Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Open Access
Discussions

Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations

Wouter Greuell, Wietse H. P. Franssen, Hester Biemans, Ronald W. A. Hutjes

Water Systems and Global Change (WSG) group, Wageningen University and Research, Wageningen, NL 6708 PB Wageningen, Netherlands

Correspondence to: Ronald Hutjes (ronald.hutjes@wur.nl)

Abstract. Seasonal predictions can be exploited among others to optimize hydropower energy generation, navigability of rivers and irrigation management to decrease crop yield losses. This paper is the first of two papers dealing with a model-based system built to produce seasonal hydrological forecasts (WUSHP: Wageningen University Seamless Hydrological Prediction system), applied here to Europe. The present paper presents the development and the skill evaluation of the system. In WUSHP hydrology is simulated by running the Variable Infiltration Capacity (VIC) hydrological model with forcing from bias-corrected output of ECMWF's Seasonal Forecasting System 4. The system is probabilistic. For the assessment of skill, we performed hindcast simulations (1981-2010) and a reference simulation, in which VIC was forced by gridded meteorological observations, to generate initial hydrological conditions for the hindcasts and discharge output for skill assessment (pseudo-observations). Skill is analysed with monthly temporal resolution for the entire annual cycle. Using the pseudo-observations and taking the correlation coefficient as metric, hot spots of significant skill in runoff were identified in Fennoscandia (from January to October), the southern part of the Mediterranean (from June to August), Poland, North Germany, Romania and Bulgaria (mainly from November to January) and West France (from December to May). The spatial pattern of skill is fading with increasing lead time but some skill is left at the end of the hindcasts (7 months). On average across the domain, skill in discharge is slightly higher than skill in runoff. This can be explained by the delay between runoff and discharge and the general tendency of decreasing skill with lead time. Theoretical skill as determined with the pseudo-observations was compared to actual skill as determined with real discharge observations from 747 stations. Actual skill is mostly and often substantially less than theoretical skill, which is consistent with a conceptual analysis of the two types of verification. Qualitatively, results are hardly sensitive to the different skill metrics considered in this study (correlation coefficient, ROC area and Ranked Probability Skill Score) but ROC areas tend to be slightly larger for the Below Normal than for the Above Normal tercile.

1 Introduction

Society may benefit from seasonal hydrological forecasts, i.e. hydrological forecasts for future time periods from more than two weeks up to about a year (Doblas-Reyes et al., 2013). Such predictions can e.g. be exploited to optimize hydropower energy generation (Hamlet et al. 2002), navigability of rivers and irrigation management to decrease crop yield losses. In this paper we will introduce WUSHP (Wageningen University Seamless Hydrological Prediction system), a dynamical (i.e. model-based) system (see Yuan et al., 2015) that was built to produce seasonal hydrological forecasts. It will be applied to Europe. The usefulness of the system depends on the level of its skill and the paper will therefore describe the system and then focus on the determination of its skill. The usual method of assessing skill of predictive systems is by analysing hindcasts, a strategy that will be adopted here as well.

It is quite common in seasonal hydrological forecasting (e.g. Shukla and Lettenmaier, 2011, Singla et al., 2012, Mo and Lettenmaier, 2014, and Thober et al., 2015) but also in medium range forecasting (Alfieri et al., 2014) to determine prediction skill by comparing the hindcasts with the output from a reference simulation. A reference simulation is a simulation made with the same hydrological model as the hindcasts, except that the forcing is taken from meteorological

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

observations or from a gridded version of meteorological observations. The reference simulation can best be regarded as a simulation that attempts to make a best estimate of the true conditions (in terms of e.g. discharge, soil moisture and evapotranspiration), using the modelling system. We will refer to the output of such a reference simulation as "pseudo-observations" (" true discharge" in Bierkens and Van Beek, 2009; "synthetic truth" in Shukla and Lettenmaier, 2011; "reanalysis" in Singla et al., 2012; "a posteriori estimates" in Shukla et al., 2014). Pseudo-observations have the advantages of being complete in the spatial and the temporal domain and to be available for all model variables. Also, they are suitable for the quantification of small sensitivities, e.g. to bias correction of the meteorological forcing, which would be hard to detect with real observations.

The downside of pseudo-observations is, of course, that they are not equal to real observations. In this paper we will determine the performance of the prediction system not only with pseudo-observations but also with real observations of discharge (like e.g. Koster et al., 2010, and Yuan et al., 2013) and compare the skill found with the two different approaches ("theoretical and actual skill", according to Van Dijk et al., 2013), which was earlier done by Bierkens and Van Beek (2009) and Van Dijk et al. (2013). Also, we will analyse conceptual differences between using pseudo- and real observations for verification. We will argue that the fact that the pseudo-observations are obtained with the same model as the hindcasts logically contributes to an overestimation of the skill when the pseudo-observations are used for verification.

During recent years, a number of systems for seasonal hydrological forecasts have been developed. Examples are the forecasting model suite for France described by Céron et al. (2010), the University of Washington's Surface Water Monitor (SWM; Wood and Lettenmaier, 2006) and the African Drought Monitor (Sheffield et al., 2014). Seasonal hydrological forecast systems for the entire continent of Europe are scarce. Thober et al. (2015) forced a mesoscale hydrological model (mHM) with meteorological hindcasts of the North American Multi-Model Ensemble (NMME) to investigate the predictability of soil moisture in Europe. Bierkens and van Beek (2009) developed an analogue events method to select annual ERA40 meteorological forcings on the basis of annual SST anomalies in the North Atlantic and then made hydrological forecasts with a global-scale hydrological model applied to Europe.

The hydrological hindcasts are produced by WUSHP by running the Variable Infiltration Capacity (VIC) hydrological model using bias-corrected output of hindcasts from ECMWF's Seasonal Forecast System 4 as meteorological forcing. The system is probabilistic. In addition, a reference simulation is carried out, in which VIC is forced by gridded meteorological observations (WATCH Forcing Data Era-Interim, i.e. WFDEI), with the aims of generating pseudo-observations and initial hydrological conditions. Details about WUSHP are provided in Sect. 2.

This paper aims to analyse to what extent WUSHP is able to predict runoff and discharge in Europe for lead times up to 7 months. The second aim is to get a better understanding of the effects of using pseudo-observations for the verification of hindcasts. We will start the result section by assessing theoretical skill of the runoff hindcasts (Sect. 3.1) and then proceed to theoretical skill of the discharge hindcasts and a comparison between theoretical skill of discharge and runoff (Sect. 3.2). Differences between theoretical and actual skill of discharge will be presented using our data (Sect. 3.3) and a conceptual analysis (Sect. 4). Additional figures are published in a supplement of this paper. In a companion paper (Greuell et al., 2016) we analyse the source of skill and the lack of skill discussed in the present paper, using two different methods. Firstly, skill in the forcing and other directly related hydrological variables like evapotranspiration are analysed. Secondly, a number of Ensemble Streamflow Prediction (ESP) and reverse-ESP experiments, which isolate different causes of predictability, are discussed. The main conclusions from the companion paper are that, in Europe, a) skill beyond the first lead month is almost exclusively caused by initial hydrological conditions and not by skill in the meteorological predictions and b) at most times and locations the initial state of soil moisture contributes more to skill than the initial state of snow.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

2 System, models, data and methods of analysis

2.1 The hindcasts and the reference simulation

We will here describe the version of the WUSHP that has been used to generate the hindcasts for the European continent. WUSHP consists of two simulation branches, namely a single reference simulation and the hindcasts themselves. In both branches, terrestrial hydrology is simulated with the Variable Infiltration Capacity model (VIC, see Liang et al., 1994), which runs on a domain extending from 25 W to 40 E and from 35 to 72 N, including 5200 land based cells of 0.5˚ x 0.5˚ (see maps in e.g. Fig. 1). In the reference simulation VIC is forced by a gridded data set of meteorological observations, namely the WATCH Forcing Data Era-Interim (WFDEI; Weedon et al., 2014). The reference simulation has a dual aim, namely to create the pseudo-observations for verification purposes and to create a best estimate of the temporally varying model state, which is then used for initialisation of the hindcasts. The second branch, the hindcasts, consists of three steps. Seasonal predictions of meteorological variables are taken from ECMWF's Seasonal Forecast System 4 (S4 hereafter). These are then corrected for bias using WFDEI again, here as the reference data set. Finally, VIC is run with the bias-corrected S4 hindcasts as forcing, taking initial states from the reference simulation. The whole system is probabilistic.

The S4 hindcasts used in the present study include 15 members, cover the period from 1981 to 2010 and consist of 7 month simulations initialised on the first day of every month (see Molteni et al., 2011 and http://www.ecmwf.int/en/forecasts/documentation-and-support/long-range/seasonal-forecast-documentation/user-guide/introduction). The ensemble is constructed by combining a 5-member ensemble analysis of the ocean initial state with SST perturbations of that state and with activation of stochastic physics.

The variables taken from the S4 hindcasts are daily values of precipitation, minimum and maximum temperature, atmospheric humidity, wind speed and incoming short- and long wave radiation since these are needed to force VIC. All of these variables were regridded with bi-linear interpolation from the 0.75 x 0.75˚ lat-lon grid of the S4 hindcasts to a 0.5˚ x 0.5˚ grid. Next, the quantile mapping method of Themeßl et al. (2011) was applied to bias-correct the forcing variables, taking the WFDEI as reference. For each variable and grid cell, 84 correction functions were established and applied by separating the data according to target month (12) and lead month (7).

VIC was run for the period of the S4 hindcasts (1981 – 2010) and in addition for spin-up periods. In the reference simulation two extra years (1979 – 1980) were simulated to spin up the states of snow, soil moisture and discharge. The hindcast simulations were initialised with states of soil moisture and snow from the reference simulation, so for these variables spin up was not needed. However, due to the set-up of the routing module of VIC, the state of discharge could not be saved and loaded. Hence to spin up discharge, each 7-month hindcast simulation was preceded by a one month simulation with WFDEI forcing. Simulations were performed on a 0.5˚ x 0.5˚ grid for all 15 members of the bias-corrected S4 hindcasts. Though the forcing consisted of daily values, the simulations were done with a three-hourly time step. Because snow may contribute significantly to the seasonal predictability of other hydrological variables, VIC was run with the option of elevation bands. This means that for each cell calculations were carried out at up to 16 different elevations, with the aim of simulating the elevational gradient of snow. Since the hindcasts cover 30 years with 12 dates of initialisation each and consist of 15 members, a total of 5400 hindcast simulations was carried out. VIC was run in naturalised flow mode meaning that river regulation, irrigation and other anthropogenic influences were not considered.

Simulations of historic discharge made with VIC and four other hydrological models were validated with observations from large European rivers by Greuell et al. (2015). For making seasonal predictions the most interesting results of that validation study are the skills of simulating interannual variability and the annual cycle. In both aspects VIC performed, on average across all basins considered, more or less in the middle of the ranking of the five models.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

2.2 Discharge observations

For the assessment of skill with real discharge observations, two data sets were acquired from the Global Runoff Data Centre, 56068 Koblenz, Germany (GRDC), namely the GRDC data set proper and the European Water Archive (EWA) data set. These data sets do not include any variable or parameter characterising the human impact. We converted these two data sets into two gridded versions with a resolution of 0.5˚ x 0.5˚ and a time step of a month. The first contained only observations for catchments larger than 9900 km^2 ("large-basins"). The second contained only observations for catchments smaller than the area of the grid cells ("small basins"). The subdivision enabled to investigate the effect of catchment size on skill.

In many cases the location of observation stations did not match with the corresponding river in the digital river network used in the routing calculations (DDM30, see Döll and Lehner, 2002). We corrected for this issue by matching the observations with the simulations by means of catchment size. The size of the model catchments ("model catchment area") was determined by the DDM30 network. The size of the catchments upstream of the observation station ("station catchment area") was taken from the meta data of the observations. First the station catchment area was compared to the model catchment area of the cell that is nearest to the station ("nearest model cell catchment area").

For large basins we then proceeded as follows:

- If the station and the nearest model cell catchment area differed by less than 15%, the observations were matched with the model calculations for the nearest model cell.
- Otherwise, the station catchment area was compared with the model catchment area of the eight cells surrounding the nearest model cell.
- The minimum of the eight differences was determined.
- If that minimum was less than 15%, the simulations for the corresponding cell were matched with the observations.
- Otherwise, the station was discarded.

For small basins we proceeded as follows:

- If the nearest model cell did not have an influx from any of the neighbouring cells, its simulations were matched with the observations.
- Otherwise, all of the eight neighbouring cells without influx were selected.
- Their simulations were averaged and matched with the observations.

We further discarded all observations with less than 21 years of data within the simulation period (1981-2010) for any of the months of the year. The final data sets contained 111 cells with observations for large basins and 636 cells with observations for small basins.

2.3 Methods of analysis

From the model output, consisting of daily means, monthly mean values were computed, which were then used for the analysis. The analysis is restricted to runoff, defined here as the amount of water leaving the model soil either along the surface or at the bottom, and discharge, defined here as the flow of water through the largest river in each grid cell. Discharge accumulates all runoff from cells that are upstream in the model river network, with delays due to transport inside cells and through the river network. Hence, whereas runoff represents only local hydrological processes, discharge aggregates hydrological processes occurring in the entire upstream catchment.

Instead of analysing skill per target season and/or for a number of consecutive lead months, we analysed skill per target and per lead month. The thus achieved higher temporal resolution of the skill metrics enables a more accurate determination of the beginning and end of periods of skill. Moreover, skill at a monthly resolution provides the possibility to determine the consistency of the skill where we define consistent skill as skill that persists during at least two consecutive target or lead months. In accordance with Hagedorn et al. (2005) we designated the first month of the hindcasts as lead month zero, so

target month number is equal to the number of the month of initialisation plus the lead month number. In discussing the results we will pay relatively little attention to lead month zero because seasonal prediction deals with forecasts beyond the first two weeks.

Three skill metrics (see Mason and Stephensen, 2008) were computed, namely i) the correlation coefficient between the observations and the median values of the simulations (shortly "correlation coefficient" or R), ii) the area beneath the Relative Operating Characteristics (ROC) graph (shortly "ROC area") and iii) the Ranked Probability Skill Score (RPSS). The ROC area is computed for three categories of the observations and hindcasts with an equal number of values, with the categories containing the one third highest, lowest and the remaining values ("above", " below" and " near-normal" category), respectively. The same subdivision of observations and hindcasts in terciles was made to compute the RPSS. All three skill metrics quantify, though in different ways, how well the ranking of the annual hindcasts matches the ranking of the observations. The ROC area indicates whether the forecast probability of an event (i.e. value falling in the considered tercile) is higher when such an event occurs compared to when not. The RPSS summarizes in a single number the skill of a forecast system to make correct forecasts of events falling in any of the defined terciles. Perfect forecasts have values of 1 for all three skill metrics. Climatological forecasts (forecasts that are identical each year) lead to values of 0 for R, 0.5 for the ROC area and 0 for the RPSS. Random forecasts were used to determine the significance of the metrics. In the case of the Ranked Probability Score (RPS), these random forecasts were generated by sampling randomly from the multinomial distribution with p = (1/3, 1/3, 1/3) and N = 15 (the number of ensemble members), which is the distribution of climatological ensemble forecasts. Each metric will be designated as significant for p-values less than 0.05.

To a large extent, we found that our results and conclusions are independent of the chosen metric. Hence and because among the three metrics the correlation coefficient is the easiest to understand, we will discuss results in terms of the correlation coefficient, which is in line with Doblas-Reyes et al. (2013). The sensitivity to the chosen metric will be discussed in Sect. 4.4.2.

Metrics will not be computed if observations or hindcasts consist for more than one third of zeros or one sixth of ties (i.e. equal values).

3 Results

In this section we present the skill of monthly mean values of hindcasted runoff and discharge. First, skill as determined with the pseudo-observations is discussed, starting with runoff (Sect. 3.1) and then continuing with a comparison between runoff and discharge (Sect. 3.2). Next, Sect. 3.3 analysis differences in skill found by using pseudo- and real observations for verification. In the first three sub-sections skill is measured in terms of the correlation coefficient between the observations and the median values of the simulations (R). Section 3.4 deals with results for other skill metrics.

3.1 Spatiotemporal variation of skill in *runoff* forecasts

Eighty-four maps of skill of the runoff hindcasts were drawn for all 12 months of initialisation and all 7 lead months. Two cross-cuts through that collection are shown in Figs. 1 (for a single initialisation month) and 2 (for a single lead month). The seven panels of Fig. 2 show the skill of the hindcasts initialised on April 1 as a function of lead time. Cells with an insignificant amount of skill are tinted yellow. In lead month 0, significant skill is found across almost the entire domain (99% of the cells). After the first lead month, the fraction of cells with significant skill gradually decreases to reach 16% at the longest lead time (lead month 6). This is more than expected for the case of completely unskilful simulations (5% of the cells), so at the end of the hindcast simulations significant skill that does not occur due to chance is still present in some regions. The general impression is that the pattern of skill does not move in space but that skill is fading, i.e. for individual grid cells R is mostly decreasing with increasing lead time.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

The twelve panels of Fig. 2 show the annual cycle of skill of the hindcasts for lead month 2. Consistent skill (persistent during at least 3 consecutive target months) is found in:

- Fennoscandia. Much skill is present during the entire year, except for November and December, and there is a dip in skill in April. On average across the entire region, skill reaches a maximum in May and June. Compared to the rest of the peninsula, there is generally less skill along the Scandinavian Mountain range.
- Poland and North Germany. The core period lasts from November to January, but it is extended with periods of less skill into October and the months from February to May.
- West France, more or less from Paris to Brittany and roughly from December to May.
- Romania and Bulgaria. The core as well as the whole period are the same as that for Poland and North Germany.
- The southern part of the Mediterranean region from June to August. The high amounts of skill are limited to the coastal parts of North Africa, Sicily, South Greece, Turkey, Syria and Lebanon.

These results can be compared to those of Bierkens and Van Beek (2009). They found maxima in predictability of winter discharge in North Sweden, Finland, the region between Moscow and the Baltic Sea, Romania and Bulgaria, and East Spain. For the winter there is crude agreement with the current study about North Sweden, Romania and Bulgaria but not about the other regions. For the summer, Bierkens and Van Beek (2009) compute maxima in skill for South Spain, Sardinia, West Turkey and South-west Finland. This pattern agrees to some extent with the locations of the summertime maxima in skill of the present study (most of Fennoscandia and southern part of the Mediterranean region).

Figure 3 displays a synthesis of Fig. 2 in the form of a map with the fraction of the 12 months of the year with significant skill for lead month 2. Many of the regions with very little or no skill all over the year are coastal regions (e.g. north coast of Spain), especially coastal regions on the western side of land masses (e.g. west coasts of Denmark, South Norway, Croatia and the British Isles), and mountain regions (e.g. the Alps, mountains in North Norway and Sweden and on the border of Poland and Slovakia). The entire British Isles exhibit very little skill, except for the east coast of Great Britain.

Figure 4 summarizes skill across the domain in terms of the fraction of cells with significant R for all initialisation and lead months. Overall there is a considerable amount of significant skill, with a minimum roughly from August to November and a maximum in May. For lead month 2 the fraction of cells with significant skill varies between 36% (September) and 76% (May). In all of the 84 combinations of initialisation and lead month, the theoretical value of no skill at all (5%) is exceeded. Individual curves show the loss of skill with increasing lead time. The exception is formed by hindcasts starting in November, December and January which gain skill when they progress from April to May. A graph similar to Fig. 4 but for the domain-averaged R instead of the fraction of cells with a significant R (not shown here) shows identical behaviour including the mentioned exception to the overall trend of skill decaying with lead time.

3.2 Spatiotemporal variation of skill in *discharge* forecasts

This sub-section compares skill for discharge with skill for runoff. The two maps of Fig. 5, which depict the skill in the runoff and the discharge hindcasts for July as lead month 2, show a high degree of similarity in terms of the patterns and the magnitude of the skill. The same holds for other target months and lead times. There are subtle differences though because rivers average the skill, or lack of skill, from the whole upstream part of their catchment. As a result, cells containing rivers with large catchments may contrast against adjacent cells if these contain rivers with a small, local catchment. Indeed, some downstream parts of large rivers stick out in the skill map for discharge but not in the skill map for runoff. An example in Fig. 5b are the reaches of the Danube along the Romanian-Bulgarian border, which show more skill than local small rivers in adjacent cells, because some upstream parts of the Danube have more skill than the region around the Romanian-Bulgarian border. An example that demonstrates the opposite is the downstream part of the Loire showing less skill than local small rivers, because upstream parts of the Loire have less skill than small, local rivers in the downstream part.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Domain summary statistics of skill also differ slightly between runoff and discharge. Figure 5c compares the annual cycle of the skill in discharge with the skill in runoff at five different lead times. Here we show the difference in the domain-averaged R instead of the fraction of cells with a significant R because in lead month 0 that fraction is close to one for both variables. In terms of the domain-averaged R, predictability is higher for discharge than for runoff for the first lead month. On average over the 12 months of the year, the difference is 0.049. We ascribe this result to the combined effect of the delay between runoff and discharge and the general tendency of decreasing skill with lead time. The curves for the different lead times in Fig. 5c show that the difference in skill between the two variables gradually disappears with increasing lead time (an annual average of 0.020 and 0.012 for lead months 1 and 2, respectively). This is compatible with the given explanation for the difference and the fact that the rate by which skill is lost gradually decreases with increasing lead time.

We finally analysed whether the difference in skill between discharge and runoff was a function of the size of the catchment (Fig. 5d). For the first lead month, when on average there is more skill in discharge than in runoff, the difference increases with the size of the catchment. Again this can be explained by the combination of the skill decaying with time and the delay between runoff and discharge, with the delay increasing with the size of the catchment. For longer lead times (lead months 2 and 4), when the domain-averaged difference in skill has become very small (panel c), panel d shows no effect of the catchment size. So, referring to the comparison between runoff and discharge in panels a and b for lead month 2, cases like the Danube (more skill than local rivers) and the Loire (less skill than local rivers) tend to cancel.

3.3 Verification of discharge with pseudo- and real observations

So far, all skill was determined by using the discharge generated with the reference simulation. i.e. with pseudo-observations. In this section, this “theoretical skill” will be compared with the skill determined with real discharge as observed at gauging stations (“actual skill”) from the GRDC and EWA data bases. Figure 6 compares the theoretical skill (panels b and d for large and small basins, respectively) with actual skill (panels c and e for large and small basins, respectively) for a single combination of a target month (May) with a lead month (2).

For the combination of target and lead month of Fig. 6, a substantial degradation in skill is found when the pseudo-observations are replaced by real observations. In terms of the fraction of cells with significant skill, the reduction is from 73 to 56 % for large basins and from 52 to 27 % for small basins and the domain-averaged R decreases from 0.48 to 0.33 for large basins and from 0.37 to 0.18 for small basins. Figure 7 compares actual with theoretical skill for all target months and two lead times by considering the domain-mean R. The reduction in skill occurs for all combinations of target and lead months and does not exhibit a clear annual cycle. On average across all target months and for lead month 2, the ratio of actual to theoretical skill is 0.667 (0.258 divided by 0.387) for large basins and 0.538 (0.156 divided by 0.290) for small basins. This can be compared to Van Dijk et al. (2013), who found a ratio of actual to theoretical skill of 0.54 for 6192 catchments worldwide in terms of the ranked correlation coefficient.

We investigated to what extent these results are affected by human interference, keeping in mind that the simulations are naturalized while the observations include human impacts to a variable but unknown degree. Human interference is expected to have a negative effect on actual skill and hence on the ratio of actual to theoretical skill. We quantified the human impact by performing two model simulations with the Lund-Potsdam-Jena managed Land (LPJmL) model (Rost et al., Schaphoff et al., 2013) that was operated at the same spatial resolution (0.5˚ x 0.5˚) and with the same river network (DDM30) as VIC. From the discharge output of a naturalized run and a run with reservoir operation and irrigation, the human impact at cell level was quantified by computing the so-called Amended Annual Proportional Flow Deviator (AAPFD, see Marchant and Hehir, 2002). Subsequently, we selected all discharge observations for large basins with an AAPF < 0.3, i.e. basins with a relatively small degree of human impact (about half of all 111 basins). For this selection the ratio of actual to theoretical skill was computed in terms of the domain mean R averaged across all target months and for lead month 2. We found a ratio of 0.686, which should be compared to a ratio of 0.667 for the entire set of large basins (see above). So, as expected the ratio is

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

larger for basins with less impact. However, since the difference between the two ratios is small we conclude that the effect of the combination of naturalised runs with observations that are affected by human interference contributes only little to the difference between actual and theoretical skill. A similar analysis was not applied to the collection of small basins with observations since these are generally smaller than the spatial resolution of the simulations.

Comparing skill for small basins with skill for large basins in Fig. 7, we notice two differences. Firstly, in terms of the domain mean R theoretical skill is higher for large basins than for small basins (0.39 and 0.29, respectively, for the annual mean and lead month 2). However, this result holds for the cells with observations. If all cells of the domain are considered, the difference almost vanishes. On average, all cells with an upstream catchment larger than 10000 km^2 have a mean R of 0.396 and all cells with an upstream catchment smaller than 2500 km^2 have a mean R of 0.384. So, the apparent difference in theoretical skill between large and small basins can be blamed almost entirely to the geographical distribution of the stations, with small basin stations being relatively more often located in regions with relatively little skill like Germany, France and the British Isles than large basin stations.

The second effect of the size of basins is that skill reduction is larger for small basins than for large basins. We speculate that this is due a combination of two effects. Firstly, there is more skill in simulations of historic streamflow in large basins than in small basins (Van Dijk and Warren, 2010). Secondly, as Van Dijk et al. (2013) demonstrated, the ratio of actual to theoretical skill is almost linear in the skill of simulating historic streamflow. Combining these two relationships confirms the relationship that we found, namely an increase in the ratio of actual to theoretical skill with basin size.

3.4 Results for other skill metrics

So far, skill was measured in terms of the correlation coefficient between the median of the hindcasts and the observations (R) only. This section compares those results with results in terms of other skill metrics. Figure 8 gives an example for one particular target month (May) and lead month 2. Panels a, b and c show the skill patterns for R, for the ROC area for Below Normal (BN) years and for the RPSS. The three patterns are similar to a large degree, noting that differences in colour are partly due to the interplay between differences in the domain-averaged magnitude of the skill metrics and the choice of the colour intervals. The pattern of the map of the ROC area for Above Normal (AN) years (not shown here) is also similar to the patterns of the three maps shown. The agreement that we find between the patterns of the different metrics is in accordance with a result mentioned in a global analysis of seasonal streamflow predictions by Van Dijk et al. (2013) who found high spatial correlation between the different skill metrics they used (among which R, the RPSS and the ranked correlation coefficient).

Though the different nature of the different metrics does not enable a quantitative comparison of the metrics, ROC areas for the different terciles can be compared among each other. For the particular combination of target month and lead month shown in Fig. 8, the domain-mean ROC area is largest for the BN tercile (0.75), slightly smaller for the AN tercile (0.73) and much lower for the near-normal (NN) tercile (0.58, not shown here; 0.5 corresponds to climatological forecasts). A similar tendency is found in the fraction of cells with a significant ROC area (69%, 63% and 21%, respectively). The fraction of cells with a significant value of the RPSS is 47%, which is somewhere between the fractions for ROC areas of the three terciles because the RPSS “mixes” the skill to make forecasts of events falling in all terciles. Finally, panels d presents a map of the difference between the BN and the AN ROC area. There is no clear regional pattern in this difference, i.e. coherent larger regions with clustered positive or negative values cannot be distinguished.

All of these results also hold for other combinations of target and lead month. Figure 9 compares the BN with the AN tercile in terms of the fraction of cells with a significant ROC area across all target and initialisation months. The main finding is that in all cases the fraction is larger for the BN than for the AN tercile. However, the two fractions tend to become equal (i) when they approach 1.0, (ii) when they approach the limit of no skill (5%) and (iii) during target months from October to January.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Open Access
Discussions

4 Discussion

For verification of the hindcasts two options were considered in this paper. We determined the skill of the hindcasts by comparing predicted discharge with the output of the reference simulation (the "pseudo-observations" leading to "theoretical skill") and with observations of real discharge ("real observations" leading to "actual skill"). To obtain a basis for understanding the differences in skill that we found, Fig. 10 presents a streamflow diagram of the three relevant physical systems, namely the real world and the model systems that generate the hindcasts and the pseudo-observations. In each system, confined in the diagram by a box, meteorological and initial conditions force and initialize hydrology, of which discharge is the relevant component here. There are two complications. First, the initial conditions themselves are generated by meteorological forcing during the spin up period, initial conditions at the beginning of the spin up period and hydrology. This is represented by the upper left branch in each box, omitting initial conditions at the beginning of the spin up period for simplicity. Second, due to measurement errors real observations of discharge generally differ from real discharge (Juston et al., 2014) as illustrated in the upper right corner of the figure.

The two essential questions are: 1) What are the conceptual differences between the physical systems that generate the pseudo- and the real discharge observations, i.e. between the model reference run and the real world. To answer this question, the components in the upper and the lower box of the diagram need to be compared. 2) What are the expected effects of these differences on skill, i.e. on the comparison with the hindcasts. To answer this question, the components that differ between the real world and the model reference run need to be compared with the model hindcasts. The rule then is that skill decreases with increasing disagreement between a component of the hindcast system and the corresponding component of one of the other systems. The following components (red text in diagram) differ between the real world and the model reference simulation, and their expected effect on skill are:

1) Real meteorology differs from the meteorology assumed in the reference simulation (WFDEI), both during the spin up period and during the hindcast period. During spin up, model reference run and hindcasts have identical meteorological forcing (namely WFDEI), which differs from real meteorology. Therefore, this difference is expected to lead to more theoretical than to actual skill. During the hindcast period, all three systems have different meteorological forcings. For cases with skill in the meteorological hindcasts, one would need to have an expectation about the agreement between the skilful hindcasts and reality, on one side, and the skilfull hindcasts and the WFDEI data set, on the other side. Unfortunately, we do not have a well-founded expectation about such a difference in agreement and, hence, we have no expectation about its effect on the difference between theoretical and actual skill. However, in Europe and beyond the first lead month almost all skill in the seasonal forecasts is due to the initial conditions, see the companion paper. Therefore, beyond the first lead month and in Europe differences in forcing during the hindcast period have a negligible effect on skill.

2) Models are imperfect, so model hydrology differs from real world hydrology. Hindcasts and the pseudo-observations are produced with the same model, so imperfections in model hydrology are expected to lead to more theoretical than actual skill.

3) In the real world a difference discharge observations differ from reality, i.e. a measurement error exists. There is no equivalent of this error in the model environment. Hence, as for differences 1) and 2) this difference is expected to lead to more theoretical than to actual skill.

Initial conditions are absent in this list of differences since in WUSHP they are not independent components but entirely determined by two components of the system listed above, namely meteorology and hydrology. Alternatively, initial hydrological conditions could be taken from observations or by assimilation of observations into model calculations. In that case, initial conditions would become an independent or semi-dependent component of the system. However, again, while model initial conditions would, of course, differ from real initial conditions, the two model system had identical initial conditions. Hence, again this difference would be expected to lead to more theoretical than to actual skill.

In summary, all of the conceptual differences between the generation of pseudo- and real observations, are expected to lead to more theoretical skill than actual skill, except for the difference in meteorology during the hindcast period, which has, in the case of Europe beyond the first lead month, a neutral effect, and otherwise an unknown effect.

A complication to this analysis is failure of the assumption implicitly made in the diagram that the catchment of the observation station and the model catchment are identical. This is not the case, see Sect. 2.2, so differences between observation and model catchment form an additional cause of differences between theoretical and actual skill. Again, this will favour theoretical skill with respect to actual skill since catchments are identical in the hindcasts and the reference simulation. In particular, differences in meteorological forcing between the catchment of the observation station and the model catchment reduce actual skill. Van Dijk et al. (2013) investigated this aspect by making simulations for Australia at different spatial resolutions and verifying with networks of observations with different spatial densities. They found that the resolution and perhaps the quality of the forcing data contributed to at least half of the difference between theoretical and actual skill.

Our data analysis broadly confirms that theoretical skill exceeds actual skill.

It is interesting to discuss what would happen in the utopian case that the system of the model reference run would converge with the real world, i.e. if model meteorological forcing and hydrology would approach perfection and if measurement errors would approach zero. Equality of the two systems would, according to the analysis above, lead to equality of theoretical and actual skill. However, we like to note that at the same time optimisation of the model system could, and would in many cases, lead to a degradation of the theoretical skill. Hence, theoretical skill is not equal to the maximum that could be accomplished if hydrological model and meteorological forcing during the reference simulation were perfect. An example proving this statement is a model that is imperfect because it accumulates too much snow. The model will do so both in the initial state of the reference simulation and the initial state of the hindcasts and since more snow leads, at some stage of the melting season, to more predictive skill, theoretical skill will be overestimated. A perfect model, accumulating less but realistic amounts of snow, would show less skill. Another example underlining the statement that theoretical skill is not the maximum that could be realized with a perfect model deals with predictive skill caused by interannual variations in the initial amount of soil moisture. A model that is imperfect because it overestimates the transport speed of soil moisture through the soil and the groundwater reservoirs will do so both in the reference simulation and the hindcasts. Predictive skill due to soil moisture initial conditions will then occur too early. Compared to the model that overestimates transport speed, a perfect model with smaller, realistic transport speed would yield less theoretical skill at the early lead times.

The version of VIC used in this study was calibrated by Nijssen et al. (2001) in a crude way, in the sense that they assumed no spatial variation of the parameters set by calibration within almost the entire European continent. Improving the calibration of VIC would be an obvious candidate for trying to improve the seasonal predictions discussed in this paper. This should lead to higher actual skill. However, the two examples discussed in the previous paragraph show that theoretical skill may actually, for certain locations, months of initialisation and lead months, decline due to the recalibration.

Many conclusions drawn from this work are valid at the scale of our domain and not necessarily at the scale of river basins. Only in some parts of our analysis, especially where we focused on the annual cycle of the skill (Fig. 2), regional patterns at a scale smaller than that of the domain were discussed. This was done in a qualitative way. In a future extension of this study, an objective method like cluster analysis could reveal regions where skill has similar signature. This could lead to an improved assessment of the physical and climatological factors that are responsible for the spatial variations in skill found in this study.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Open Access
Discussions

5 Conclusions

This paper is the first of two papers dealing with a model-based system built to produce seasonal hydrological forecasts (WUSHP: Wageningen University Seamless Hydrological Predictions). The present paper presents the development and the skill evaluation of the system for Europe, the companion paper provides an explanation of the skill or the lack of skill.

First, "theoretical skill" of the runoff hindcasts was determined taking the output of the reference simulation as "pseudo-observations". Using the correlation coefficient (R) as metric, hot spots of significant skill were found in Fennoscandia (from January to October), the southern part of the Mediterranean (from June to August), Poland, North Germany, Romania and Bulgaria (mainly from November to January) and West France (from December to May). There is very little or no significant skill all over the year in some coastal and mountain regions. The entire British Isles exhibit very little skill, except for the east coast of Great Britain. If the entire domain is considered, the annual cycle of skill has a minimum roughly from August to November and a maximum in May.

Runoff and discharge show a high degree of similarity in terms of the spatial patterns and the magnitude of the skill. However, when averaged over the domain and the year, predictability is slightly higher for discharge than for runoff for the first lead month (by 0.049 in terms of R). The difference then decreases with increasing lead time. These tendencies can be ascribed to the combined effect of the delay between runoff and discharge and the fact that skill decreases with lead time. We also found that the difference between discharge and runoff skill increases with the size of the catchment.

Theoretical skill as determined with the pseudo-observations was compared to actual skill as determined with real discharge observations. On average across all target months and for lead month 2, the ratio of actual to theoretical skill in terms of the domain-mean R is 0.67 (0.26 divided by 0.39) for large basins and 0.54 (0.16 divided by 0.29) for small basins. So, skill reduction due to replacing pseudo- by real observations is larger for small basins than for large basins.

Skill patterns for the different skill metrics considered in this study (correlation coefficient, ROC area and Ranked Probability Skill Score) are similar to a large degree. ROC areas tend to be slightly larger for the Below Normal than for the Above Normal tercile but not during target months from October to January.

References

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., & Salamon, P. (2014). Evaluation of ensemble streamflow predictions in Europe. Journal of Hydrology, 517, 913-922.

Bierkens, M. F. P., & Van Beek, L. P. H. (2009). Seasonal predictability of European discharge: NAO and hydrological response time. Journal of Hydrometeorology, 10(4), 953-968.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. (2013). Seasonal climate predictability and forecasting: status and prospects. Wiley Interdisciplinary Reviews: Climate Change, 4(4), 245-268.

Döll, P., & Lehner, B. (2002). Validation of a new global 30-min drainage direction map. Journal of Hydrology, 258(1), 214-231.

Greuell, W., Andersson, J. C. M., Donnelly, C., Feyen, L., Gerten, D., Ludwig, F., ... & Schaphoff, S. (2015). Evaluation of five hydrological models across Europe and their suitability for making projections of climate change. Hydrol Earth Syst Sci Discuss, 12, 10289-10330.

Greuell, W., W. H. P. Franssen, H. Biemans and R. W. A. Hutjes. Seasonal streamflow forecasts for Europe – II. Explanation of the skill. Submitted to Hydrol. Earth Syst. Sci.

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. Tellus A, 57(3), 219-233.

Hamlet, A. F., Huppert, D., & Lettenmaier, D. P. (2002). Economic value of long-lead streamflow forecasts for Columbia River hydropower. Journal of Water Resources Planning and Management, 128(2), 91-101.

Juston, J., Jansson, P. E., & Gustafsson, D. (2014). Rating curve uncertainty and change detection in discharge time series: case study with 44-year historic data from the Nyangores River, Kenya. Hydrological Processes, 28(4), 2509-2523.

Koster, R. D., Mahanama, S. P., Livneh, B., Lettenmaier, D. P., & Reichle, R. H. (2010). Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. Nature Geoscience, 3(9), 613-616.

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research: Atmospheres (1984–2012), 99(D7), 14415-14428.

Marchant, R., & Hehir, G. (2002). The use of AUSRIVAS predictive models to assess the response of lotic macroinvertebrates to dams in south-east Australia. Freshwater Biology, 47(5), 1033-1050.

Mason, S. J., & Stephenson, D. B. (2008). How do we know whether seasonal climate forecasts are any good?. In Seasonal Climate: Forecasting and Managing Risk (pp. 259-289). Springer Netherlands.

Mo, K. C., & Lettenmaier, D. P. (2014). Hydrologic prediction over the conterminous United States using the national multi-model ensemble. Journal of Hydrometeorology, 15(4), 1457-1472.

Molteni, F, Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., Vitart, F. (2011). The new ECMWF seasonal forecast system (System 4). ECMWF Technical Memorandum 656.

Nijssen, B., O'Donnell, G. M., Lettenmaier, D. P., Lohmann, D., & Wood, E. F. (2001). Predicting the discharge of global rivers. Journal of Climate, 14(15), 3307-3323.

Rost, S., Gerten, D., Bondeau, A., Lucht, W., Rohwer, J., & Schaphoff, S. (2008). Agricultural green and blue water consumption and its influence on the global water system. Water Resources Research, 44(9), doi 10.1029/2007WR006331.

Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., & Lucht, W. (2013). Contribution of permafrost soils to the global carbon budget. Environmental Research Letters, 8(1), 014026, doi:10.1088/1748-9326/8/1/014026.

Sheffield, J., Wood, E. F., Chaney, N., Guan, K., Sadri, S., Yuan, X., ... & Ogallo, L. (2014). A drought monitoring and forecasting system for sub-Sahara African water resources and food security. Bulletin of the American Meteorological Society, 95(6), 861-882.

Shukla, S., McNally, A., Husak, G., & Funk, C. (2014). A seasonal agricultural drought forecast system for food-insecure regions of East Africa. Hydrology and Earth System Sciences, 18(10), 3907-3921.

Singla, S., Céron, J. P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., & Vidal, J. P. (2011). Predictability of soil moisture and river flows over France for the spring season. Hydrology & Earth System Sciences Discussions, 8(4).

Themeßl, M. J., Gobiet, A., & Leuprecht, A. (2011). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. International Journal of Climatology, 31(10), 1530-1544.

Van Dijk, A. I., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., & Beck, H. E. (2013). Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. Water Resources Research, 49(5), 2729-2746.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. Water Resources Research, 50(9), 7505-7514.

Wood, A. W., & Lettenmaier, D. P. (2006). A test bed for new seasonal hydrologic forecasting approaches in the western United States. Bulletin of the American Meteorological Society, 87(12), 1699.

Yuan, X., Wood, E. F., Luo, L., & Pan, M. (2013). CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. Journal of Climate, 26, 4828-4847.

Yuan, X., Wood, E. F., & Ma, Z. (2015). A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. Wiley Interdisciplinary Reviews: Water, 2(5), 523-536.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

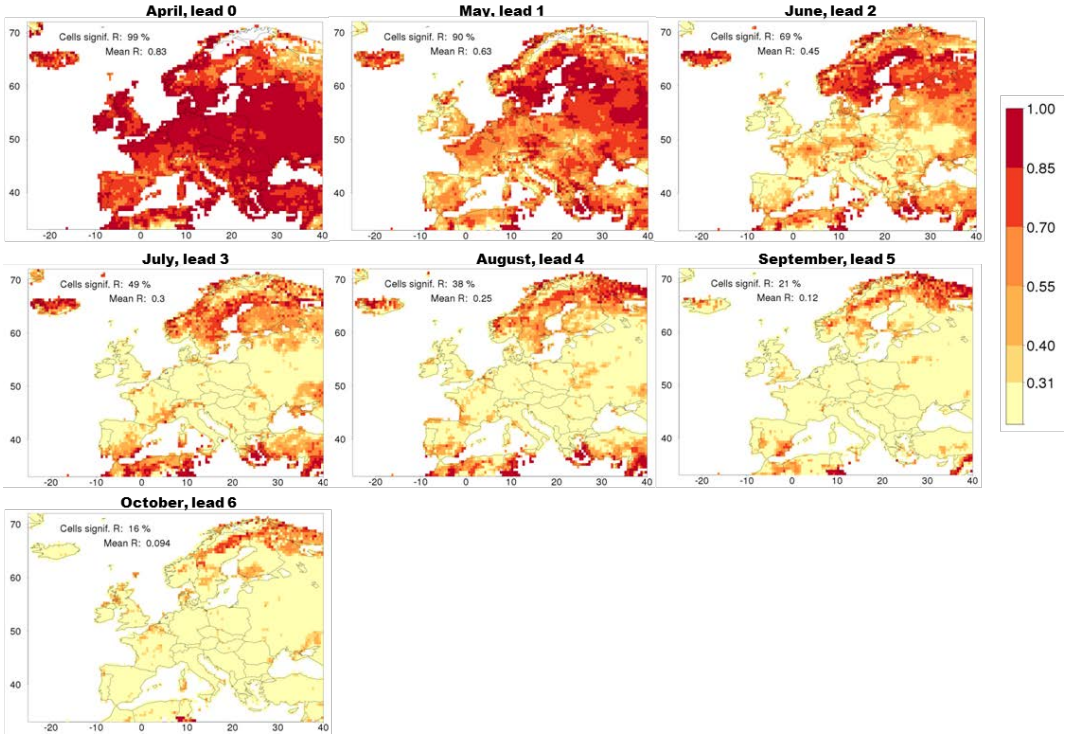

Figure 1: Skill of the runoff hindcasts initialised on April 1 for all seven lead months. Skill is measured in terms of the correlation coefficient between the median of the hindcasts and the observations (R). White, terrestrial cells correspond to cells where observations or hindcasts consist for more than one third of zeros or one sixth of ties. The threshold of significant skill lies at 0.31, so yellow cells have insignificant skill. The legend provides the fraction of cells with significant values of R and the domain-averaged value of R.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

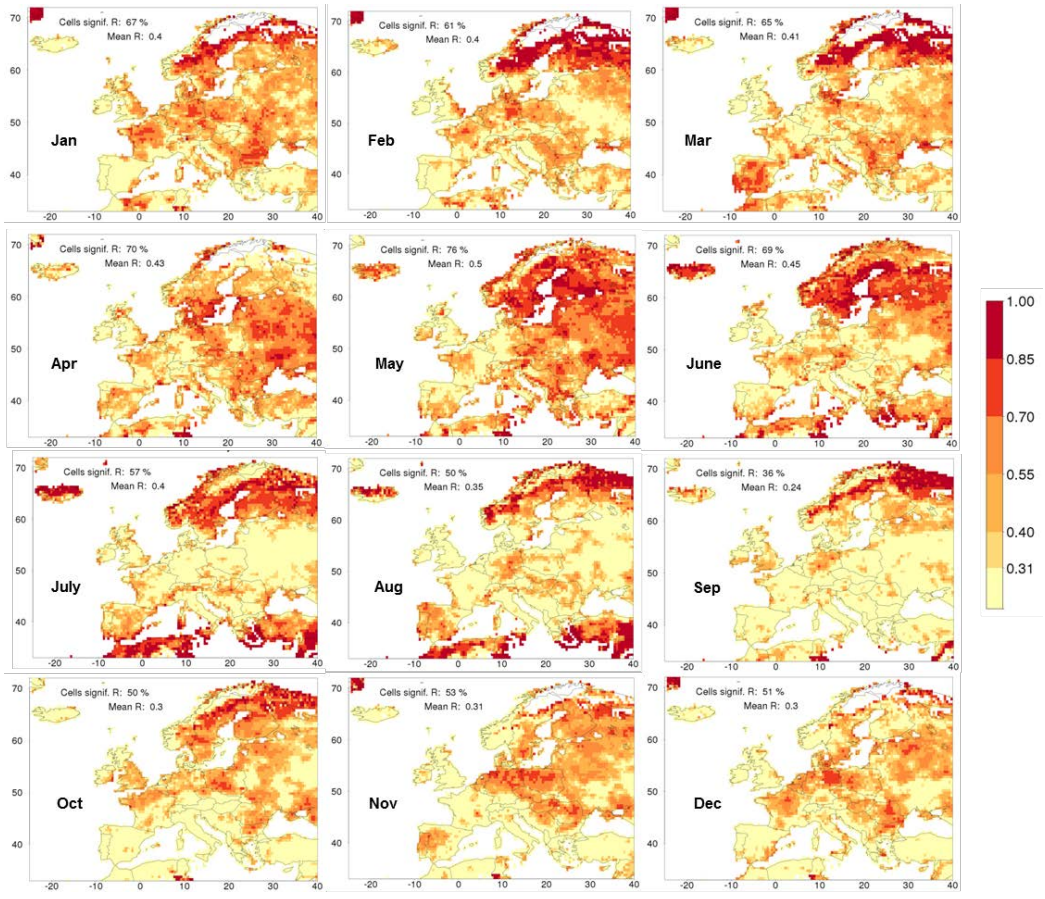

Figure 2: Annual cycle of skill (R) of runoff hindcasts of lead month 2. More explanation is given in the caption of Fig. 1.

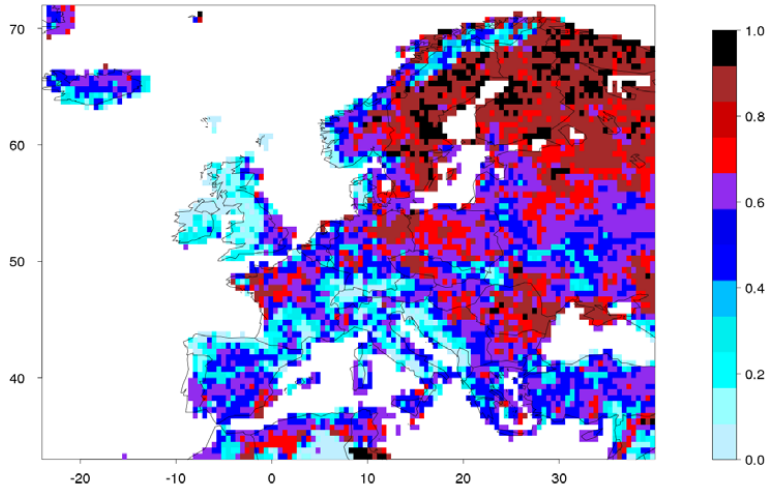

Figure 3: Fraction of the 12 months of the year with significant skill (R) in the runoff forecasts of lead month 2

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions

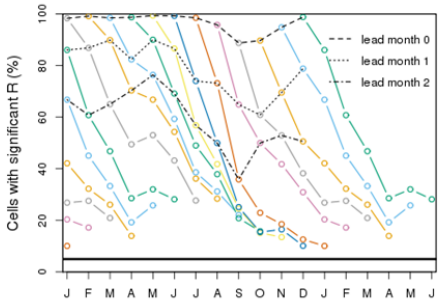

Figure 4: Fraction of cells with significant skill (in terms of R) in the runoff hindcasts, as a function of initialisation month and lead time. Each coloured curve corresponds to the hindcasts initialised in a single month. For better visualisation, parts of the curves that end in the next year are shown twice, namely at the left hand and the right hand side of the graph. Black lines (dashed, dotted and dashed-dotted) connect the results for identical lead times. The horizontal line gives the expected fraction of cells with significant skill, in the case that the hindcasts have no skill at all (5%).

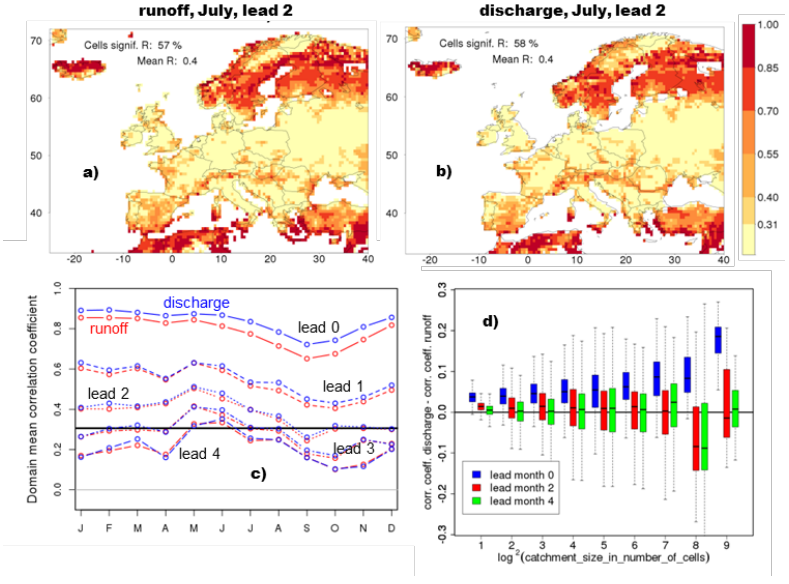

Figure 5: Comparison of the skill of the hindcasts of discharge and runoff. The two maps display R for runoff (a) and discharge (b) for hindcasts initialised on May 1 and lead month 2 (July), see further explanation in Fig. 1. Panel c depicts the annual cycle of the domain-averaged R for runoff (red) and discharge (blue) for lead months 0 to 4. The horizontal line at 0.31 is the threshold of significance for a single cell. Panel d is a box plot of the difference between R for discharge and runoff as a function of the catchment size. Each bin *i* contains the results for all catchments with a maximum of 2^i cells and more than $2^{(i-1)}$ cells, e.g. bin 4 is for all catchments with a size from 10 to 16 cells. Boxes represent the interquartile range and whiskers extend by 1.5 times the interquartile range from the box top and bottom. All values are averages over the twelve months of the year and results are shown for three different lead times.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions

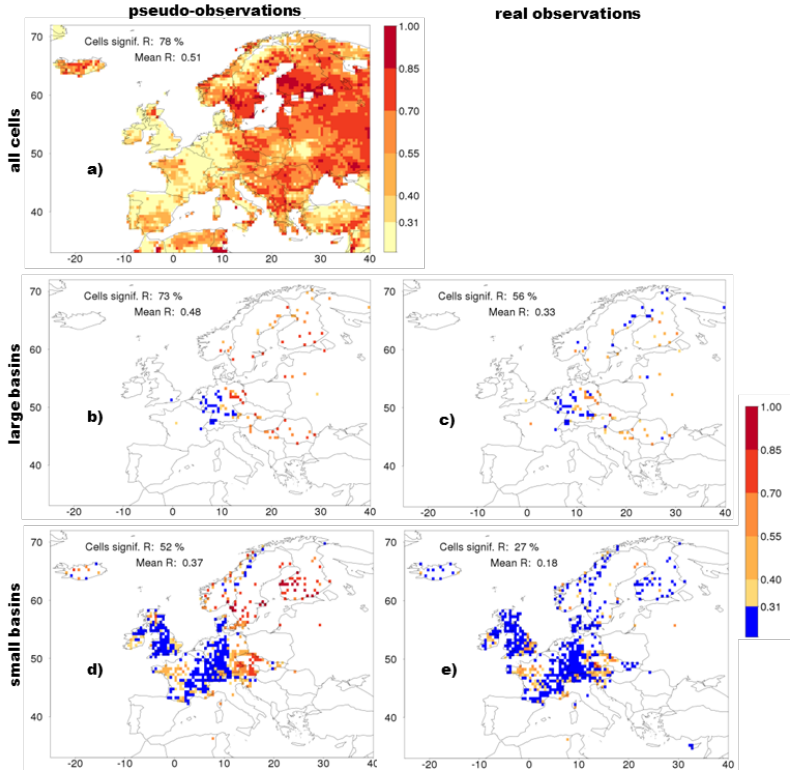

Figure 6: Skill (R) of the discharge hindcasts for May as lead month 2 (initialisation on March 1). In sequence: a) discharge verified with pseudo-observations, b) as a but for cells representing large basins only, c) discharge verified with real observations for large basins. The two final panels (d and e) are identical to b and c, respectively, but for cells representing small basins. More explanation is given in the caption of Fig. 1 but in panels d-e cells with insignificant skill are coloured blue instead of yellow for better contrast.

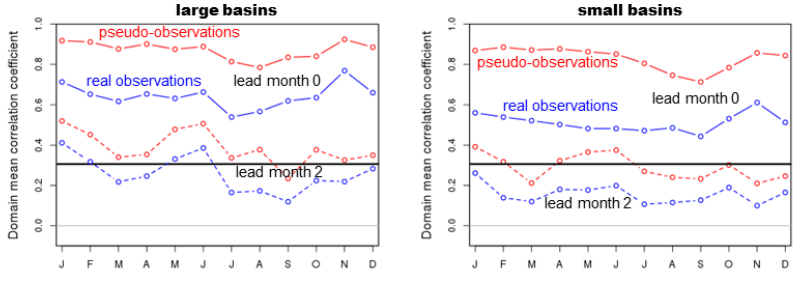

Figure 7: Comparison between verification of discharge with pseudo- (red) and real (blue) observations in terms of the annual cycle of the domain mean R. The horizontal line at 0.31 is the threshold of significance for a single cell. Results are shown for cells representing large basins (left) and cells representing small basins (right). Both panels depict cycles for lead months 0 and 2 only.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

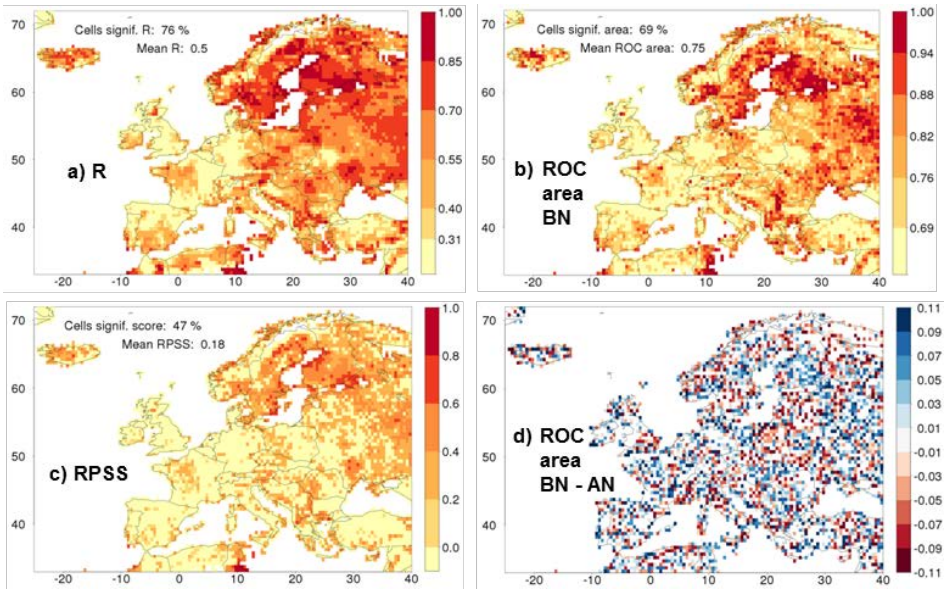

Figure 8: Maps of different skill metrics for one combination of a target month (May) and a lead month (2) of the runoff hindcasts. Panels show a) R, b) the ROC area for the Below-Normal tercile, c) the Ranked Probability Skill Score (RPSS) and d) the difference in ROC area between the Below-Normal and the Above-Normal tercile. In panels a, b and c skill is not significant in cells with a yellow colour. Legends provide the fraction of cells with significant values of the metric and the domain-averaged value of the metric.

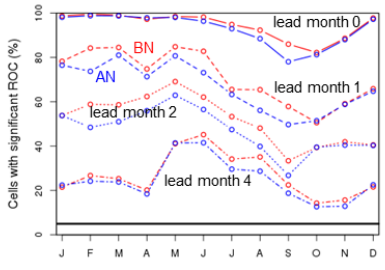

Figure 9: Skill of the runoff hindcasts in the Below Normal (BN) minus skill of the runoff hindcasts in the Above Normal (AN) tercile. The plot depicts annual cycles of the fraction of cells with a significant ROC area for the two terciles and for four lead months.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

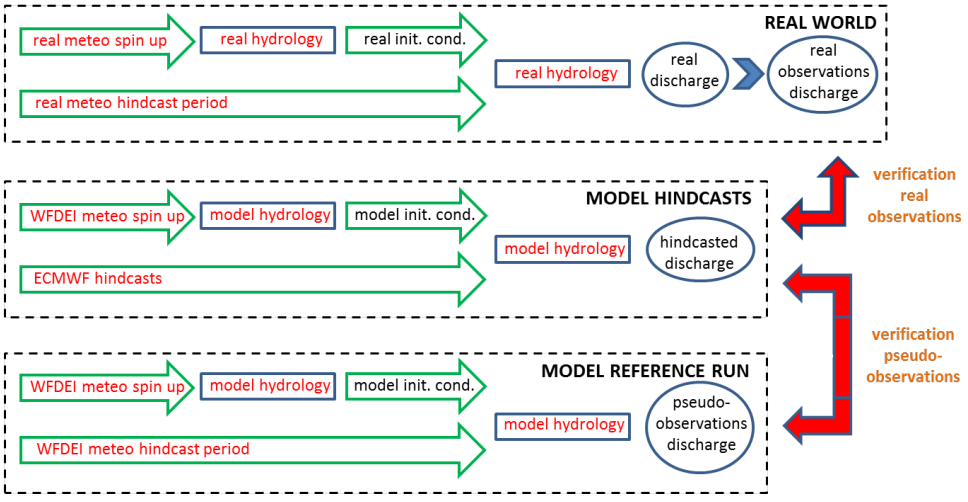

Figure 10: Diagram illustrating conceptual differences between verification of hindcasts (in the middle) with pseudo observations (bottom) and with observations of real discharge (top). See the text for a detailed explanation.