Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-603-RC2, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations" *by* Wouter Greuell et al.

C. Prudhomme (Referee)

chrp@ceh.ac.uk

Received and published: 9 February 2017

General

The paper is the first of 2 companion papers on a pan-European seasonal streamflow forecasting system. This paper focuses on the verification of the re-forecast for a 30-year period (1981-2010).

Streamflow forecasting beyond medium range is still a relatively new area of research in Europe, and has received more attention in the past few years, following the availability of seasonal climate re-forecasts. Skilful hydrological forecasts at monthly to seasonal lead time would have great potential use in Europe as it would help planning

and management of water resources for a huge variety of sectors including transportation, agriculture, public and domestic water supply or energy. Whilst the skill of dynamic rainfall forecasts is relatively limited at lead times over 10 days in temperate climates such as Europe, the existence hydrological memory due to catchment storage raised the question of potential higher skill in hydrological seasonal forecast than in its climate forcing data. As such, the paper addresses a topical subject with a large readership interest. I have however some concerns about some of the analyses undertaken here, detailed below. I hence suggest a major revision.

The streamflow forecasting system developed and used in this paper relies on two major sources of information and tool: 1) climate forcing data, here based on the ECMWF System 4 re-forecasts; and 2) a gridded hydrological model that transforms the weather signal into runoff and routed discharge. Inherent to any modelling exercise, simulations and re-forecasts are likely to be associated with bias and errors.

The authors run a gridded hydrological model forced by observed climate for 1 month, as spin-up to set-up initial conditions, and run the model with re-forecast climate forcing. They then evaluate the skill of the re-forecasts by comparing the results with 1) hydrological simulations forced by observed climate (runoff and routed discharge; called 'theoretical skill'); 2) observed discharge (called 'actual skill'). For actual skill, they use discharge time series from the GRDC and EWA database, and match the location of the river gauges with the routed network used in the model (at a 0.5°x0.5°resolution, i.e. $\sim$ 50km) so that gauged flows can be compared with the correct modelled discharge. Three metrics are used for the theoretical skill assessment, but most discussion is based on correlation coefficients, also applied to actual skill. The seasonal variation of the spatial distribution of the theoretical skill is described and compared for runoff and discharge, mainly for a 2-month lead time. Overall pan-European theoretical and actual skill compared for 2 classes of catchment size, and some causes of degradation between theoretical and actual skill discussed, but not formally tested.

Whilst the findings of pan-european hydrological seasonal forecasting skill are really

relevant, I have some reservation regarding some methodological decisions and interpretations presented in the paper, detailed below.

- Actual skill analysis. The analysis must be better justified, and the discussion strengthened. Below are some points that need to be added to the paper:

o Is simulated discharge comparable to actual discharge? There is no data assimilation at the beginning of the forecast to reduce potential bias in the simulated discharge. So the hydrological re-forecasts include both hydrological modelling errors and climate forcing errors, without any attempt to reduce the former.

o Is the catchment matching exercise working? The hydrological model has a relatively coarse resolution, and a catchment area error of up to 15% (for large catchments) is deemed acceptable [the choice of this threshold should be justified]. For small catchments, there is no attempt to scale the discharge from the hydrological model scale to the gauged catchment scale. This could introduce some discrepancies between simulated and observed discharge. In fact p8 l3-4, the authors do state that '[the] small basins (. . .) are generally smaller than the spatial resolution of the simulations'

o Is the hydrological model performance influencing the actual skill results? Poor hydrological model performance will introduce errors for both initial states and re-forecasts. One hypothesis is for 'actual skill' to be much lower for seasons and locations where the hydrological model is known not to reproduce well the hydrological processes. Comparison of hydrological model performance and actual skill is necessary for a meaningful interpretation of the results. This is only mentioned briefly in the discussion (2.5 lines) as second point (p9 l31-33). This should be the first point of the analysis when regarding actual skill.

- Re-forecast simulations

o Is the spin-up period long enough? It is not clear what actual spin up is used, with 1-month spin-up period suggested (p3 l29), but this sounds really short compared to

expected storage in some parts of Europe (e.g. snow pack in high latitude/ high elevation and/or groundwater storage in large aquifers).

- General methodology

o How are the catchments classified as small/ large? There is no surface area mentioned, and not physical justification, but size is the only physical measure used to attempt explaining the difference between theoretical and actual skill

o What is the justification for the non-calculation of skill metrics? (p5l23-24). In particular, zero flow simulations can be extremely important to depict droughts. Why excluding them?

o How is a skilful forecast defined? (p5 l37-38; p6 l1-2) What is the threshold used to define a re-forecast as 'skilful'? Is this based on statistically significance test? Is it the value of 0.31 quoted in caption fig 1? This needs to be made clearer within the text

o Human influence analysis. This is fully based on the assumption that LPJmL has identified and reproduces accurately all the human interventions, and the derived Amended Annual Proportional Flow Deviator is a realistic representation of the degree of influence. This is a strong assumption that needs to be caveated in the text. This modelling exercise needs to be described in the methods section and not so late in the paper (p7 l34-36)

- Analysis/ interpretation

o Influence of catchment size on theoretical vs actual skill (p8 l4-17). I found the analysis difficult to follow, the paragraph confusing, and the language used is inappropriate 'apparent difference in (. . .) skill (. . .) can be blamed almost entirely to the geographical distribution of stations'. What does 'this results holds for the cells with observations' mean? Is the difference between 'large basins' skills (0.396) and 'small basins' skills (0.384) significant? Is this to be linked with the scale of the hydrological modelling? The analysis would be more thorough if conducted by looking at relationships with

catchment sizes, rather than dividing the sample in 2 categories. It also needs to be linked with the modelling performance.

o Section 3.4 (p8). Is this conducted on pseudo observations? Why is this not after section 3.2? What is the implication of the findings? Can a physical explanation be given? Can the authors recommend skill metrics following their analysis?

o Discussion (p9-10). I found it unclear and difficult to follow, and some description of methods (model calibration technique) don't fit well (this should be in methods). The authors here describe some hypotheses for the difference between theoretical and actual skill: this should come at the beginning of the paper, and being tested within the study. Moreover, the analysis between theoretical and actual skill is short and not very thorough, yet is discussed at length; this does not reflect well the study. Some points are not clear (e.g. p9 l26-30; p9 l39-42)

o Statements not justified. There is a lack of evidence of the authors' claim that 'optimisation of the model system could, and would in many case, lead to a degradation of the theoretical skill'. What is the reason for that? What is the evidence? Have the authors conducted a sensitivity analysis? I agree that perfect theoretical skill does not adequate with perfect re-forecast, when main processes are not accounted for in the models. But the whole section needs careful re-wording, and better scientific justification, references, or suggestions for further analysis for verification of the hypotheses.

Main points of suggested improvement

Science

- There is no information on the hydrological model performance, albeit it is written to be 'on average across all basins considered, more or less in the middle ranking of the five models' [p3l39-40]. This is not enough and does not provide any information of the actual performance (it could be middle ranking of an ensemble with very low skill). Reference of a paper is not enough in this case. This is critically important when

the re-forecast skills are compared with what the authors call real- observations, as it would be expected that lower hydrological modelling performance would result in lower skill in reproducing the real observations.

- There is not enough discussion on the role of initial conditions, hydrological memory and catchment storage that can bring predictability: catchment storage could include groundwater, lakes, and snow pack. At the very least, reference to some of the findings of part 2 could be made.

- There is a lot of discussion about the quality of measurements and their implication on lower actual skill, and much less on modelling error. I found this out of proportion.

- Current conclusion is a summary of the research. I would expect the discussion to be opened to future research and application.

- The reference to the companion paper (page 2) is very limited, and it is difficult to see the link between both. At least the conclusions could be brought in the discussion, rather than exposed in the introduction and not referred to later onto justify the writing up of the study in 2 parts.

Structure

- The title does reflect the bulk of the paper. The analysis of 'real-discharge' is only done in section 3.3, out of 4 analysis sections.

- The structure is not logic: 3.1, 3.2 and 3.4 all analyse the results in a 'pseudo-observations' [modelled] world whilst 3.3 looks at the results in 'real- observations' world.

- Description of the model set-up/ calibration is given in the discussion (p10 l29-33), but this should be in the methods section when the model is introduced

Other points

Science

- The explanation of matching gauges locations with the 0.5 grid needs to be improved

Structure/ description

- Introduction: Most of the introduction is dedicated to the methods, data and tools used in the paper, and is not a review and discussion of the state of the art, with a judgment of the conclusions obtained from previous studies, and how to move forward. A typical example is p2 l9-15, with a list of papers without any discussion, and a description of some of the analysis, and even a discussion of the results, which should not be in introduction. I found this very confusing. The whole section needs to be greatly improved, with a more traditional layout of state of the art, research gaps identified, and then at the end aims of the paper, without details of the methods and tools used.

- Section 3.1: Inconsistency in figure references; first sentence of page 6 does not describe what fig 2 shows.

- Figure 3 is excellent.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016.