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

Wouter Greuell[1)], Wietse H. P. Franssen[1)], Hester Biemans[2)] and Ronald W. A. Hutjes[1,2)]

1) Water Systems and Global Change, Wageningen University, Droevendaalsesteeg 3, NL 6708 PB Wageningen, Netherlands
2) Water and Food, Wageningen Environmental Research, Droevendaalsesteeg 3, NL 6708 PB Wageningen, Netherlands

correspondence to 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 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 in hindcasting runoff and discharge is analysed with monthly temporal resolution, up to 7 months of lead time, 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, northern Germany, Romania and Bulgaria (mainly from November to January) and western France (from December to May). Generally skill decreases with increasing lead time, except in spring in regions with snow-rich winters. In some areas some skill persists even at the longest lead times (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. ctual skill is mostly and often substantially less than theoretical skill. This effect is stronger for small than for large basins, which is consistent with a conceptual analysis of the structural differences between the two types of verification. Qualitatively, the use of different skill metrics (correlation coefficient, ROC area and Ranked Probability Skill Score) lead to broadly similar spatio-temporal patterns of skill, but the level of skill decreases, and the area of skill shrinks, in the order correlation coefficient, ROC area below normal tercile, ROC area above normal tercile, Ranked Probability Skill Score and finally, ROC near 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 in low flow conditions (Li, et al., 2008) and irrigation management (Mushtaq et al. 2012; Ghile and Schulze 2008) to decrease crop yield losses. In order to be of any value in decision making processes of such sectors, forecasts must be credible, i.e. be skilful in predicting anomalous system states, as well as being relevant and legitimate to the decision making process (e.g. Bruno Soares and Dessai, 2016). In this paper we will introduce WUSHP (Wageningen University Seamless Hydrological Prediction system), a dynamical, model-based system (see Yuan et al., 2015) that was built around the Variable Infiltration Capacity (VIC) hydrological model and ECMWF's Seasonal Forecast System 4, to produce seasonal hydrological forecasts. It will be applied to Europe. The usefulness of the system depends partially on the level of its skill and the paper will therefore focus on an extensive assessment of the skill of WUSHP. The usual method of assessing skill of predictive systems is by analysing hindcasts, a strategy that will be adopted here as well.

During recent years, a number of systems for seasonal hydrological forecasts have been developed. Examples are the forecasting model suite he 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 (Bierkens and van Beek, 2009; Thober et al., 2015), but a few more concentrate on smaller domains such as the British Isles (Svensson et al., 2015), Iberia (Trigo, 2004) or France (Céron et al., 2010; Singla et al., 2012).

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 continental Europe, excluding the British Isles and Fennoscandia. Evaluating a number of forecasting techniques that produced distinct variations in the magnitude of skill, they found that spatial patterns in skill were remarkably similar among the different techniques, as well as comparable to the spatial patterns of the autocorrelation (persistence) of reference soil moisture. High skill was found in eastern Germany and Poland, Romania, southern Balkans and eastern Ukraine as well as north-western France. Less skill was found in the mountainous areas of Alps and Pyrenees, the northern Adriatic and Atlantic Iberia. Most skill was found for winter months (DJF), least for autumn (SON), this minimum shifting to summer (JJA) at long lead times (6 months).

Bierkens and van Beek (2009) developed an analogue events method to select annual ERA40 meteorological forcing on the basis of annual SST anomalies in the northern Atlantic and then made hydrological forecasts with a global-scale hydrological model applied to Europe. Evaluating only winter and summer half year aggregated skill, they found wintertime skill in large parts of Europe with maxima in eastern Spain and a zone from the southern Balkans and Romania through eastern Poland and western Russia to the Baltic states and Finland. Summertime skill was lower, generally by about 50% and even more around the Alps and the Adriatic. NAO based climate forecast added significant skill only in limited areas, such as Scandinavia, the Iberian Peninsula, the Balkans, and around the Black Sea.

Svensson et al. (2015) found skilful winter river flow forecasts across the whole of the UK due to a combination of skilful winter rainfall forecasts for the north and west, and strong persistence of initial hydrological conditions in the south and east. Strong statistical correlations between NAO and winter precipitation in Iberia lead to skilful forecasts of JFM river flow and hydropower production (Trigo et al., 2004). Ceron et al. (2010) and Singla et al. (2012) set up a high resolution river flow forecasting system (8 km) over France, for which seasonal climate forecast improved MAM skill over northern France, but worsened it over southern France (compared to a river flow model with proper initialisation of soil moisture, snow etc., but random atmospheric forcing). Demirel et al. (2015) found that both two physical models and one neural network over-predict runoff during low-flow periods using ensemble seasonal meteorological forcing for the Moselle basin. As a result forecasts of more extreme low flows are less reliable than forecasts of more moderate ones.

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 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” (misleadingly named “true discharge” in Bierkens and Van Beek, 2009; more appropriately “synthetic truth” in Shukla and Lettenmaier, 2011; “reanalysis” in Singla et al., 2012; “a posteriori estimates” in Shukla et al., 2014). We prefer the term “pseudo-observations” over “re-analysis” since the latter has a meteorological connotation that often implies the use of some form of (variational) data assimilation. We did not attempt any form of assimilating observed hydrological variables, such as discharge, in our reference run.

Pseudo-observations have the important 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. Finally, assessment of skill based on pseudo observations excludes model errors from the analysis, which is especially useful when addressing various sources of skill (Wood et al., 2016), something we will do in the companion paper.

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). Such a comparison was previously made by Bierkens and Van Beek (2009) and Van Dijk et al. (2013). We will analyse and discuss conceptual differences between using pseudo- and real observations for verification. We hypothesise 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.

This paper aims to analyse to what extent WUSHP is able to predict runoff and discharge in Europe over the full annual cycle and for lead times up to 7 months. We aim to assess skill at maximum resolution, i.e. at monthly resolution instead of seasonal or semi-annual aggregates. Where many studies use correlation coefficient as main skill metric we will assess skill also for the more probabilistic metrics ROC area and RPSS (see Sect. 2.3). The second aim of the paper is to get a better understanding of the effects of using pseudo-observations, as opposed to using actual observations, for the verification of hindcasts. In the next section we describe the concept and details of our modelling (Sect. 2.1) and analysis approach (Sect. 2.2 and 2.3). 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 in Sect. 3.2. Differences between theoretical and actual skill of discharge will be presented (Sect. 3.3) followed by an analysis of differences in skill determined with various metrics in Sect. 3.4. The discussion starts with a conceptual analysis of reasons for differences in actual and theoretical skill (Sect. 4.1), followed by a discussion of uncertainties (Sect. 4.2) and implications (Sect. 4.3).

In a companion paper (Greuell et al., 2016) we analyse the reasons for the presence or 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 experiments similar to the conventional Ensemble Streamflow Prediction (ESP) and reverse-ESP, which isolate different causes of predictability, are discussed. In the present results and discussion sections we will occasionally look forward to the identified causes of skill.

2 System, models, data and methods of analysis

To assess the forecast quality of our system, two approaches for verification of the hindcasts are used in this paper. First, we determine the skill of the hindcasts by comparing predicted discharge with the output of a reference simulation (the “pseudo-observations” leading to “theoretical skill”), allowing an evaluation that is continuous in space and time. Secondly, we quantify skill with respect to observations of real discharge (“real observations” leading to “actual skill”), allowing evaluation at a limited number of locations (discharge stations) on the river network only. Fig. 1 presents a flow diagram of the three relevant systems, namely the real world and the two model systems that generate the hindcasts and the pseudo-observations respectively.

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 three components that cause differences between actual and theoretical skill. First, the initial conditions 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, observations of discharge generally differ from real discharge (Juston et al., 2014) due to unavoidable measurement errors as illustrated in the upper right corner of the figure. Third, obviously a difference exists between real hydrology and model hydrology, central in each box. Since the hindcasted discharge and pseudo observations share the same model hydrology and the same initial conditions, and both are free from any observational errors, theoretical skill will always be larger than actual skill.

For now we simply accept, and even stress this a-priori ‘superiority’ of theoretical over actual skill. In Sect. 4.1 we will come back to this and further discuss, at least in qualitative terms, how each of the differences between the three systems affect skill assessment.

In the following subsections we will describe each component.

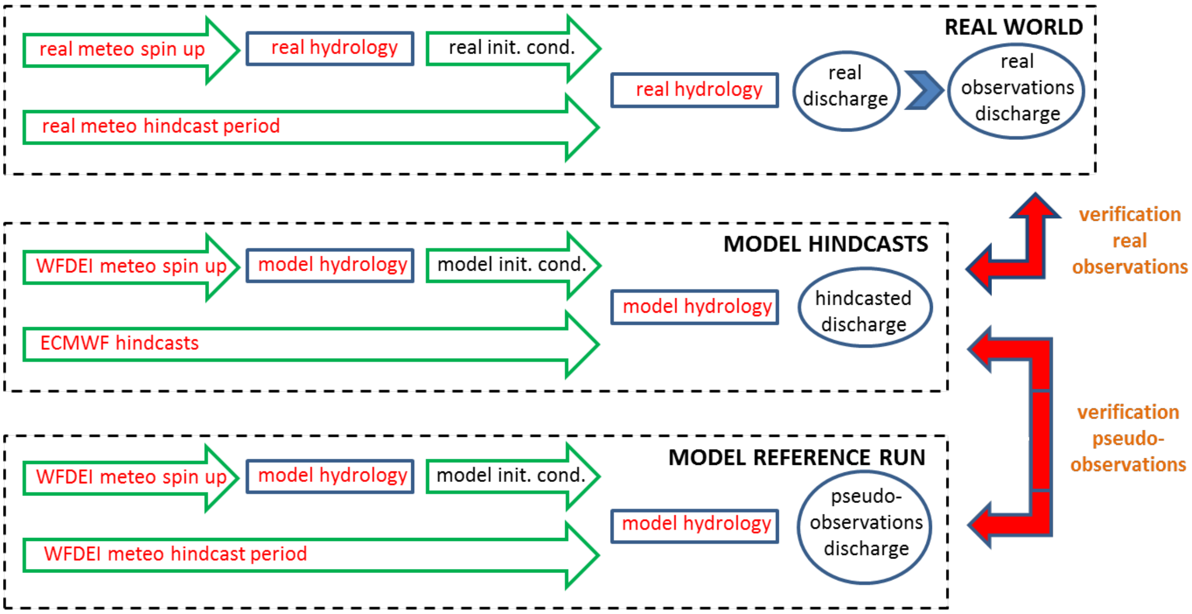

Figure 1: Conceptual setup of the present study, showing differences between verification of hindcasts (in the middle) with pseudo observations (bottom) and with observations of real discharge (top). See the text in this section for further explanation and Sect. 4.1 for further discussion.

2.1 The hindcasts and the reference simulation

WUSHP consists of two simulation branches: 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. 2. VIC is forced by a gridded data set of daily meteorological data. VIC is run in so-called ‘energy balance mode’ which requires resolving the diurnal cycle. Therefore, internally the model temporally disaggregates the daily input to 3-hourly data and runs at 3 hourly time step. Output of all variables is again at daily resolution. Because snow may contribute significantly to the seasonal predictability of other hydrological variables, VIC was run with the option of subgrid elevation bands. This means that for each gridcell calculations were carried out at up to 16 different elevations, with the aim of simulating the elevation gradient of snow. VIC was run in naturalised flow mode, i.e. river regulation, irrigation and other anthropogenic influences were not considered.

In the reference simulation VIC is forced by the WATCH Forcing Data Era-Interim (WFDEI; Weedon et al., 2014) for the period of 1979-2010, of which the first two years were used to spin up the states of snow, soil moisture and discharge, and not used in further analysis. The reference simulation has the dual aim 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 the initialisation of the hindcasts.

The second branch, the hindcasts, consists of three steps. Seasonal predictions of daily meteorological variables are taken from ECMWF's Seasonal Forecast System 4 (S4 hereafter). These are then bias-corrected using WFDEI 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 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 ECMWF Seasonal Forecast User Guide, online). The S4 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 whole system is thus probabilistic.

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 all 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. Since bias correction generally improves forecasting skill, 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). Such empirical distribution mapping of daily values has been successful in improving especially forecast reliability (rather than sharpness and accuracy; Crochemore et al., 2016).

VIC was run for the period of the S4 hindcasts (1981 – 2010). Additionally, for 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 one month simulation with WFDEI forcing. Since the hindcasts cover 30 years with 12 initialisation dates each and consist of 15 members, a total of 5400 hindcast simulations was carried out.

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) and Roudier et al. (2016). VIC exhibits a fairly small average bias (across 46 stations) of +23 mm/yr (= 7%) and overall differentiates well between low and high runoff basins with a spatial correlation coefficient of 0.955. However, specific discharge was overestimated in the Mediterranean and underestimated in northern Fennoscandia.

Annual cycles are fairly well reproduced across Europe, though VIC somewhat overestimates their amplitude. In northern Fennoscandia the spring peak is too late and too long. Annual cycles are best reproduced for rain-fed rivers in central Europe while those for rivers with significant snow dynamics are good (Alps). However, the annual cycle is more poorly reproduced in basins with strong soil freezing dynamics (northern Fennoscandia) or strong damping of discharge amplitudes by large lakes (southern Finland).

Perhaps more relevant in the present context is the model‘s capability to reproduce inter-annual variations in discharge. The standard deviation of simulated annual discharge was 9% higher than observed and the correlation between the two 0.935. Like most models, VIC is better in simulating high flows (95 percentile: Q95) than low flows (Q5); the first is slightly overestimated, the second more seriously underestimated. The inter-annual variation in Q5 is overestimated in central Europe and the Alps, but underestimated in Fennoscandia (overall correlation across Europe 0.40). The inter-annual variation in Q95 shows no clear spatial pattern and the overall correlation is 0.7.

All validation results discussed in these two paragraphs are for the VIC model forced by E-OBS (v9, Haylock et al. 2008). Our forcing, WFDEI, shows higher precipitation (+104 mm/yr) across most of Europe, except the Alps, Scotland and westernmost Norway. According to Greuell et al. (2015) this leads to higher mean discharge, higher inter annual variability and higher Q95 (not Q5) of simulated discharge for almost all stations.

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): the GRDC data set proper and the European Water Archive (EWA) data set. We mapped these two station data sets onto the VIC grid with a resolution of 0.5˚ x 0.5˚ and a time step of a month. To enable the investigation of the effect of size on some of our results, we made two sub-classes of observations. The first comprised observations for basins larger than 9900 km^2 (“large basins”), the second basins smaller than the area of the grid cells, i.e. smaller than about 2530 km^2 in southern Europe (at 35^o N) or < 1050 km^2 at 70^o N (“small basins”).

Initially, 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 basin size. The size of the model basins (“model basin area”) was determined by the DDM30 network. The size of the basins upstream of the observation stations (“station basin area”) was taken from the meta data of the observations. First the station basin area was compared to the model basin area of the cell that is nearest to the station (“nearest model cell basin area”).

Then, the mapping procedure for each observation varied slightly between the two classes of basins.

For large basins we then proceeded as follows:

- If the station and the nearest model cell basin area differed by less than 15%, the observations were matched with the model calculations for the nearest model cell.
- Otherwise, the station basin area was compared with the model basin 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 set within our European domain contained 111 cells with observations for large basins and 636 cells with observations for basins smaller than a model gridcell.

These data sets do not include any variable or parameter characterising the level of human impact. To enable analysis of the effect of anthropogenic flow modifications on predictive skill, we quantified the human impact by performing two model simulations with the Lund-Potsdam-Jena managed Land (LPJmL) model (Rost et al., 2008; Schaphoff et al., 2013). This model was operated at the same spatial resolution (0.5˚ x 0.5˚) and with the same river network (DDM30) as VIC, but the former does include dams (GRanD database; Lehner et al;. 2011) and associated reservoir management. From the discharge output of a naturalized LPJmL run and an LPJmL 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). For the analysis in Sect. 3.3, we selected all discharge observations for large basins with an AAPFD < 0.3, i.e. basins with a relatively small degree of human impact (about half of all 111 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 basin upstream of a particular cell.

Instead of analysing skill per target season and/or for a number of consecutive lead months, we analysed skill for every combination of 12 target and 7 lead months. 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.

Three skill metrics (see Mason and Stephensen, 2008, for a good discussion of the why and how of these) were computed: i) the correlation coefficient between the observations and the *median* values of the hindcasts (shortly “correlation coefficient” or R), ii) the area beneath the Relative Operating Characteristics (ROC) curve (shortly “ROC area”) and iii) the Ranked Probability Skill Score (RPSS). The ROC area is computed for each month separately and 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 (upper, lower and middle tercile, resp.; above, below and near-normal, AN, BN and NN categories). The same subdivision of observations and hindcasts in terciles was made to compute the RPSS. Since none of these metrics is sensitive to systematic biases in the forecasting system, no attempt was made to correct simulated runoff or discharge for any such errors prior to computing the skill metrics. So we focus our evaluation on the models capability to predict river flow anomalies rather than absolute rover flows.

All three skill metrics quantify, though in different ways, how well the ranking of the hindcasts matches the ranking of the observations. The correlation coefficient is a measure of the association between (pseudo-) observation and forecast ensemble median; we used the Pearson correlation coefficient. The ROC area is a measure of resolution or discrimination and 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 is a measure of accuracy and summarizes in a single number the skill of a forecast system to make forecasts with the correct percentage of ensemble members falling in any of the defined terciles. Perfect forecasts have values of 1 for all three skill metrics. Climatological forecasts (probabilistic forecasts that in our case each year predict a 0.33 chance of a high or low anomaly occurring) 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 RPSS, these 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. This implies association is significant for R > 0.31, resolution is significant for ROC area > 0.69 and accuracy is significant for RPSS > 0.

To a large extent, we found that our results and conclusions, in terms of spatio temporal patterns of skill, 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 mostly in terms of the correlation coefficient, which is in line with Doblas-Reyes et al. (2013). The sensitivity to the chosen metric and significant differences between these metrics will be discussed in Sect. 4.2.

All metrics were computed using the low and high level R packages “SpecsVerification” (Siegert et al., 2014) and “easyVerification” (Bhend et al., 2016) , respectively. Metrics cannot be computed if observations or hindcasts within the entire 30 year period consist for more than one third of zeros or one sixth of ties (i.e. equal values). Such skill gaps (i.e. the white terrestrial cells in Fig. 2 and 3) only occur in the far North due to rivers that are frozen for at least a month in winter.

3 Results

3.1 Spatiotemporal variation of skill in runoff forecasts

Eighty-four maps of skill of the runoff hindcasts were drawn for all 12 initialisation months and all 7 lead months (all are presented in supplementary material S1). Two cross-cuts through that collection are shown in Fig. 2 (for a single initialisation month) and 2 (for a single lead month). The seven panels of Fig. 3 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; cells where no metric could be computed remain white. 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.

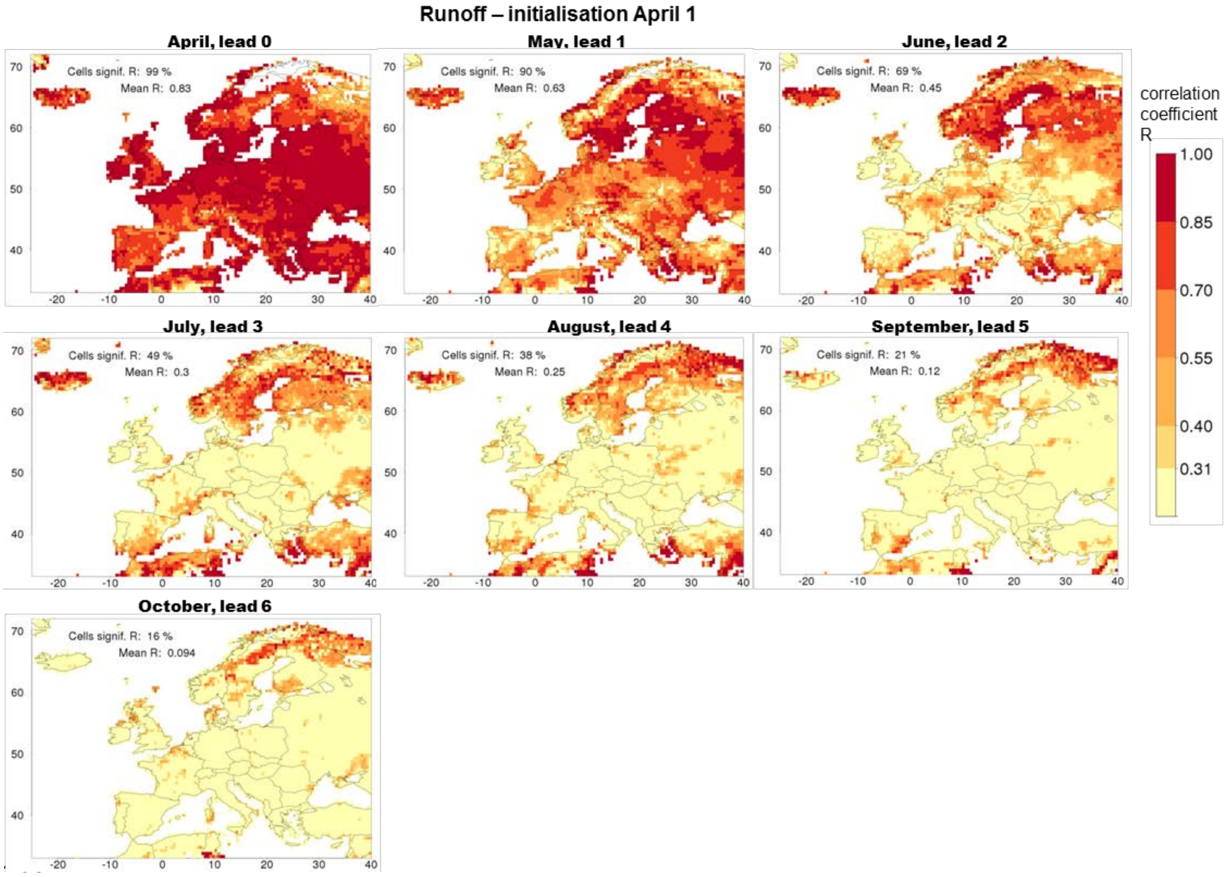

Figure 2: Skill of the runoff hindcasts initialised on April 1 for all seven lead months. Skill is measured in terms of the Pearson correlation coefficient between the median of the hindcasts and the observations (R). The threshold of significant skill lies at 0.31, so yellow cells have insignificant skill, (dark) red cells have (most) skill. White, terrestrial cells correspond to cells where observations or hindcasts consist for more than one third of zeros or one sixth of ties. The legend provides the fraction of cells with significant values of R (at the 5% level) and the domain-averaged value of R.

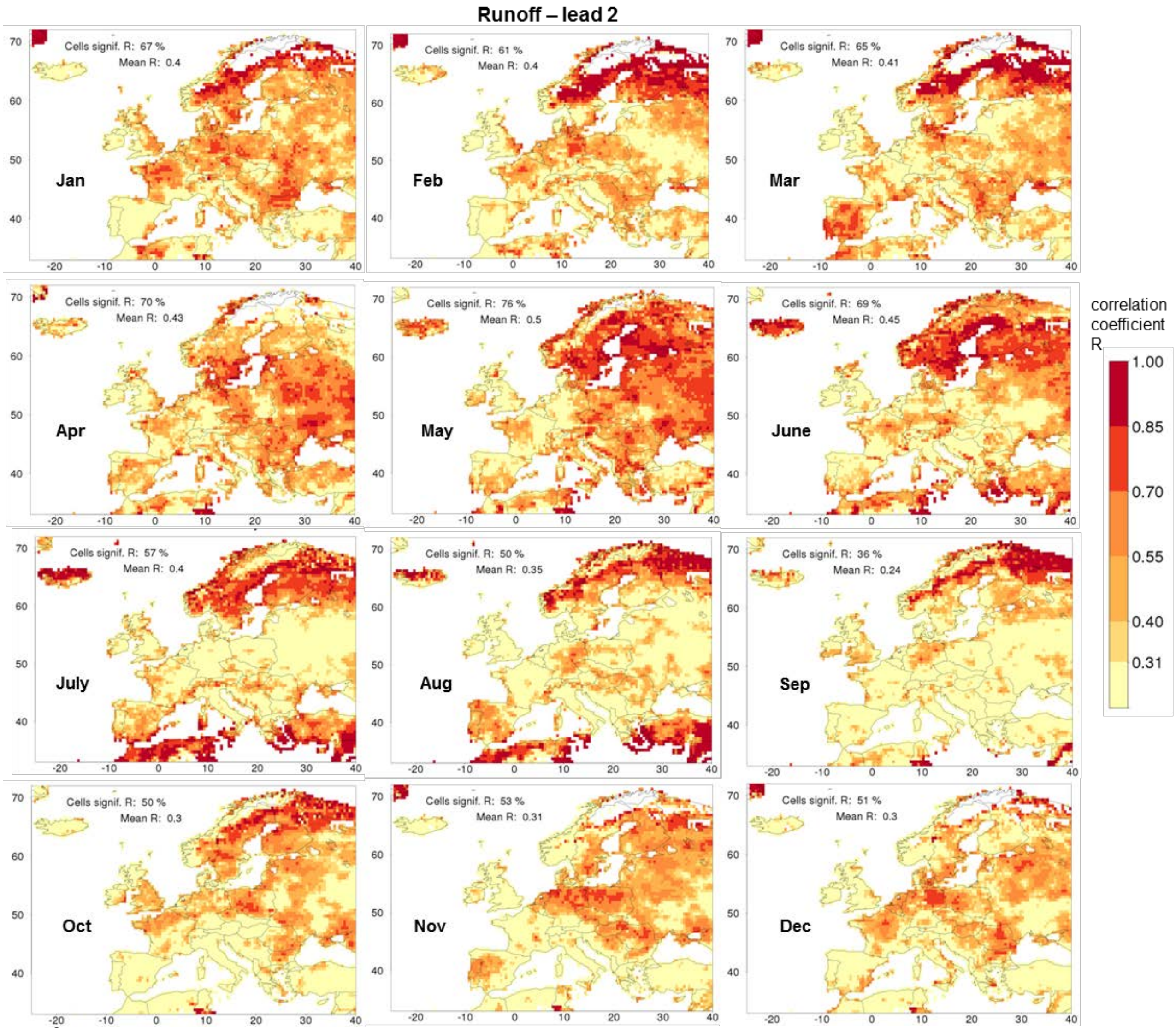

Figure 3: Annual cycle of skill (R) of runoff hindcasts of lead month 2. More explanation is given in the caption of Fig. 2.

The twelve panels of Fig. 3 show the annual cycle of skill of the hindcasts for lead month 2. Consistent skill (persistent during at least 2 consecutive target months) is found in (causes of skill are reproduced here from the companion paper, Greuell et al., 2016):

- 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, i.e. the end of the melting season, and –as shown in the companion paper- largely due to initialising snow. Compared to the rest of the peninsula, there is generally less skill along the Scandinavian Mountain range. The companion paper shows some evidence that this may be due to high variability of orographic rain, ill-represented in the S4 re-forecasts.
- Poland and northern 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. Here both initialisation of soil moisture and snow are important for skill .

- western France, more or less from Paris to Brittany and roughly from December to May. Skill derives from initialisation of soil moisture.
- The eastern side of the British Isles from January to April up to lead month 2. Also here skill derives from soil moisture initialisation.
- Romania and Bulgaria. The core as well as the whole period are the same as that for Poland and northern Germany. In addition to causes mentioned there, in this part of Europe also summer precipitation and evapotranspiration are forecasted fairly well.
- The southern part of the Mediterranean region from June to August. The high amounts of skill are limited to the coastal parts of northern Africa, Sicily, southern Greece, Turkey, Syria and Lebanon.
- The Iberian peninsula from January to March up to lead month 2, and July and August like the other parts of the Mediterranean mentioned before. Skill derives from soil moisture in initialisation and in winter also from some skill in precipitation.

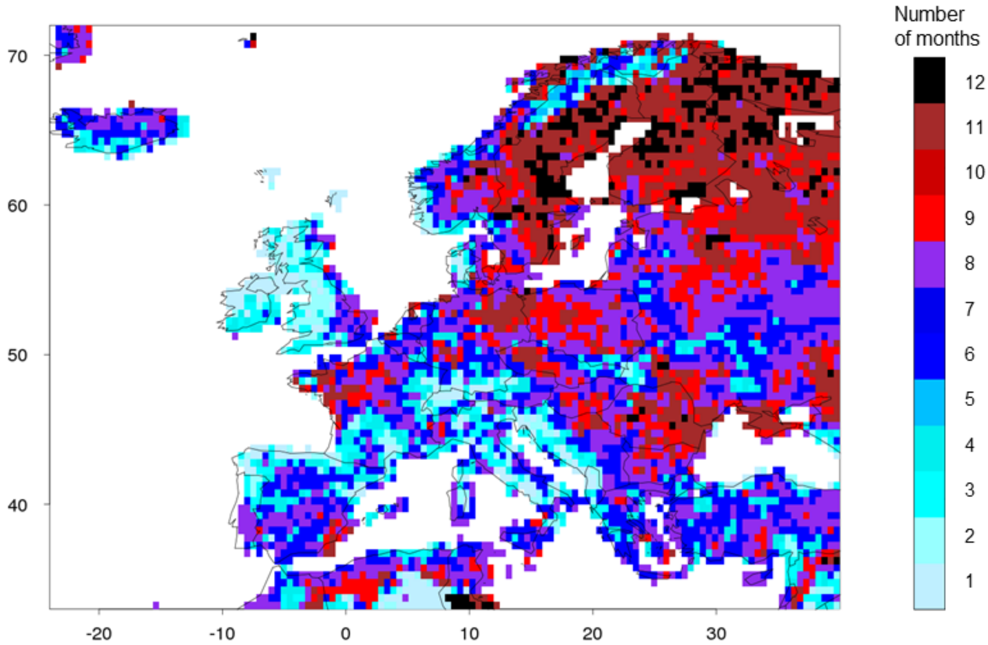

Figure 4: Number of months in a year with significant skill (R) in the runoff forecasts of lead month 2.

Figure 4 displays a synthesis of Fig. 3 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. northern coast of Spain), especially coastal regions on the western side of land masses (e.g. western coasts of

Denmark, southern Norway, Croatia and the British Isles), and mountain regions (e.g. the Alps, mountains in northern Norway and Sweden and on the Tatra on the border of Poland and Slovakia). The British Isles exhibit little skill, except for the eastern coast of Great Britain in late winter and early spring (JFMA). The companion paper shows that for regions with skill during a large part of the year, this skill is derived from complementary periods of skill due to initial conditions of snow and/or soil moisture.

These pan-European results can be compared to those of Bierkens and Van Beek (2009). They found maxima in predictability of winter discharge in Northern Sweden, Finland, the region between Moscow and the Baltic Sea, Romania and Bulgaria, and Eastern Spain. For the winter there is crude agreement with the current study about Northern Sweden, Romania and Bulgaria, but not about the other regions. For the summer, Bierkens and Van Beek (2009) compute maxima in skill for Southern Spain, Sardinia, Western Turkey and South-western Finland, a pattern that broadly agrees with the locations of the summertime maxima in skill (most of Fennoscandia and southern part of the Mediterranean region) we find.

Singla et al. (2012) found considerable skill in the Seine basin for low flows from June – September, a bit more eastern from the region where we found skill. Trigo et al. (2004) using a statistical model based on December NAO indices found skill for JFM discharge (and hydropower production) for the Douro, Tejo and Guadiana basins covering most of central and western Iberia. We confirm this skill which last till about May here, when initialised in January. In addition (not analysed by Trigo) we find skill beyond lead zero also in summer but then more concentrated around the south eastern coast of Iberia. Svensson et al. (2015) using a statistical model, based on NAO indices and river flow persistence, found good skill for winter river flows on the eastern side of the British Isles, consistent with our findings, and barely significant skill on its western coast that we do not reproduce.

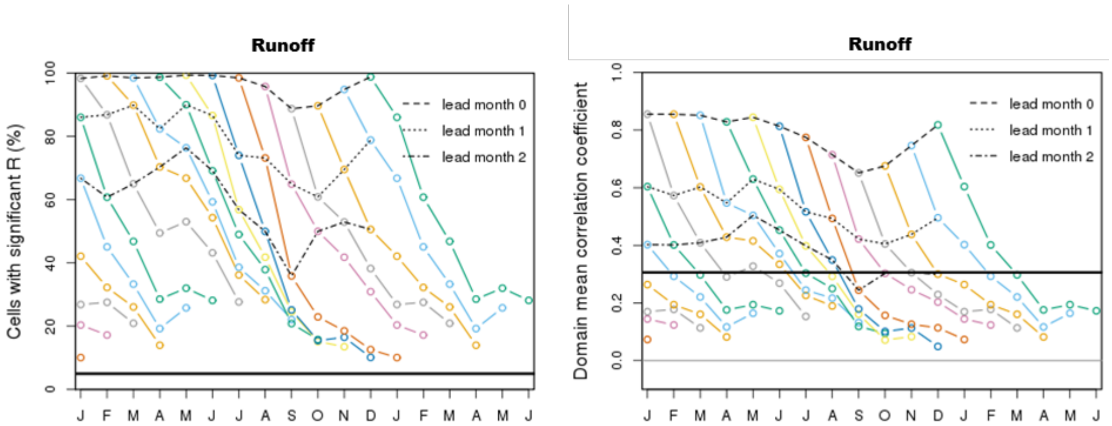

Figure 5: At left a) Fraction of cells with significant skill (in terms of R), and at right b) domain average correlation in the runoff hindcasts, as a function of initialisation month and lead time. Each coloured curve corresponds to the hindcasts initialised in a single month. For better visualisation, 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 in a) shows the expected fraction of cells with significant skill, in the case that the hindcasts have no skill at all (5%), in b) the minimal magnitude of the correlation of a single cell for it to be statistically significant

Figure 5a 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, implying there are (small) pockets of skill even at lead month seven. 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 phenomenon caused by initial conditions of snow that takes longer or shorter to melt in (late) spring. For details, see the companion paper. Fig. 5b shows decay and gain trends of the domain-averaged R. It shows that a forecast initialised in February exhibits higher domain average skill into June (5 lead months), than one starting in July has for August (2 lead months). Similar summary plots for the other skill metrics are presented in the Fig. S2, and discussed in Sect. 3.4.

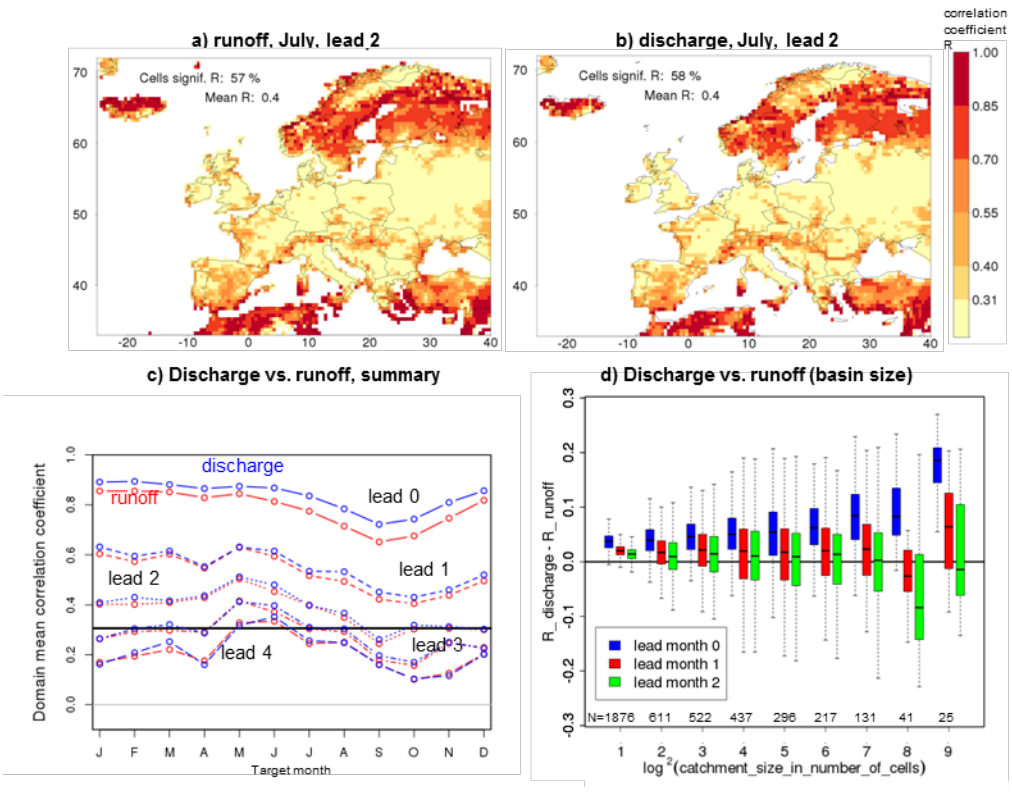

Figure 6: Comparison of the skill of the hindcasts of discharge and runoff. The two maps display R for runoff (a) and discharge (b) for hindcasts initialised on May 1 and target month July (see further explanation in Fig. 1). Panel c depicts the annual cycle of the domain-averaged R for runoff (red) 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 basin size. Each bin *i* contains the results for all basins with a maximum of 2^i cells and more than $2^{(i-1)}$ cells, e.g. bin 4 is for all basins with a size from 10 to 16 cells. Boxes represent the interquartile range and the median; whiskers extend to minimum and maximum values found in the bin. All values are average differences over the twelve months of the year and results are shown for three different lead times. The value above the abscissa give the number of cells in each bin.

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. 6, which depict the skill in runoff and 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 (not shown). There are, however, subtle differences because rivers aggregate the skill, or lack of skill, from the whole upstream part of their basin. As a result, cells containing rivers with large basins may contrast against adjacent cells if these contain rivers with a small, local basin. 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. 6b 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.

Domain summary statistics of skill also differ slightly between runoff and discharge. Figure 6c 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. 6c 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 with 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 basin (Fig. 6d). 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 basin. 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 basin. For longer lead times (from lead month 1 on), when the domain-averaged difference in skill has become very small, the figure shows no effect of the basin size. Referring to the comparison between runoff and discharge in panels Fig. 6a and 6b for lead month 2, cases like the Danube (more skill than local rivers) and the Loire (less skill than local rivers) tend to cancel when the entire domain is considered. .

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 7 compares the theoretical skill (Fig. 7b and 7d for large and small basins, respectively) with actual skill (Fig. 7c and 7e for large and small basins, respectively) for a single combination of a target month (May) with a lead month (2).

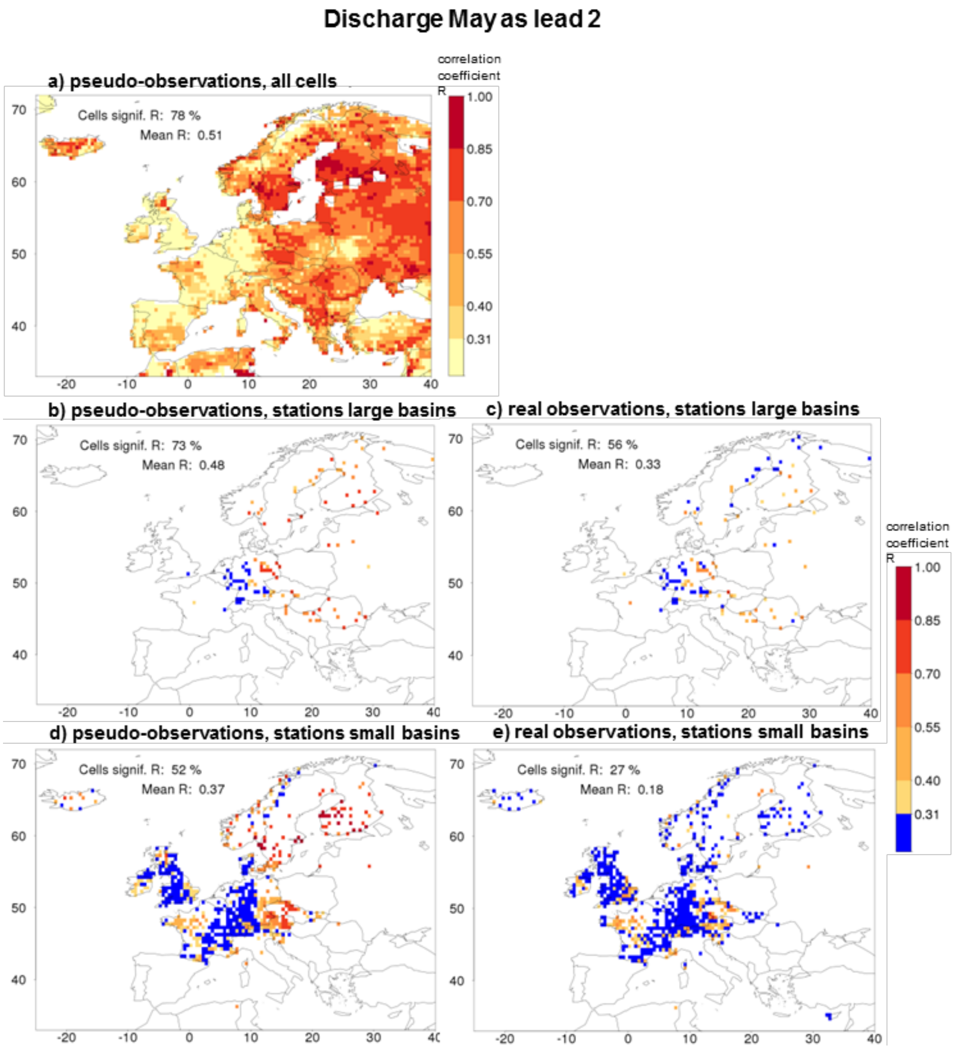

Figure 7: Skill (R) of the discharge hindcasts for May as lead month 2 (initialisation on March 1). In sequence: a) discharge verified with pseudo-observations, b) as a but for cells representing large basins only, c) discharge verified with real observations for large basins. Panels d) and e) are identical 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) and e) cells with insignificant skill are coloured blue instead of yellow for better contrast.

For this combination of May forecasts initialised in March, 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. The larger basins, especially those in northern Fennoscandia lose all skill when using actual observations, a region where VIC also performed poorly in reproducing historic flows. There the specific discharge was underestimated and the annual cycle was poorly reproduced; especially the spring peak occurred too late and too long (Greuell et al., 2015). In central

Europe useful skill remains when using real observations, a region where VIC well reproduced annual cycles, though interannual variation in low flows where overestimated in that area.

Figure 8 compares actual with theoretical skill for all target months and two lead times by considering the domain-mean R. Similar figures for the other skill metrics are presented in Fig. S4 and discussed in the next Sect. 3.4. 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 is comparable to Van Dijk et al. (2013), who found a ratio of actual to theoretical skill of 0.54 for 6192 basins worldwide in terms of the ranked correlation coefficient.

Comparing skill for small basins with skill for large basins in Fig. 8, 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, this difference becomes insignificantly small. So, the apparent difference in theoretical skill between large and small basins can be attributed almost entirely to the geographical distribution of the discharge monitoring stations, with stations on small basins being relatively more often located in regions with relatively little skill like Germany, France and the British Isles than large basin stations.

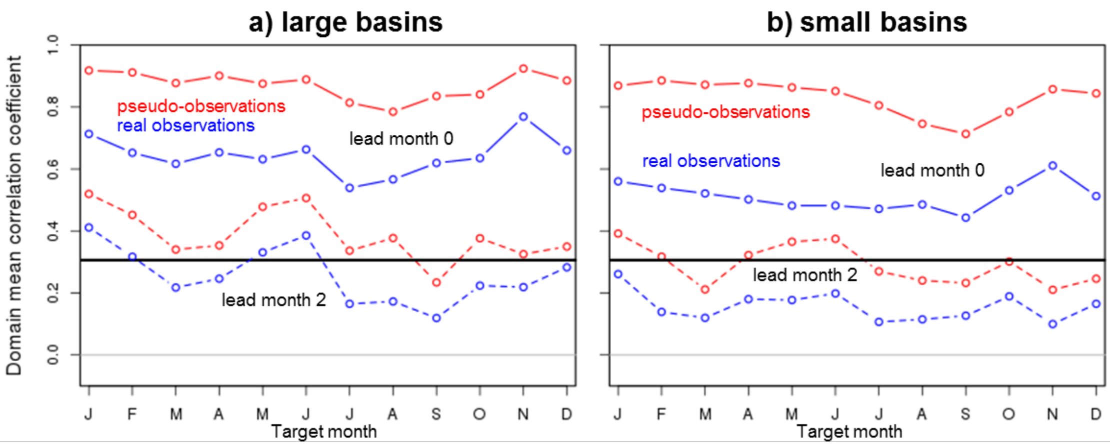

Figure 8: Comparison between verification of discharge with pseudo- (red) and real (blue) observations in terms of the annual cycle of the domain mean R. The horizontal line at 0.31 is the threshold of significance for a single cell. Results are shown for cells representing large basins (left) and cells representing small basins (right). Both panels depict cycles for lead months 0 and 2 only.

The second effect of the size of basins is that reduction between theoretical and actual skill is larger for small basins than for large basins. This is perhaps even more clear from Fig. S4 in the supplementary material. We speculate that this is due to 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, confirmed for VIC in Europe by Greuell et al. 2015). 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.

Finally, 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. For relatively natural basins (AAPFD < 0.3; see end of Sect. 2.2), 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 larger for basins with less impact. However, since the difference between the two ratios is small we conclude that the effect of evaluating naturalised runs against observations that are obviously 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 smaller than the spatial resolution of the simulations.

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, for runoff, with results in terms of other skill metrics. Figure 9 gives an example for one particular target month and lead month, i.e. target May initialised in March (lead 2). Fig. 9a, 9b and 9c show the skill patterns for R, for the ROC area for Below Normal (BN) years and for the RPSS. The three patterns are spatially similar to a large degree, though the magnitudes and number of significant cells do differ. The pattern of the map of the ROC area for Above Normal (AN) years (see Fig. S1) is also similar to the patterns of the three maps shown. On average, across all lead and target months, 89% of the cells that have significant R also have significant ROC scores for the BN tercile, 84% also for the ROC scores for the AN tercile. Finally, 65% of the cells that have significant R also have significant RPSS scores. The fraction of cells with no significant R, but with significant ROC or RPSS remains below the 5% level across all target and lead months, and thus such cases are likely due to chance.

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).

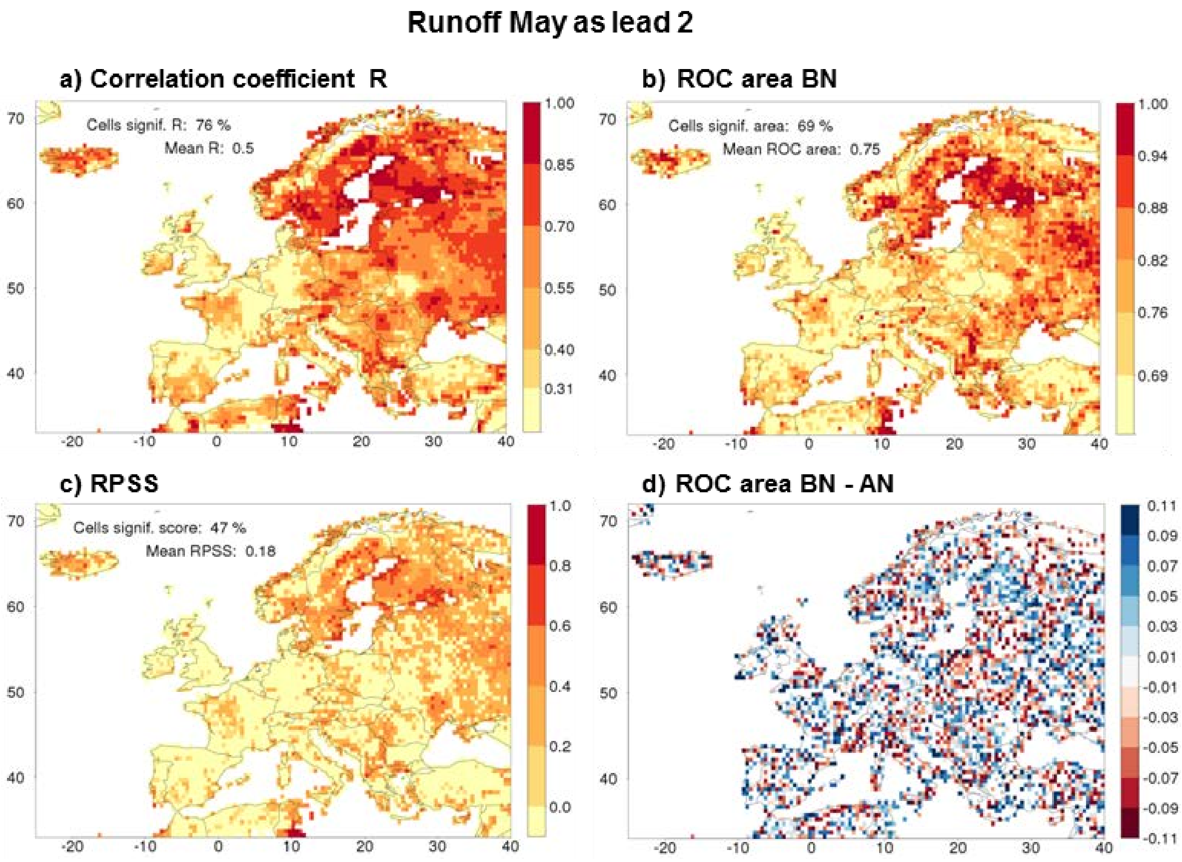

Figure 9: Maps of different skill metrics for one combination of a target month (May) and a lead month (2) of the runoff hindcasts. Panels show a) R, b) the ROC area for the below normal tercile, c) the Ranked Probability Skill Score (RPSS) and d) the difference in ROC area between the BN and AN terciles. 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.

Although 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 May target month and lead month two shown in Fig. 9, 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, see Fig. S2c and d; 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 represents the skill across all terciles. All metrics show a minimum value in the annual cycles in either September or in October, irrespective of lead time; maxima are attained in February for lead month 0 shifting to May at longer lead times (Fig. S2). Finally, Fig. 9d presents a map of the difference between the BN and the AN ROC area. BN ROC values are larger than AN (blue colours) in southern Finland and central Sweden, western France, Hungary and Serbia and large parts of Russia. The reverse (ROC AN > ROC BN, red colours) is true in eastern Poland and the Baltic states, southern eastern France (Rhone basin) and eastern UK.

For other combinations of target and lead months the results of this analysis are similar, though numbers may vary. See supplementary figures.

Figure 10 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 combinations of lead and target month the fraction significant cells is larger for the BN than for the AN tercile. This is perhaps not as expected from the VIC performance in reproducing historic flows, which is better for high flows than for low flows (Greuell et al., 2015; recall that their high/low flows are defined as Q95 and Q5, respectively, while here they are Q67 and Q33; see also Sect. 2.1). However, the AN and BN fractions tend to become equal (i) when these ROC areas approach 1.0, (ii) when they approach the limit of no skill (5%) and (iii) during target months from October to January.

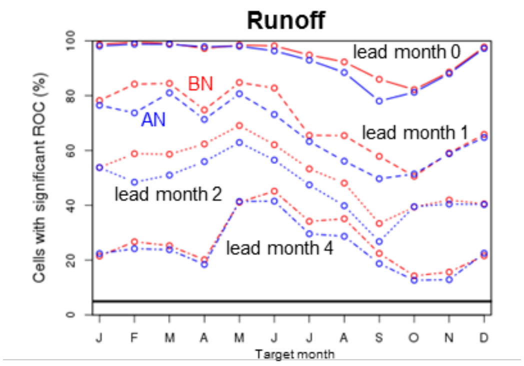

Figure 10: Skill of the runoff hindcasts in the Below Normal (BN) compared to the skill of the runoff hindcasts in the Above Normal (AN) tercile. The plot depicts annual cycles of the fraction of cells with a significant ROC area for the two terciles and for four lead months.

4 Discussion

4.1 Theoretical versus actual skill

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 Fig. 1) 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 (WFDEI), which differs from real meteorology. Therefore, this difference is expected to lead to more theoretical than actual skill. During the hindcast period, all three systems have different meteorological forcing. 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 skilful hindcasts and the WFDEI data set, on the other side. Unfortunately, we do not have a well-founded expectation about such a disagreement 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, in terms of physics and in terms of spatial and temporal discretisation, 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. One assumption implicitly made in the diagram is that the basin of the observation station and the model basin are identical. This is not the case (see Sect. 2.2), so differences between observation and model basin form an additional cause of disagreements between theoretical and actual skill. Again, this will favour theoretical skill with respect to actual skill since basins are identical in the hindcasts and the reference simulation. In particular, differences in meteorological forcing between the basin of the observation station and the model basin 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 at least half to the difference between theoretical and actual skill.

3. In the real world discharge observations are subject to measurement errors. Measurement errors of discharge are not constant over time (due to varying cross sectional areas, following erosion and sedimentation) and therefore add noise to the data; noise always reduces skill. 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.
4. 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, while model initial conditions would, of course, differ from real initial conditions, the two model system had identical initial conditions. Hence, this difference would again 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.

Our data analysis, Sect. 3.3, 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 can lead to a degradation of the theoretical skill if the hydrological models have unrealistic memory time scales in their storage compartments. If this memory, from stored water in either snow, soil or aquifer (or man-made reservoirs behind dams) , is too strong then skill will reduce with calibrating the model towards more realistic storage accumulation. However, if this memory is too small before improving the model, then, of course, the reverse may happen and skill increases with optimization.

An example proving this statement is a model that 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 more realistic amounts of snow, would exhibit less skill. Another example is predictive skill caused by interannual variations in the initial amount of soil moisture and/or groundwater. A model that is imperfect because it overestimates the transport speed of water 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.

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.

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.

4.2 Results and uncertainties

There seems to be a broad correspondence between the probabilistic forecast verification presented here and the model validation presented in Greuell et al. 2015; and Roudier et al. 2016. These studies found that average discharge and inter-annual variations therein are well reproduced against observations, consistent with our result that all skill scores -also against real observations (see Fig. S4 for the lead 0 results)- are good for large parts of Europe in the first lead month. Their finding that high flows are generally better reproduced than low flows seems to contradict with our fact that BN forecasts are more skilful than AN forecasts (although by a small margin, and for lead 0 mostly so in southern Europe, Fig. S1 e.g. Initialisation April). This discrepancy may be due to different definitions of high or low flows between these studies and the present one. They define high and low flows by Q95 and Q5 based on daily discharge, respectively, while here we use Q66 and Q33based on monthly discharge, much less extreme values. Also, their study showed that the variability in Q5 was more overestimated than the variability in Q95, which may be a reason for the higher skill we find in the lower tercile (skill requires variability, see discussion of companion paper), though this inference is hard to prove. This prior work also invokes some warnings. Greuell et al. (2105) found that seasonal flow cycles show a too late and too broad spring peak in (northern) Fennoscandia. This suggests that our theoretical forecast skills may also be too high at too long lead times in that region and season, (as was also already revealed by comparing Figure 7b vs 7c).

In a future extension of our work, an objective method like cluster analysis could reveal regions where skill has a 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 and its companion paper.

There also seems to be a broad correspondence between the regions and seasons with skill identified in the present work, with that from more spatially or temporally confined studies based on entirely different physical or even statistical models. Without repeating the more detailed description in the Introduction and closer comparison in Sect. 3.1, we restate here that the results of Bierkens and van Beek, (2009) and Thober et al. (2015) were similar at the European domain. These pan-European studies, like ours, confirm more regional studies such as for the British Isles (Svensson et al., 2015), Iberia (Trigo, 2004) or France (Céron et al., 2010; Singla et al., 2012). Though a high resolution study like the latter may add much spatial detail, this does not change the region and season of skill.

Our results are based on a forcing with the 15 member, monthly initialized, 7 month forecast version of ECMWF System 4, basically because at the start of this work that hindcast was the only one accessible to us but also because it allows verification at the highest temporal resolution. Alternatively, we could have used the 51 member seasonally initialised (4 times per year), 7 month forecast version of the same model. That would have provided us with better constrained, more precise statistics (larger sample size), or would have allowed assessment of more percentiles (e.g. quintiles instead of terciles) at similar precision. However, the variation of skill over a year would not have been resolved with such detail as in the present work. Finally, a 15 member, seasonally initialized, 12 month forecast version is available. Our results show that for some regions at lead month 6 still a few, small pockets of persistent skill remain, suggesting that extending the forecast for our domain might be worth exploring.

Other seasonal forecasting systems, based on different coupled ocean-climate models, exist that could have been used, such as CFSv2 (Saha et al., 2014), Glosea5 (MacLachlan et al., 2014). Given that, at least at large scales, multi model ensembles exhibit better climate forecast skill, it is interesting to investigate if that additional skill also propagates into river flow forecasts. While this seems to be true for the Eastern United States (Luo & Wood, 2008) it is not known if similar conclusions could be drawn for Europe. A similar reasoning can also be extended to the hydrological models: using a multi climate model ensemble to force a multi hydrological model ensemble might also provide improved skill, as the latter models may be complementary in the regions and seasons of best model performance. Bohn et al. (2010) showed some advantage of using an ensemble of three hydrological models (but with a single forcing), over using only the best of the three, but only after bias correcting the hydrological output and making a linear combination of them with monthly varying weights.

4.3 Implications and recommendations

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.

For applications of these seasonal forecasts in decision making processes at (sub) basin level, a more detailed skill analysis is recommended for that specific (sub)basin, preferably after a better model calibration for that same basin. The facts presented in this study that anomaly correlations and ROC scores for the AN and BN terciles are significant for large parts of the domain several lead months in advance, supported by (fairly) positive validation results for interannual variability of high and low flows (Greuell et al., 2015; Roudier et al. 2016), suggest these anomaly forecasts are good enough to be used as such. However, areas of significant RPSS are much smaller and remain significant for shorter lead times. Spatially distributed calibration of VIC model parameters, or distribution based calibration of modelled discharge to observed, or both, might also increase the RPSS. This might then allow forecasting of absolute discharge magnitudes and thus inform decision making processes that involve certain absolute discharge thresholds.

In the respective Result sections we already discussed the probable reasons for skill, which are much elaborated on in the companion paper. In general that paper shows that for most areas skill in runoff is caused by initialising snow and /or soil moisture properly, only in few areas and seasons skill in precipitation or skill in temperature and evapotranspiration adds to that beyond the first lead month. This has two implications: one is that, if ever the skill of seasonal climate forecasts improves for Europe, this may well translate to improved seasonal river flow forecast too. The second is that better initial conditions of snow water equivalent and soil moisture from observations may do the same, but the latter only if the spatial distribution of the soil moisture storage capacity is more realistic too (see Sect. 4.1).

Overall the present analysis shows that especially in winter, spring and early summer, there is potentially good skill to forecast runoff and discharge in large parts of Europe, with considerable lead time. While this broadly confirms previously published work, the present study (while being specific to our model setup) gives much more spatial and temporal (season and lead time) details. As such it provides a good basis to support operational forecasts and to add information about skill to seasonal forecasts, which is very important for proper value assessment and decision making.

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, northern Germany, Romania and Bulgaria (mainly from November to January) and western 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 eastern 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 basin.

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. For 10 day flow forecasts Alfieri et al. (2014) also found that, especially in mountain areas. performance drops significantly in river basins with upstream area smaller than 300 km^2.

Spatio-temporal 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.

6 Author Contributions

Greuell and Hutjes designed the experiments, with suggestions from the other co-authors. Franssen and Greuell developed the workflow scripts and performed all the simulations. Greuell and Franssen developed the analyses and plotting scripts in R. Biemans did the LPJmL work on AAPFD. All co-authors participated in repeated discussions on interpretations of results and suggested ways forward in the analysis. Greuell prepared the first version of the manuscript with contributions from all co-authors. Hutjes prepared the revision of the manuscript with contributions from all co-authors.

7 Conflicting Interests

The authors declare that they have no conflict of interest.

8 Acknowledgments

This study was financially supported by the EUPORIAS project (EUropean Provision of Regional Impact Assessment on Seasonal-to-decadal timescale); grant agreement No. 308291, funded by the European Commission (EU) project in the Seventh Framework Programme. We thank the valued suggestions and insightful comments from two (anonymous) reviewers that contributed to an improved version of the manuscript.

9 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.

Bhend, J., Ripoldi, J., Mignani, C., Mahlstein, I., Hiller, R., Spirig, C., Liniger, M., Weigel, A., Bedia Jimenez, J., De Felice, M., Siegert, S., (2016) easyVerification: Ensemble Forecast Verification for Large Data Sets. https://CRAN.R-project.org/package=easyVerification

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.

Bohn, T. J., Sonessa, M. Y., Lettenmaier, D. P. (2010). "Seasonal Hydrologic Forecasting: Do Multimodel Ensemble Averages Always Yield Improvements in Forecast Skill?" Journal of Hydrometeorology 11(6): 1358-1372.

Bruno Soares, M. and S. Dessai (2016). Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe. Climatic Change 137(1): 89-103.

Céron, J. P., Tanguy, G., Franchistéguy, L., Martin, E., Regimbeau, F., Vidal, J. -P. (2010). "Hydrological seasonal forecast over France: feasibility and prospects." Atmospheric Science Letters 11(2): 78-82.

Crochemore, L., Ramos, M. H. and Pappenberger, F.. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrol. Earth Syst. Sci. 20(9): 3601-3618.

Demirel, M.C., Booij, M.J. and Hoekstra, A.Y. (2015).The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. Hydrol. Earth Syst. Sci., 19, 275–291, 2015

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.

ECMWF Seasonal Forecast User Guide, retrieved from: http://www.ecmwf.int/en/forecasts/documentation-and-support/long-range/seasonal-forecast-documentation/user-guide/introduction

Ghile, Y. B., and Schulze, R. E., 2008: Development of a framework for an integrated time-varying agrohydrological forecast system for southern Africa: Initial results

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, doi: 10.5194/hessd-12-10289-2015

Greuell, W., W. H. P. Franssen, H. Biemans and R. W. A. Hutjes. Seasonal streamflow forecasts for Europe – II. Explanation of the skill. (2016, in revision) to Hydrol. Earth Syst. Sci., doi: 10.5194/hess-2016-604

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.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. Journal of Geophysical Research: Atmospheres (1984–2012), 113(D20).

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.

Lehner, B., Reidy Liermann, C., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J., Rödel, R., Sindorf, N., Wisser, D. (2011): High resolution mapping of the world's reservoirs and dams for sustainable river flow management. Frontiers in Ecology and the Environment 9(9): 494–502.

Li, H.,, Luo, L. and Wood, E.F. (2008). Seasonal hydrologic predictions of low-flow conditions over eastern USA during the 2007 drought. Atmospheric Science Letters **9**(2): 61-66.

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.

Luo, L. and E.F. Wood, 2008: Use of Bayesian Merging Techniques in a Multimodel Seasonal Hydrologic Ensemble Prediction System for the Eastern United States. J. Hydrometeor., 9, 866–884, doi: 10.1175/2008JHM980.1

MacLachlan, C., A. Arribas, K. A. Peterson, A. Maidens, D. Fereday, A. A. Scaife, M. Gordon, M. Vellinga, A. Williams, R. E. Comer, J. Camp, P. Xavier and G. Madec, 2014. Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. QJR Meteorol Soc, doi:10.1002/qj.2396.

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.

Mushtaq, S., Chen, C., Hafeez, M., Maroulis, J. and Gabriel, H., 2012: The economic value of improved agrometeorological information to irrigators amid climate variability. Int. J. Climatol., 32, 567–581.

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.

Roudier, P., Andersson, J.C.M., Donnellly, C., Feyen, L., Greuell, W. and Ludwig, F., (2016). "Projections of future floods and hydrological droughts in Europe under a +2°C global warming." Climatic Change 135(2): 341-355.

Saha, Suranjana and Coauthors, 2014: The NCEP Climate Forecast System Version 2 Journal of Climate J. Climate, 27, 2185–2208. doi: 10.1175/JCLI-D-12-00823.1

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. and Lettenmaier, D. P. (2011). "Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill." Hydrol. Earth Syst. Sci. 15(11): 3529-3538.

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.

Siegert, S., Bhend, J., Kroener, I., De Felice, M. (2014). SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate. https://CRAN.R-project.org/package=SpecsVerification

Singla, S., Céron, J. P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., & Vidal, J. P. (2012). Predictability of soil moisture and river flows over France for the spring season. Hydrology & Earth System Sciences, 16: 201-216.

Svensson, C., Brookshaw, A., Scaife, A. A., Bell, V. A., Mackay, J. D., Jackson, C. R., Hannaford, J., Davies, H. N., Arribas A., Stanley, S. (2015). "Long-range forecasts of UK winter hydrology." Environmental Research Letters 10(6): 064006.

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.

Thober, S., Kumar, R., Sheffield, J., Mai, J., Schäfer, D., and Samaniegoet, L. (2015). "Seasonal Soil Moisture Drought Prediction over Europe Using the North American Multi-Model Ensemble (NMME)." Journal of Hydrometeorology 16(6): 2329-2344.

Trigo, R. M., Pozo-Vázquez, D., Osborn, T.J., Castro-Díez, Y., Gámiz-Fortis, S., Esteban-Parra, M.J. (2004). "North Atlantic oscillation influence on precipitation, river flow and water resources in the Iberian Peninsula." International Journal of Climatology 24(8): 925-944.

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 basins worldwide. Water Resources Research, 49(5), 2729-2746.

Van Dijk, A. I. J. M., and G. A. Warren (2010), AWRA Technical Report 4, Evaluation Against Observations, WIRADA/CSIROWater for a Healthy Country Flagship, Canberra. http://www.clw.csiro.au/publications/waterforahealthycountry/2010/wfhc-awras-evaluationagainst-observations.pdf

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.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J. and Clark, M. (2016). Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill. Journal of Hydrometeorology 17(2): 651-668.

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.