configured to produce reliable ensemble forecasts (e.g. Madadgar and Moradkhani 2013; Wang and Robertson 2011). In addition, other experimental dynamical forecast systems have attempted to explicitly deal with problems related to reliability (Yuan 2016; Bennett et al. 2016). (While Yuan (2016) is mentioned, the authors do not note a crucial difference between this system and WHUSP - that is, Yuan's hydrological post-processing step that attempts to ensure reliable ensembles.) A paragraph explaining how the WHUSP approach compares to existing systems, including any differences in the aims of WHUSP compared to other systems, would be a useful addition to the introduction. (For example, it is fine to validate against pseudo-

reality if this is how the system is to be used in operation, but if the aim of WHUSP is to predict actual inflows to reservoirs then it should be validated against observations.)

8) Quantile mapping has several limitations as a method for post-processing ensemble forecasts - in particular that it does not correct for errors in reliability because it ignores information that is available from correlations between hindcasts and observations (see Wood and Schaake 2008; Madadgar et al. 2014). A statistical calibration (e.g. Gneiting et al. 2005; Schepen et al., 2014) is probably preferable. This should be acknowledged somewhere here.

9) It was not clear to me what was used as the reference forecast when CRPSS was calculated. I suggest a cross-validated measure of climatology that varies with month (e.g. an ensemble of resampled historical streamflow, or similar). Please clarify.

Specific (minor) comments

Page 2

Line 18 'The term ESP refers to...hindcasts'. Not only hindcasts - ESP systems are widely used to produce forecasts

Line 20 'reference simulation'. As noted above, 'reference' forecasts in forecasting literature are frequently used to denote a benchmark for performance. I suggest using a different term than 'reference simulation' here, because it is not really being used as a reference. Suggest simply 'simulation'

Line 21 'identical'. ESP experiments are often cross-validated (depending on the aims of the experiment), so forcings may not be 'identical'.

Page 4

Line 21 'This is not surprising'. It's not surprising if you consider correlation as synonymous with skill. As I argue above, I don't believe this is justified. As Murphy (1988) points out, skill scores have components that consider, e.g. conditional and uncon-

ditional biases, which correlations ignore. Clearly bias correction will influence these aspects of skill.

Page 5

Lines 1-2 'By selecting the forcing from S4 hindcasts, the ESP experiments remain as close as possible to the Full Hindcasts'. I do not understand what the authors mean by 'select' here. It could imply that the authors simply used S4 forecasts (i.e. the same as the 'Full Hindcasts'), but this does not make sense (see comment #5). Please clarify.

Lines 14-15 '...which is important since ensemble size affects skill metrics'. I would say what's of more concern is that a small ensemble of 15 members is likely to mean that your skill metrics are subject to considerable sampling uncertainty.

Page 7

Line 5 '...is linked to climate change...' change to '...could be linked to climate change...'

Line 5 '...by detrending the data...' A brief summary of what is being detrended, the detrending technique and the trend significance test is needed (either a description, which could go in an appendix, or a reference)

Line 6 '...is insignificant across most of the domain...' I assume the trends were analysed only for the hindcast periods. Please note this somewhere

Line 35 '(revESP) always causes much less significant skill' I accept that this will probably be true, but could this result be partly caused by the way in which that the model has been initialised? (i.e with states that may not be correlated - see comment #4)

Page 8

Lines 1-3 'We explain the enhanced skill in runoff by an indirect effect of the skill of the precipitation forcing in the first lead month, which gradually adds some skill to the model states of soil moisture and snow.' I undestand what the authors are getting at here, but I think this is poorly phrased. The forcing doesn't really 'add skill' to the

model states. It's simpler to say that runoff forecasts are generally more skillful than climate forecasts because they aggregate skill from initial conditions and skill from meteorological forcings.

Line 31-32 '... where the first skill occurs...' What is meant by 'first skill'?

Page 9

Line 20 'skill in evapotranspiration' - This section is somewhat out of character with the first part of the paper. I am not suggesting it is not important work, but it perhaps would have been better in its own paper.

Line 21 '...hindcasts of evapotranspiration have intrinsic value...' suggest '...hindcasts of evapotranspiration have value indepedent of streamflow forecasts...'. Also, please provide a brief summary somewhere of how VIC calculates ET.

P 10

Line 24-26 'Initial conditions of snow and/or soil conditions lead to skill in the temperature hindcasts of the climate model (S4) and initial conditions of snow and soil moisture lead to skill in the evapotranspiration hindcasts of the hydrological model (VIC).' Is there generally agreement in the VIC snow/soil moisture states and the (I presume) observations assimilated by S4?

P11

Line 18 '...their semi statistical forcing is more skilful than the S4 forcing...'. Is is also possible that your hydrological model is more efficient, thereby giving you relatively more skill from initial conditions?

Line 28 '...to what extent they are due to a lack of interannual variability in the processes that eliminate the skill?' As discussed in comment #1, this would be straightforward to answer if you used skill scores calculated against a suitable climatological reference forecast (comment #9)

Line 32 'which is an important skill-eliminating factor'. This won't show up in correlations.

P12

Line 25 'The logical answer is "yes" but such a strategy should then be reconsidered regularly'. As noted in comment #8, statistical calibration methods are available to post-process climate outputs. One of the benefits of these methods is that in the absence of demonstrable forecast skill, they return climatology forecasts. So a more effective alternative than using two forecasting systems might be to use calibrated S4 forecasts as forcing, which would then effectively give an ESP-like forecast when skill isn't there. See Peng et al. (2016) for an example of this applied to S4.

Line 36 'This study demonstrates the power of using pseudo-observations for verification.' I agree, but the usefulness of pseudo-reality also depends on the ultimate aims of the forecasting system. If the aim of the system is to give accurate streamflow forecasts where quantities matter (e.g., forecasting inflows to reservoirs), then the system must be verified accordingly (i.e., against gauged streamflows). I think the authors should acknowledge this.

Typos/grammar

Page 1

Line 6 delete 'among others'

Page 4

Line 37 '...of the an...' delete 'the'

Page 8

Line 27-28 '...snow stops to be available...' change to 'snow is not available'

Page 9

Line 34 delete 'quite'

Page 10

Line 11 '...April as...' make '...April at...'

Line 30 change '...Mediterranean in due...' to '...Mediterranean is due...'

Page 12

Line 5-6 '...two hotspot region.' Make it 'region'

Line 33 '...not the case practical applications.' Add 'in'

References

Bennett, J. C., Q. J. Wang, M. Li, D. E. Robertson, and A. Schepen (2016), Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resources Research, 52, 8238–8259, doi: 10.1002/2016wr019193.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Monthly Weather Review, 133(5), 1098-1118, doi: 10.1175/mwr2904.1.

Madadgar, S., and H. Moradkhani (2013), A Bayesian Framework for Probabilistic Seasonal Drought Forecasting, Journal of Hydrometeorology, 14(6), 1685-1705, doi: 10.1175/jhm-d-13-010.1.

Madadgar, S., H. Moradkhani, and D. Garen (2014), Towards improved post-processing of hydrologic forecast ensembles, Hydrological Processes, 28(1), 104-122, doi: 10.1002/hyp.9562.

Murphy, A. H. (1988), Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient, Monthly Weather Review, 116(12), 2417-2424, doi: 10.1175/1520-0493(1988)116<2417:ssbotm>2.0.co;2.

Peng, Z., Q. J. Wang, J. C. Bennett, A. Schepen, F. Pappenberger, P. Pokhrel, and Z. Wang (2014), Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China, Journal of Geophysical Research (Atmospheres), 119, 7116–7135, doi: 10.1002/2013JD021162.

Raftery, A. E. (2016), Use and communication of probabilistic forecasts, Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(6), 397-410, doi: 10.1002/sam.11302.

Schepen, A., Q. J. Wang, and D. E. Robertson (2014), Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs, Monthly Weather Review, 142(5), 1758-1770, doi: 10.1175/mwr-d-13-00248.1.

Wang, Q. J., and D. E. Robertson (2011), Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, Water Resources Research, 47, W02546, doi: 10.1029/2010WR009333.

Wood, A. W., and J. C. Schaake (2008), Correcting Errors in Streamflow Forecast Ensemble Mean and Spread, Journal of Hydrometeorology, 9(1), 132-148, doi: 10.1175/2007jhm862.1.

Yuan, X. (2016), An experimental seasonal hydrological forecasting system over the Yellow River basin – PartÂă2: The added value from climate forecast models, Hydrology and Earth System Sciences, 20(6), 2453-2466, doi: 10.5194/hess-20-2453-2016.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016.