Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions

Seasonal streamflow forecasts for Europe – II. Explanation of the skill

Wouter Greuell, Wietse H. P. Franssen, 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 second 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. Whereas the first paper presents the development and the skill evaluation of the system, this paper provides explanations for the skill. In WUSHP hydrology is simulated by running the Variable Infiltration Capacity (VIC) hydrological model with meteorological forcing from bias-corrected output of ECMWF's Seasonal Forecasting System 4 (S4). WUSHP is probabilistic. For the assessment of skill, hindcast simulations (1981-2010) were carried out. To explain skill, we first looked at the forcing and found considerable skill in the precipitation forecasts of the first lead month but hardly any significant skill for later lead months. Seasonal forecasts for temperature have more skill. Skill in summer temperature is related to climate change and more or less independent of lead time. Skill in February and March is unrelated to climate change. Sources of skill in runoff were isolated with Ensemble Streamflow Prediction (ESP) experiments. These revealed that beyond the second lead month simulations with forcing that is identical for all years (ESPall) produce more skill in runoff than the simulations forced with S4 output (Full Hindcasts). This occurs because interannual variability of the S4 forcing has insufficient skill while it adds noise. Other ESP-experiments show that in Europe initial conditions of soil moisture form the dominant source of skill in runoff. From April to July, at the end of the melt season, initial conditions of snow contribute significantly to the skill, also when forecasts start much earlier. Some remarkable skill features are due to indirect effects, i.e. skill due to forcing or initial conditions of snow and soil moisture at an earlier stage is stored in the hydrological state (snow and/or soil moisture) of a later stage, which then contributes to persistence of skill. Finally, predictability of evapotranspiration was analysed in some detail, leading among others to the conclusion that it is due to all potential sources of skill but mostly to forcing.

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.

This is the second paper about 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. WUSHP basically consists of simulations carried out with the Variable Infiltration Capacity (VIC) hydrological model, which uses bias-corrected output of forecasts from ECMWF's Seasonal Forecast System 4 (S4) as meteorological forcing. The system is probabilistic. In the first paper (Greuell et al., 2016), the set-up of WUSHP was described and spatial and temporal variations of skill and lack of skill in runoff and discharge in Europe were established by means of hindcasts. In general, considerable amounts of significant skill were found. Hot spots of significant skill in runoff are situated 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)

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

and West France (from December to May). The spatial pattern of skill in runoff is fading with increasing lead time but significant skill is left even at the end of the hindcasts (7 months).

The current paper deals with the causes of the skill and the lack of skill in WUSHP along two lines. The first line is an analysis of the skill in the most important variables of the S4 meteorological forcing. For S4 this was done earlier by Kim et al. (2012) for the boreal winter months with initialisation at the first of November. For that case, they found that S4 has no skill in the precipitation forecasts and some skill in the temperature forecasts for South Sweden, South Finland, the region south-east of Saint Petersburg and North Germany. Scaife et al. (2014) analysed the skill for the same target months and start period but with another prediction system, namely the Met Office Global Seasonal forecast System 5 (GloSea5). One of their maps (Fig. 4c) exhibits hardly any significant skill in the temperature forecasts for Europe but the adjacent map demonstrates that in North and South Europe forecasts of the North Atlantic Oscillation are correlated significantly with observed temperatures. This means that there is untapped predictability in the GloSea5 forecasts.

The second line of analysing sources of predictability will consist of so-called Ensemble Streamflow Prediction (ESP) experiments, experiments that are designed to isolate the effect of the different sources of predictability, which are in the case of the present study meteorological forcing, the initial conditions of soil moisture and the initial conditions of snow. ESP experiments were widely used in earlier studies (e.g. Wood and Lettenmaier, 2008, Shukla and Lettenmaier, 2011, Singla et al., 2011) and separated the sources of skill of runoff or discharge forecasts by considering forcing and the initial conditions. Koster et al. (2010) separated the soil moisture and snow initial conditions to skill..

The term ESP refers to a modelling technique that produces probabilistic hydrological hindcasts, which start from deterministic initial conditions that are as realistic as possible and vary from year to year. These initial conditions of the hindcasts are generated with a reference simulation, which is a simulation forced with meteorological observations. In ESP experiments, the hindcasts are forced with an ensemble of meteorological time series that is identical for each year. Usually the forcing is a selection from historic observations. With identical forcing for each year, all skill in ESP simulations derives from the initial conditions. ESP has been and is widely used in studies dealing with seasonal hydrological forecasting (e.g. by Day, 1985, Wood et al., 2005, Shukla and Lettenmaier, 2011, Singla et al., 2011, and Van Dijk et al., 2013).

Reverse-ESP experiments were introduced by Wood and Lettenmaier (2008) and also produce probabilistic hindcasts. In reverse-ESP simulations the deterministic forcing varies from year to year, is as realistic as possible and is usually derived from observations. The initial conditions consist of an ensemble of initial conditions that is identical for each year of the hindcasts. Usually this ensemble is selected from snapshots of the model state generated during a reference simulation. All skill in reverse-ESP experiments derives from the forcing.

Wood and Lettenmaier (2008), and Shukla and Lettenmaier (2011) compared the ESP and reverse-ESP simulations with a reference simulation. Their aim was to quantify what can be gained if the meteorological forcing or the initial conditions are improved from containing no information at all (in ESP and reverse-ESP, respectively), which is the case when they have a climatological distribution, to being quasi-perfect, i.e. when they are equal to the meteorological observations and the initial state of the reference simulation. Wood et al. (2016) extended this type of analysis by determining sensitivities of the streamflow to changes in information on meteorological forcing or the initial conditions. All of these studies basically looked at uncertainty in seasonal forecasts.

However, like Wood et al. (2005), Bierkens and Van Beek (2009) and Koster et al. (2010), we take a different philosophy by analysing skill. The skill of the ESP simulations will be compared with that of the standard WUSHP hindcasts, i.e. the simulations that are forced with S4 output and start from initial conditions generated with the reference simulation. We will refer to these simulations as “Full Hindcasts” (“climate-model-based hindcasts” according to Yuan et al., 2015). The ESP and the

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

reverse-ESP experiments will then quantify the effect on the total skill of removing one or more sources of skill. It should be noted that ESP experiments isolate skill due to initial conditions while they reveal the uncertainty due to forcing and reverse-ESP experiments isolate skill due to forcing while they reveal uncertainty due initial conditions.

Bierkens and van Beek (2009) investigated sources of skill for Europe. They found that in winter initial conditions dominate while in summer forcing and initial conditions are equally important. Singla et al. (2011) assessed the skill of hydrological predictions for France and concluded that over most plains the predictability of hydrological variables primarily depended on forcing, whereas it mainly depended on snow cover over high mountains. The Seine catchment area was an exception as the skill mainly came from the initial state of its large and complex aquifers. In summary, the relative contributions of the different sources of skill varies strongly among these studies. These inter-study differences are partly due to differences between the regions and the seasons that are investigated but also inter-study differences between the quality of the forcing and the hydrological models play a role.

To complement the analysis, we will also look at skill in another output variable of the hydrological model, namely evapotranspiration. Predictions of evapotranspiration are useful for planning of water level control in polders and of water use for irrigation and fertiliser application. As for runoff, we will exploit the ESP-experiments to isolate the different sources of predictability of evapotranspiration.

Thus, the objective of the present paper is to analyse, mostly at a pan-European scale, the sources of probabilistic skill of seasonal hydrological forecasts produced by WUSHP. The spatio-temporal patterns of skill themselves have been analysed and presented in a companion paper (Greuell et al. 2016), while the present paper focusses on skill attribution. The next section will describe the seasonal prediction system itself, the analysis approach as well as details of the various ESP experiments performed. We will present the skill in three variables of the climate forcing, followed by skill in runoff as present in the various ESP experiments, which allows attribution to either forcing or different initial conditions, and finally an analysis of skill in evaporation. We conclude with a discussion and conclusions. Additional figures are published in a supplement of this paper.

2 System and methods

2.1 The forecast system

The forecasts of WUSHP combine three elements, namely meteorological forcing from ECMWF's Seasonal Forecast System 4 (Molteni et al., 2011), bias correction with the quantile mapping method of Themeßl et al. (2011) and simulations with the Variable Infiltration Capacity (VIC) hydrological model (Liang at al., 1994). The skill of the system was assessed with hindcasts. These cover the period 1981-2010, were initialised on the first day of each month and have a length of seven months. The system is probabilistic (15 members), so a total of 30 (years) x 12 (months) x 15 (members), i.e. 5400 simulations, was carried out. In addition a single reference simulation was performed, in which VIC was run with a gridded data set of model-assimilated 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 initialisation states for each of the hindcasts. To spin up discharge, each 7-month hindcast simulation was preceded by a one month simulation with WFDEI forcing. Simulations were performed in naturalised flow mode on a 0.5˚ x 0.5˚ grid. More details about the set-up of the system and the hindcasts can be found in the companion paper (Greuell et al., 2016).

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

2.2 Methods of analysis and observations

In this paper we distinguish runoff, defined as the amount of water leaving the model soil either along the surface or at the bottom, from discharge, defined here as the flow of water through the largest river in each grid cell.

Skill is measured in terms of the correlation coefficient between the median of the hindcasts and the observations (R). We will designate R-values as significant for p-values less than 0.05. We also considered other metrics, namely Relative Operating Characteristics (ROC) area for terciles and the Ranked Probability Skill Score (RPSS). In the companion paper different metrics were compared and it was found that for all combinations of target and lead month the skill patterns in the maps are similar to a high degree. However, there was a subtle difference between ROC areas for the two outer terciles, with more significant skill found in the Below Normal than in the Above Normal tercile from February to September.

Unless mentioned otherwise, prediction skill of the hydrological variables is determined against the pseudo-observations (see Sect. 2.1). These have the advantages of being complete in the spatial and the temporal domain and to be available for all model variables. We will refer to this type of skill as "theoretical skill". In the companion paper theoretical skill for discharge was compared to "actual skill", which is the skill assessed with real observations. Real observations of discharge were acquired from the Global Runoff Data Centre, 56068 Koblenz, Germany (GRDC), gridded onto the 0.5˚ x 0.5˚ model grid and subdivided into observations for catchments larger than 9900 km^2 ("large-basins") and observations for catchments smaller than the area of the grid cells ("small basins"). Regarding the use of different types of "observations" for verification, Greuell et al. (2016) concluded that, in terms of R and on average across all target months and for lead month 2, the ratio of actual to theoretical skill was 0.67 for large basins and 0.54 for small basins.

For the determination of the skill of the meteorological variables of the S4 forecasts we used the WFDEI data. Here we took the non-bias-corrected version of the S4 data so that the resulting skill is that of the un-post-processed S4 data. We compared skill of the non-bias-corrected forcing with skill of the bias-corrected forcing and found negligible differences. This is not surprising because the bias corrections do not change the ranking of the values and the skill metrics mainly measure the ranking of the hindcasts relative to the ranking of the observations.

Like in the companion paper, skill was analysed on a monthly and not on a seasonal basis with the aim of achieving relatively high temporal resolution. 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.

In all result sections we will first analyse and explain at the level of the entire domain. We will then take out the most remarkable details of the summary plots and provide an explanation for them, e.g. for the summer peak in skill of predicting atmospheric temperature, for the May peak in skill of predicting runoff due to snow initial conditions and for the July peak in skill of predicting evapotranspiration.

2.3 The ESP experiments and surface water initialisation

In this paper the term "ESP" will include both the traditional ESP experiments and a reverse-ESP experiment. In total four ESP experiments were performed:

1) "ESPall" isolates the skill due to all initial conditions (soil moisture and snow). It takes the annually varying initial conditions from the reference simulation while for each year the atmospheric forcing consists of the an ensemble of the same fifteen S4 hindcasts. More specifically, we selected member 1 from the 1981 hindcasts, member 2 from the 1983 hindcasts, etc. We did not select atmospheric forcings from observations (e.g. WFDEI), which is the strategy employed in

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

most published ESP experiments. By selecting the forcing from the S4 hindcasts, the ESP experiments remain as close as possible to the Full Hindcasts. This way we also avoided reproducing the reference simulation as a member of the hindcasts, which is an issue since the reference simulation is used as pseudo-observation for verification of the hindcasts.

2) "ESPsoilm" isolates the skill due to the initial conditions of soil moisture. ESPsoilm is identical to ESPall but in all ESPsoilm simulations snow initial conditions are taken as the average of the snow initial conditions of all 30 years of the hindcasts.
3) "ESPsnow" isolates the skill due to the initial conditions of snow. ESPsnow is identical to ESPall but in all ESPsnow simulations soil moisture initial conditions are taken as the average of the soil moisture initial conditions of all 30 years of the hindcasts.
4) "revESP" isolates the skill due the meteorological forcing. It takes the annually varying forcing from the probabilistic S4 hindcasts while the initial soil moisture and snow conditions are equal to the average of the soil moisture and snow initial conditions of all 30 years of the hindcasts. Unlike in many other published reverse-ESP experiments, atmospheric forcing is not taken from observations but from the S4 hindcasts so that forcing is identical to the forcing of the Full Hindcasts.

Thus, all experiments produce, like the Full Hindcasts, a 15 member probabilistic hindcast, which is important since ensemble size affects skill metrics (Richardson, 2001). All of these simulations were preceded by a one month run with reference forcing (WFDEI) with the single aim of initialising the amount of discharge in the rivers. This has no effect on most of the analyses of the paper, since these are made in terms of runoff. Where discharge is analysed the effect of discharge initialisation is, due to the limited residence time of water in the rivers, restricted to the first lead month of the simulations (see Yuan, 2016).

3 Explanations of skill in hydrological variables

3.1 Skill in the meteorological forcing

In this sub-section, the skill of the meteorological output variables of S4 will be analysed, limiting the attention to the three most important input variables of VIC, namely precipitation, two-meter temperature and incoming short-wave radiation. The WFDEI data are use as reference.

A summary of skill in the precipitation hindcasts is given in Fig. 1b, which plots the fraction of all coloured cells in Fig. 1a that have statistically significant R values. During the entire year, there is considerable skill for lead month 0 but skill declines very rapidly. In fact, after the first lead month the fraction of cells with significant skill approaches its theoretical value for the case that the hindcasts have no skill at all (5%). Hence, from lead month 1 on, skill is almost negligible. Regarding lead month 0, January, February and March have more skill than the other months. Figure 1a shows the map of skill for January. For the other target months maps are not shown here but hot spots of consistent skill, i.e. with a duration of at least three target months but in this case only for lead month 0, are situated on the Iberian Peninsula from November to March, in West Norway from January to April, in Greece and West Turkey from December to February and in Scotland from December to March. All these occurrences of consistent skill are restricted to the winter half of the year and coastal regions, suggesting an effect of the initial state of the sea surface temperature.

Figure 2 shows important aspects of skill in the two-meter temperature hindcasts. One aspect is the possible contribution of a 30-year trend, which could be related to greenhouse warming, to the skill. This aspect was investigated by a separate analysis of skill, in which both the hindcasts and the observations were detrended in a pre-processing step. For lead month 0 the hindcasts have significant skill in the largest part of the domain (panel a). At longer lead times, the percentage of cells with significant skill quickly drops towards the theoretical limit of no skill (5%) but there are quite a few exceptions, namely:

- For lead month 1, February and March temperatures are predicted with significant skill in a considerable part of the domain (44% in February; 53% in March). In both months the region with skill is more or less contiguous and comprises the Russian part of the domain, the Ukraine and the regions bordering the southern part of the Baltic Sea (panels d and e). In February the region of skill extends towards Central Europe. In March it also comprises North Fennoscandia. This skill hardly diminishes by detrending the data (panels b and c), suggesting that the skill is not related to climate change. Indeed, in February and March the observed trend (in the WFDEI data set) is insignificant across most of the domain (11% of the domain in February and 18% in March) and, more importantly here, it is insignificant in the regions with significant skill in the temperature hindcasts (panel f demonstrates this for March). We conclude that the temperature skill in February and March as lead month 1 must be due to initial conditions of the climate model (see also the discussion on Fig. 10).
- The three summer months exhibit significant skill at all lead times in much more than 5% of the domain (a range from 22 to 56% for all combinations of the three summer months and all lead months beyond lead month 0), see panel a. Also, the fraction of cells with significant skill is not a function of lead time, which is the type of behaviour that Yuan (2016) also found for the Yellow River basin. Since panels b and c demonstrate that the skill more or less vanishes when the temperature hindcasts and observations are detrended, we conclude that this skill is related to climate change.

  If hindcasts and observations are correlated due to a trend, as is the case here in parts of the domain during the summer months, a prerequisite is that there is a trend in both the hindcasts and in the observations. Indeed, both types of time series have a maximum in significant trends in summer. In the hindcasts and on average over all lead times beyond the first month, the summer months exhibit significant trends in almost the entire domain (95%), versus 79% in the other months of the year, on average. Similarly, observed trends are significant during the three summer months in 67% of the domain, versus only 24% in the other months of the year, on average. These percentages also show that significant trends occur much more in the hindcasts than in the observations. So, in summer it is mainly the observations that limit the occurrence of climate-change-related skill in the temperature hindcasts. This is illustrated by panels g-i. Panel h shows that the trends of the hindcasts for July as lead month 5 are significant across almost the entire domain (99% of the domain). However, according to panel i only 69% of the domain has a significant trends in the observed July temperatures. Indeed, the patterns of significance of panel g (skill in the temperature hindcasts) and i (significance of observed trends) agree to a large extent. We finally like to note here that the fraction of cells with significant trends in the hindcasts is a function of the target month but independent of lead time (with the exception of the first lead month, when significant trends occur in a smaller part of the domain). As a result, skill is not dependent on lead time and we conclude that skill that is independent of lead time is an indication that skill might be due related to climate change.
- April, May and September mix the behaviour of February and March, which have skill due to initial conditions of the climate model, with the skill of the summer months, which show skill related to climate change (panel c).
- January has a considerable amount of significant skill but only for lead month 2 (42% across the domain). This skill occurs in a stroke of land reaching from England to Russia, which vaguely coincides with the region in which Kim et al. (2012) found skill in the S4 temperature hindcasts for the three winter months. However, as this skill is not found in adjacent lead and target months, we speculate that this skill is spurious.

Since short-wave incoming radiation is important for evapotranspiration, we finalise this sub-section with a short analysis of its predictability (Fig. 3). Skill in the hindcasts of short-wave incoming radiation is limited to a small fraction of the domain for lead month 0.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

3.2 Sources of skill in runoff and discharge

While Sect. 3.1 dealt with predictability of the meteorological forcing, this sub-section analyses the effects of skill in the forcing and of other sources on the predictability of runoff and discharge (discharge is only considered in Fig. 4). We first address the question how much of the skill in the runoff hindcasts is linked to climate change. Similarly to the analysis for atmospheric temperature, the pseudo-observations and the median of the hindcasts of runoff were detrended and the skill was compared to that of the undetrended data sets. We found that for lead month 2 and averaged over all months of the year, the fraction of cells with a significant R decreased from 58.7 to 57.4% due to detrending, a difference of 1.3%. This difference is much smaller than the decrease for temperature (11.8%). We conclude that climate change contributes very little to skill in runoff. All analyses of this sub-section hereafter pertain to undetrended data.

In the remainder of this sub-section the skill in the hindcasts of runoff during the various Ensemble Streamflow Prediction (ESP) experiments is discussed. Sect. 3.2.1 deals with the ESPall simulations (skill entirely due to model initialisation of both soil moisture and snow). Section 3.2.2 discusses results of the other ESP experiments, namely those that isolate skill due to soil moisture initialisation (ESPsoilm), snow initialisation (ESPsnow) and the meteorological forcing (revESP). Unless indicated otherwise, the pseudo-observation are used for verification.

3.2.1 The relative importance of initial hydrological conditions versus meteorological forcing

Figure 4 compares the ESPall simulations with the Full Hindcasts in terms of the fraction of cells with a significant R. Panels a and b show the result for runoff and discharge, respectively, using the pseudo-observations, so calculations for all cells of the domain contribute to the result. While the behaviour hardly differs between runoff and discharge (see the companion paper), systematic differences between skill in the Full Hindcasts and ESPall are revealed. In lead month 0, skill is higher in the Full Hindcasts than in the ESPall simulations for all target months of the year. Beyond lead month 1, the reverse occurs. Lead month 1 is transitional with the order of skill depending on the time of the year. Panels c (for large catchments) and d (for small catchments) compare actual discharge skill, i.e. skill determined with real discharge observations, of the Full Hindcasts with that of the ESPall simulations. As discussed in the companion paper, domain-average actual skill is less than domain-average theoretical skill but the reversal of skill after lead month 1 found with the pseudo-observations is confirmed with real observations.

We hypothesize that the reason for the reversal lies in the competition of signal and noise in the forcing of the Full Hindcasts. The ESPall forcing is the same for each year, so its interannual variation does not contain a signal nor noise. However, the forcing of the Full Hindcasts varies from year to year. During the first lead month this forcing has considerable skill (see Sect. 3.1), so the signal-to-noise ratio of the forcing is relatively high. This enhances skill in the Full Hindcasts with respect to ESPall. At long lead times the interannual variation in the forcing hardly contains a signal, with the exception of some limited skill in the temperature hindcasts (see Sect. 3.1), so the signal-to-noise ratio is low. Noise in the meteorological hindcasts reduces skill in the Full Hindcasts with respect to ESPall. Figure 4 demonstrates that averaged across the domain, the reversal between skill enhancement due to the signal in the forcing and skill reduction due to noise in the forcing occurs at some time between the first and the third lead month, with the exact timing of the reversal depending on the target month.

Figure 5 shows that, averaged across the domain, meteorological forcing alone (revESP) always causes much less significant skill than the initial conditions of soil moisture and snow together (ESPall). However, skill in runoff due to forcing exceeds the skill in the forcing variable to which runoff is most sensitive, precipitation (compare Fig. 5 with Fig. 1). Whereas predictability of precipitation is limited to the first lead month (on average over the year, 6% of the domain has significant skill in lead months 1 and 2, just 1% more than the percentage of cells in the case of no true skill at all), significant skill in runoff due to forcing is

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

more widespread for lead months 1 and 2 (on average over the year in 23 and 15 % of the domain, respectively). We explain the enhanced skill in runoff by an indirect effect of the skill of the precipitation forcing in the first lead month, which gradually adds some skill to the model states of soil moisture and snow. This, in turn, leads to the mentioned skill in runoff during later lead months. Also, the skill in the hindcasts of temperature (Fig. 2) contributes to the skill in runoff.

3.2.2 The relative contributions of soil moisture and snow initial conditions

Figure 5 compares the skill in the hindcasts of runoff of all four ESP experiments for two lead months (0 and 2). At both lead times, initialisation of soil moisture is the dominant source of skill in Europe. This is true for all lead times (not shown here). From April to July, a considerable part of Europe has significant skill derived from snow initialisation but, in terms of the pan-European metric used here, this part never becomes larger than the part with significant skill due to soil moisture. Skill due to snow initialisation reaches a maximum in May and June (for lead months from 1 to 5), resulting in a maximum in skill in the ESPall-simulations and the Full Hindcasts for these months (at most lead times). This rapid rise in skill due to snow initialisation at the transition from April to May explains a remarkable feature that we noticed in the companion paper, namely an increase in runoff skill with lead time at this time of year.

Figure 6 illustrates that skill due to snow and soil moisture initialisation are more or less additive. Copies of the patterns of skill due to soil moisture initialisation e.g. in Africa, on the Iberian Peninsula and in West France (panel a) are found in the map of skill due to both soil moisture and snow initialisation (panel c). Small regions with considerable skill due to snow initialisation (panel b) like those near Stockholm, in South-east Czechia and South-east Austria also stick out as foci of skill in the combined initialisation map (panel c). Where both soil moisture and snow initialisation cause moderate skill, e.g. in South Finland, the combined experiment exhibits more significant skill. The additive behaviour of skill in the two initialisation components is also visible in Fig. 5. The skill in the ESPall experiment exceeds the skill in the ESPsoilm experiment from April to July, when snow contributes considerably to predictability. For target months from August to March, when snow contributes little to predictability, the percentages of cells with significant skill in ESPall and ESPsoilm are almost identical.

Figure 7 zooms in on the experiment that isolates skill due to snow initialisation (ESPsnow), giving the example of a time series of skill as a function of lead time, after initialisation on March 1. One observation is that skill does not gradually decrease with time but has a maximum during the snow melt season. We like to note that locally skill is hardly generated during the part of the melt season when there is snow in each year since in VIC the rate of snow melt is hardly sensitive to snow pack thickness (Sun et al., 1999). Skill is generated towards the end of the melt season when snow melt differs from year to year because snow stops to be available for melt at different dates due to different initial amounts of snow. So, the initial snow conditions cause skill because of interannual variation in the duration of time that snow is present at the end of the melt season and not because of interannual variation in melt during the central part of the melt season. Of course, the timing of the end of the melt season differs regionally and with elevation, which largely explains the patterns of skill visible in the maps of Fig. 7. An example is Scandinavia, where the first skill occurs near the coasts of South Norway and Sweden in April and the latest skill occurs in the Norwegian mountains in July (we ascribe the skill in South-east Sweden in July and August to chance). It is relevant to note further that the maps of Fig. 7 are affected by the fact that VIC has higher vertical resolution than its horizontal resolution may suggest, by performing simulations in multiple elevation bands within each grid cell, accounting for sub-grid variations in topography. Therefore, sub-grid topography leads to spreading of the snow skill signal of individual cells over longer periods of time.

To finish the analysis of the ESPsnow experiment, Fig. 8 analyses a remarkable feature. In ESPsnow, hindcasts for May have less skill when the hindcasts are initialised on May 1 (panel a) compared to initialisation during preceding months (February, March or April, panel b is for initialisation on April 1). Similar counterintuitive results are found for June and July as target

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

months. This result is counterintuitive because in simulations with initialisation on May 1 there is, due to the use of pseudo-observations for verification, perfect knowledge about snow conditions on that date. With initialisation on April 1, snow conditions on May 1 differ from those of the pseudo-observations, which by itself must lead to less skill in May runoff. However, there is compensation for this direct effect by an indirect effect through soil moisture. In ESPsnow, soil moisture has no skill on the date of initialisation, e.g. May 1. However, the knowledge of the snow conditions on April 1 leads via skill in snow melt in April to some skill in soil moisture on May 1 (panel c), which then leads to skill in runoff in May. Since we find more skill in May runoff after snow initialisation on April 1 than after snow initialisation on May 1, the gain of skill in the runs starting on April 1 due to the indirect effect via knowledge of soil moisture on May 1 overcompensates for the loss of skill in the same runs due to the direct effect of less knowledge of the amount of snow on May 1.

Returning to Fig. 5, we notice that meteorological forcing always causes less significant skill than the initial conditions of soil moisture. During the first lead month there is more skill due to the forcing than due to snow initial conditions. For later lead months this order depends on the season.

Finally, the ESP experiments were exploited to attribute the hotspots of significant skill for lead month 2, listed in the companion paper, to the different potential sources of skill. This was done by inspecting maps of skill (like those of Fig. 6) of the Full Hindcasts and the ESP experiments. Results are summarised in Table 1. Almost all of the significant skill in the hotspot regions is due to the initial conditions of soil moisture. Exceptions are formed by the target months from April to July when skill is caused by a mix of the initial conditions of snow and soil moisture in regions with significant snow melt skill. In these cases the relative contributions of the two sources varies in time and space but soil moisture is more important than snow, except in Fennoscandia where in June snow dominates and in July both sources are of about equal importance.

3.3 Skill and source of skill in evapotranspiration

Because hindcasts of evapotranspiration have intrinsic value (see the introduction) and in order to demonstrate the power of the use of the pseudo-observations and the ESP experiments, this section analyses skill in the hindcasts of evapotranspiration. First the annual cycle of skill in evapotranspiration will be analysed, and then the skill for two months with more than annual average skill, April and July, will be decomposed.

Levels of predictability (Fig. 9a) are higher than for precipitation (Fig. 1), similar to those for temperature (Fig. 2) and lower than those for runoff (Fig. 4a). Figure 9b isolates the diverse contributions to skill for lead months 0 and 2 by showing the skill for the Full Hindcasts and three ESP experiments. Averaged over the year, forcing contributes more and initial soil moisture less to predictability in evapotranspiration than to predictability in runoff. Initial snow is the least important of the three sources of skill.

Focusing on lead month 2, there is hardly any skill in the hindcasts from November to March (9% of the domain, on average over these months), with the exception of January (18%) when the region of skill (Germany and Benelux) is part of a larger region of skill in the temperature hindcasts for the same target and lead month. We blame the winter minimum of skill in evapotranspiration to the low levels of evapotranspiration and the low levels of skill in the temperature forecasts for the same period. The next month (April) exhibits the highest level of skill of all months (44% of the domain), mainly due to forcing and with smaller contributions by the initial conditions of soil moisture and snow. From May to September there is quite some significant skill (23% of the domain, on average over these months). Whereas in May forcing is still the most important contributor to skill, initial conditions of soil moisture form the main contributor from June to October. We speculate that this shift in the order of importance between forcing and soil moisture is due to the amount of variability in soil moisture. In Europe in spring (April, May), soil moisture variations are relatively small and hence hardly contribute to variations in evapotranspiration. Later in the year (June to September), soil moisture is often available in limited amounts, so variations are

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

larger and hence contribute more to variations in evapotranspiration. Snow initial conditions contribute to skill only from April to July.

The contribution of trends to predictability of evapotranspiration is summarised in panel c, for lead months 0, 1 and 2. For lead month 2 and averaged over all target months of the year, detrending leads to a decrease in the fraction of cells with a significant R from 17.6 to 13.8%, a difference of 3.8%. The contribution of climate change to skill in evapotranspiration is less than its contribution to skill in temperature (a difference of 11.8%) but larger than its contribution to skill in runoff (a difference of 1.3%). Climate change affects the skill in evapotranspiration during the same part of the year as it affects the skill in atmospheric temperature (Fig. 2c), namely from April to September and in November (for lead month 0). However, whereas during the three summer months the skill in the temperature hindcasts is almost exclusively linked to climate change, a considerable part of the domain still exhibits skill in evapotranspiration after detrending.

To provide a deeper understanding of the skill, the skill in April and July is analysed in some detail. Fig. 10 deals with April as lead month 2, showing the skill in evapotranspiration from the Full Hindcasts in panel a and from the revESP-experiment in panel b. Regions of skill, mainly a stroke of land from South Fennoscandia to the Black Sea, are the same in the Full Hindcasts and in revESP though skill is somewhat degraded in revESP. This indicates that meteorological forcing causes most, though not all, of the skill. Indeed, Fig. 2e (skill in temperature for March as lead month 1) and Fig. 10c (skill in temperature for April as lead month 2) show that the temperature forecasts of the preceding lead month and the lead month considered contain skill in the same regions. We conclude that much of the skill in evapotranspiration is due to skill in the temperature hindcasts. The remaining part of the skill is due to initial conditions (Fig. 9b shows this for the entire domain). We found limited amounts of skill in the ESPsnow and the ESPsoilm simulations for April in the same stroke of land from South Fennoscandia to the Black Sea (not shown here). This means that in that region initial conditions of the hydrological model on February 1 provide some skill to the hindcasts of evapotranspiration. We like to note that this could be consistent with the conclusion in Sect. 3.1 that the skill in the temperature hindcasts of February and March in this same region are due to the initial conditions of the climate model. These initial conditions could e.g. be sea surface temperatures but also the local state of snow and/or soil conditions. In the latter case, the two types of predictability in the mentioned regions have the same or a similar source. Initial conditions of snow and/or soil conditions lead to skill in the temperature hindcasts of the climate model (S4) and initial conditions of snow and soil moisture lead to skill in the evapotranspiration hindcasts of the hydrological model (VIC).

During the summer months and for all lead times, skill in evapotranspiration occurs in two regions, namely the southern part of the Mediterranean, and West and North Norway. Fig. 11 shows target month July as lead month 5, as an example. Whereas panel a is for the Full Hindcasts, panels b-d depict the maps for the three ESP experiments and panel e shows skill for the Full Hindcasts after detrending. From the ESP experiments it is concluded that the skill in the Mediterranean in due to soil moisture initial conditions (panels b-d). So, in this particular case, knowledge of soil moisture conditions on February 1 still yields skill in evapotranspiration in July. This skill in the Mediterranean is not affected by detrending (compare panels a and e), so it does not have a climate change component.

The skill in Norway has a more complicated origin. The ESP experiments show that it is due to a mix of initial snow conditions and forcing. The effect of the initial snow conditions (on February 1) can be understood with the help of the analysis of runoff skill in the ESPsnow simulation (Fig. 7), from which it was concluded that July is the end of the melt season in much of Norway. Therefore, in this country and in July the timing of the disappearance of snow cover varies from year to year. This then has a considerable effect on evapotranspiration since bare soil has, compared to snow, higher surface temperatures and hence more evapotranspiration in summer. The contribution to skill by forcing (panel d) fades with but is not removed by detrending (not shown here), so it has a part that is related to climate change and a part that is unrelated to climate change. The climate-change-

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

related skill due to forcing resides in the temperature hindcasts, which have significant skill in this region at all lead times (Fig. 2g). The non-climate change related skill in the revESP simulation for July is likely an indirect effect of the skill in the forcing (especially precipitation) in the first lead month (February). This leads to skill in snow water equivalent towards the end of February, which fades but has not disappeared completely on July 1(panel f) and then causes skill in evapotranspiration at the end of the melt season.

6 Discussion

The Ensemble Streamflow Prediction (ESP)-experiments of this study show that in Europe initial conditions of soil moisture are the largest source of skill in the seasonal streamflow forecasts produced with WUSHP. In terms of domain averages, this is true for all lead and target months. Contributions to skill by the initial conditions of snow and by the meteorological forcing are mostly much smaller. To our knowledge, two other studies analysed the skill of hydrological seasonal forecasts for Europe with ESP-like experiments, namely Bierkens and Van Beek (2009) and Singla et al. (2011). Results of these two studies were summarised in the introduction. However, the conclusions of Singla et al. (2011) are incomparable with those of the present study as they used the uncertainty strategy to analyse results while we analysed skill (see Sect. 1).

Comparing our results with those of Bierkens and van Beek (2009), the difference is that meteorological forcing contributes more to skill in Bierkens and van Beek (2009), at least in summer. This difference might be due to the quality of the forcing. Bierkens and van Beek (2009) developed an analogue events method to select, on the basis of annual SST anomalies in the North Atlantic, annual ERA40 meteorological forcings, which they used as forcing for their hydrological model. One might speculate that in Europe their semi-statistical forcing is more skilful than the S4 forcing used in WUSHP. This speculation is consistent with Scaife et al. (2014), who compared the skill of the dynamical GloSea5 temperature hindcasts with statistical hindcasts based on (dynamical) GloSea5 hindcasts of the North Atlantic Oscillation. Over the European continent the statistical hindcasts produced more significant skill than the dynamical hindcasts. Scaife et al. (2014) concluded that there is untapped predictability in the dynamical GloSea5 forecasts. Probably the same is true for the S4 hindcasts.

The dominance of soil moisture initial conditions also extends to the hotspot regions and periods of skill (Table 1). The question of the cause of skill linked to soil moisture can be deepened by another level as in Shukla and Lettenmaier (2011). The underlying idea is that this type of skill increases with the interannual variability of soil moisture at the date of initialisation and that this skill is gradually eliminated during the course of the simulation by interannual variability in processes like rain fall and snow melt. The question is to what extent hotspots of skill linked to soil moisture initialisation are due to the cause of the skill and to what extent they are due to a lack of interannual variability in the processes that eliminate the skill? Figure 12 helps answering this question for the skill found in the runoff hindcasts of August as lead month 2 with a simple method of analysis. Panel a shows the standard deviation of total modelled soil moisture (σ_{SM}) on the day of initialisation (June 1), taken from the reference simulation. Panel b depicts the standard deviation of total rain fall (σ_{RF}) during the course of the simulation (June – August), taking from the WFDEI data set, which is an important skill-eliminating factor. These two quantities were combined into an estimate of the skill (S_{est}):

$$S_{est} = \exp\left(-\frac{{\sigma_{RF}}^2}{{\sigma_{SM}}^2}\right) \qquad (1)$$

This estimate (panel c) needs to be compared with the skill of the hindcasts, mapped in panel d in terms of R. The two maps are not expected to be exactly equal, not only because of the simplicity of the estimation method but also because S_{est} is not a correlation coefficient. However, in the limits S_{est} has the desired properties. It is equal to zero for the cases of constant initial amounts of soil moisture or infinite variability in rain fall. It is equal to one for the cases of infinite variability in soil moisture or

constant rain fall. The correlation coefficient between the patterns in panels c and d is highly significant (0.67) and the hotspot regions of skill are the same in both panels, namely the northern part of Fennoscandia and the southern part of the Mediterranean. So, in the case of August as lead month 2 the estimation method is reasonably successful in computing the pattern of skill in the hindcasts with the simple means of the WFDEI data set and model calculations from the reference simulation. The additional merit of the estimation method is the deeper understanding of the cause of the skill in the two hotspot region. Northern Fennoscandia is a hotspot because the amount of interannual variability in initial soil moisture is larger than elsewhere (panel a). The southern part of the Mediterranean is a hotspot because the amount of interannual variability in rainfall is less than elsewhere (panel b).

A remarkable result of our work is the reduction of the skill beyond lead month 1, when the annually varying S4 forcing (Full Hindcasts) replaces forcing that is identical for all years (ESPall). This result is counter-intuitive but, as we discussed, a logical consequence of forcing with interannual variation that has no or insufficient skill, as the S4 forcing. We found no other publication that confirmed this result, though some studies (e.g. Singla et al., 2011, and Mackay et al., 2015) found little overall difference in skill between Full Hindcasts and ESPall simulations. In conflict with our result, skill is enhanced due to adding meteorological hindcasts, also at longer leads, according to the studies of Yuan et al. (2013), Thober et al. (2015) and Yuan (2016). This conflict can be explained if the meteorological hindcasts of the mentioned studies are more skilful than those of the present study. Indeed, Europe is a region with relatively little skill in meteorological hindcasts (Kim et al., 2012, Scaife et al., 2014, and Baehr et al., 2015). Effects of regional differences in the skill of the forcing on the relative skill of full hindcasts and ESP simulations are mentioned by Wood et al. (2005), who reported that full hindcasts for the western United States have practically no skill improvement over ESP, except for some regions and seasons with predictability of the forcing due to ENSO.

This simple method of analysis helped to bring the understanding of the skill in northern Scandinavia and the southern Mediterranean to a deeper level but it was less successful for the other hotspots. A more thorough analysis along these lines and a deeper understanding of skill in the hindcasts is left for future work.

This superiority of the ESPall-hindcasts with respect to the Full Hindcasts raises the question whether one should issue ESPall-based hindcasts and not the Full Hindcasts in an operational version of WUSHP, for forecasts beyond the first two lead months. The logical answer is “yes” but such a strategy should then be reconsidered regularly. The S4 forecasts could, and most likely will, become more skilful with time in the future (see Scaife et al., 2014). That would then lead to improved Full Hindcasts, which might surpass the ESPall-based hindcasts in terms of forecast skill. We like to note here, that regarding the effect of improvements in forcing on hydrological forecasts, small increases in forcing skill may lead to larger increases in discharge skill (Wood et al., 2016).

In this study we analysed the effect of trends on predictability by comparing skill for undetrended and detrended observations and hindcasts. The effect of trends was almost negligible for runoff, considerable and in summer even dominant for atmospheric temperature and of intermediate magnitude for evapotranspiration. While for academic purposes the distinction between climate-change related and non-climate change related skill may be relevant, this is not the case practical applications. To our knowledge, operational forecast systems issue their predictions without considering trends, i.e. predictions are issued by taking a historic period covered by observations or hindcasts as reference.

This study demonstrates the power of using pseudo-observations for verification. These data cover all cells of the entire domain and they are available for all hydrological model variables, e.g. runoff and evapotranspiration, which have no equivalent (runoff) or are sparse (evapotranspiration) in the realm of real observations. Many features of skill would not have been detectable with real observations. At the same time, we like to note that actual skill obtained with real observations is generally less than theoretical skill obtained with pseudo-observations, as discussed extensively in the companion paper.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

This paper mainly dealt with domain-averaged skill though we also zoomed in on local features, e.g. in Figs. 2, 7 and 11. There is a multitude of other local features in the skill plots that could be explained. More detailed local analyses, e.g. at basin level, are left for future work.

7 Conclusions

The present paper explains skill in the hindcasts of WUSHP, a seasonal hydrological forecast system, applied to Europe. We first analysed the meteorological forcing and found considerable skill in the precipitation forecasts of the first lead month but negligible skill for later lead times (Fig. 1). Seasonal forecasts for temperature have more skill (Fig. 2). Skill in summer temperature is related to climate change occurring in both the observations and the hindcasts, and more or less independent of lead time. Skill in North-East Europe in February and March is unrelated to climate change and must hence be due initial conditions of the climate model (System 4). There is hardly any skill in the hindcasts of incoming short-wave radiation (Fig. 3).

Sources of skill in runoff were isolated with Ensemble Streamflow Prediction (ESP) experiments. These revealed that, beyond the second lead month, simulations with forcing that is identical for all years but with "perfect" initial conditions (ESPall) produce, averaged across the model domain, more skill in runoff than the simulations forced with S4 output (Full Hindcasts), see Fig. 4. This occurs because interannual variability of the S4 forcing adds noise while it has hardly any skill. Other ESP-experiments (Fig. 5) show that in Europe initial conditions of soil moisture form the dominant source of skill in runoff. From April to July, initial conditions of snow contribute significantly, with a domain-mean maximum in May and June. The timing of that maximum (Fig. 7) varies spatially and coincides with the end of the melt season when snow melt differs from year to year because snow stops to be available for melt at different dates. Similar to the dominance of soil moisture at the scale of the entire domain, all regional and temporal hotspots of skill in runoff found in the companion paper are due to initial conditions of soil moisture, with smaller or larger contributions by the initial conditions of snow from April to July in hotspot regions with snow fall in earlier months. We further show that skill due to snow and soil moisture initialisation is more or less additive (Fig. 6).

Some remarkable skill features are due to indirect effects, i.e. skill due to forcing or initial conditions of snow and/or soil moisture is, during the course of the model simulation, stored in the hydrological state (snow and/or soil moisture), which then contributes to skill. An example (Fig. 8) occurs in the ESPsnow (skill is only caused by initial conditions of snow) hindcasts for the months of May, June and July, which have less skill when initialised at the beginning of the month compared to initialisation during preceding months. Another example is the skill in runoff in the revESP (skill only due to forcing) experiment, which exceeds the skill in the forcing variable to which runoff is most sensitive, precipitation. This can largely be explained by the skill of the precipitation forcing during the first lead month, which adds skill to the model states of soil moisture and snow. This then leads to skill in runoff during later lead months.

Finally, predictability of evapotranspiration was analysed in some detail. Levels of predictability (Fig. 9a) and the annual cycle of skill are similar to those for temperature. For most combinations of target and lead months, forcing forms the most important contributor to skill but for lead month 2 initial conditions of soil moisture dominate from June to October (Fig. 9b). In April, a stroke of land from South Fennoscandia to the Black Sea exhibits skill (Fig. 10), which is mainly due to skill in the temperature forecasts with smaller contributions from the initial conditions of snow and soil moisture. During the three summer months and for all lead times, skill in evapotranspiration occurs in two regions (Fig. 11). Skill in the southern part of the Mediterranean in due to soil moisture initial conditions. Skill in Norway is due to a mix of initial snow conditions and forcing.

The applied methods of analysis are not suitable for giving quantitative advice on what would be the best investment for increasing the amount of skill of WUSHP. However, since initial soil moisture is the dominant source of predictability, a large

effect can be expected from assimilation of soil moisture observations with the modelled state of soil moisture (see e.g. Draper and Reichle, 2015). In addition, snow water equivalent could be assimilated with the modelled state of snow (see e.g. Griessinger et al., 2016). Improving the calibration of VIC would be another obvious road towards improvement of the seasonal predictions discussed in this paper. This should lead to higher actual skill but not necessarily to more theoretical skill, see the discussion section of the companion paper.

References

Baehr, J., Fröhlich, K., Botzet, M., Domeisen, D. I., Kornblueh, L., Notz, D., ... & Müller, W. A. (2015). The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model. Climate Dynamics, 44(9-10), 2723-2735.

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.

Day, G. N. (1985). Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management, 111(2), 157-170.

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.

Draper, C., & Reichle, R. (2015). The impact of near-surface soil moisture assimilation at subseasonal, seasonal, and inter-annual timescales. Hydrology and Earth System Sciences, 19(12), 4831.

Greuell, W., W. H. P. Franssen, H. Biemans and R. W. A. Hutjes. Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations. Submitted to Hydrol. Earth Syst. Sci.

Griessinger, N., Seibert, J., Magnusson, J., & Jonas, T. (2016). Assessing the benefit of snow data assimilation for runoff modelling in Alpine catchments. Hydrol. Earth Syst. Sci., 20, 3895-3905.

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.

Kim, H. M., Webster, P. J., & Curry, J. A. (2012). Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. Climate Dynamics, 39(12), 2957-2973.

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.

Mackay, J. D., Jackson, C. R., Brookshaw, A., Scaife, A. A., Cook, J., & Ward, R. S. (2015). Seasonal forecasting of groundwater levels in principal aquifers of the United Kingdom. Journal of Hydrology, 530, 815-828.

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.

Richardson, D. S. (2001). Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Quarterly Journal of the Royal Meteorological Society, 127(577), 2473-2489.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., ... & Hermanson, L. (2014). Skillful long-range prediction of European and North American winters. Geophysical Research Letters, 41(7), 2514-2519.

Shukla, S., & Lettenmaier, D. P. (2011). Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill. Hydrology and Earth System Sciences, 15(11), 3529-3538.

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).

Sun, S., Jin, J., & Xue, Y. (1999). A simple snow-atmosphere-soil transfer model. Journal of Geophysical Research: Atmospheres, 104(D16), 19587-19597.

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., & Samaniego, L. (2015). Seasonal soil moisture drought prediction over Europe using the North American Multi-Model Ensemble (NMME). Journal of Hydrometeorology, 16(6), 2329-2344.

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., Kumar, A., & Lettenmaier, D. P. (2005). A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States. Journal of Geophysical Research: Atmospheres, 110(D4).

Wood, A. W., & Lettenmaier, D. P. (2008). An ensemble approach for attribution of hydrologic prediction uncertainty. Geophysical Research Letters, 35(14).

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., & 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.

Yuan, X. (2016). An experimental seasonal hydrological forecasting system over the Yellow River basin – Part 2: The added value from climate forecast models, Hydrol. Earth Syst. Sci., 20, 2453-2466, doi:10.5194/hess-20-2453-2016.

Yuan, X., Ma, F., Wang, L., Zheng, Z., Ma, Z., Ye, A., & Peng, S. (2016). An experimental seasonal hydrological forecasting system over the Yellow River basin – Part 1: Understanding the role of initial hydrological conditions. Hydrol. Earth Syst. Sci., 20, 2437-2451, doi:10.5194/hess-20-2437-2016.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

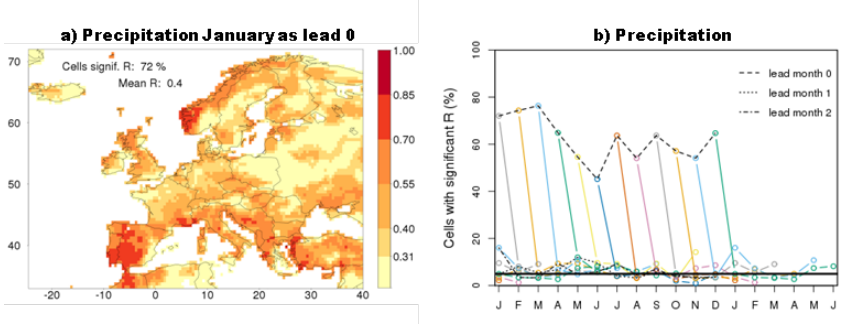

Figure 1: Skill of the precipitation hindcasts. Panel a shows a map of the correlation coefficient between the observations and the median of the hindcasts (R), for target month January as lead month 0. The threshold of significant skill lies at 0.31, so yellow cells have insignificant skill. The legend provides the percentage of cells with significant values of R and the domain-averaged value of R. Panel b depicts the percentage of cells with significant skill in terms of R, as a function of the initialization, target and lead month. Each coloured curve corresponds to the hindcasts starting in a single month. For better visualisation the 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 due to chance in the case that the hindcasts have no skill at all (5%).

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions

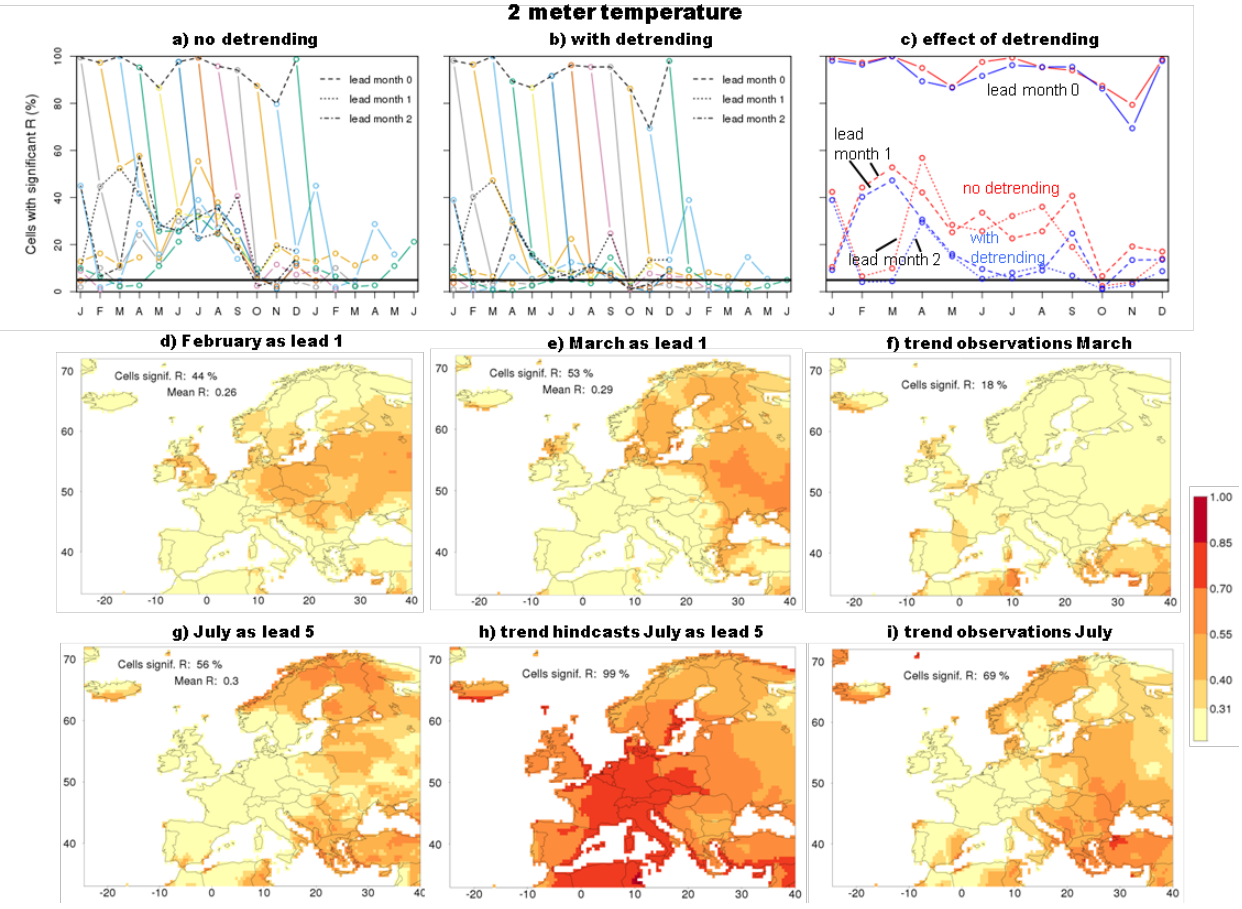

Figure 2: As Fig. 1 but for skill of the two-meter temperature hindcasts. Panels a and b give the percentage of cells with significant values of R for the undetrended (a) and the detrended (b) temperature hindcasts. Panel c compares annual cycles of undetrended with detrended skill for the first three lead months. Three panels show maps of R for the undetrended temperature hindcasts for target months February (d) and March (e) as lead month 1 and July as lead month 5 (g). The remaining three panels depict the correlation coefficient of the trend (not the trend itself!) of the observed monthly mean temperature for March (f) and July (i), and of the trend in the median of the hindcasted temperature for July as lead month 5 (h).

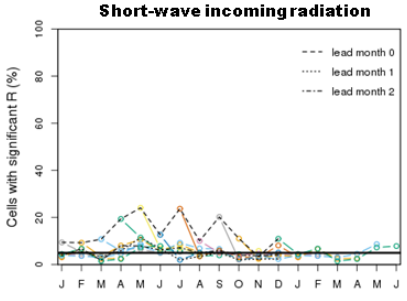

Figure 3: As Fig. 1b, but for skill in the hindcasts of incoming short-wave radiation.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

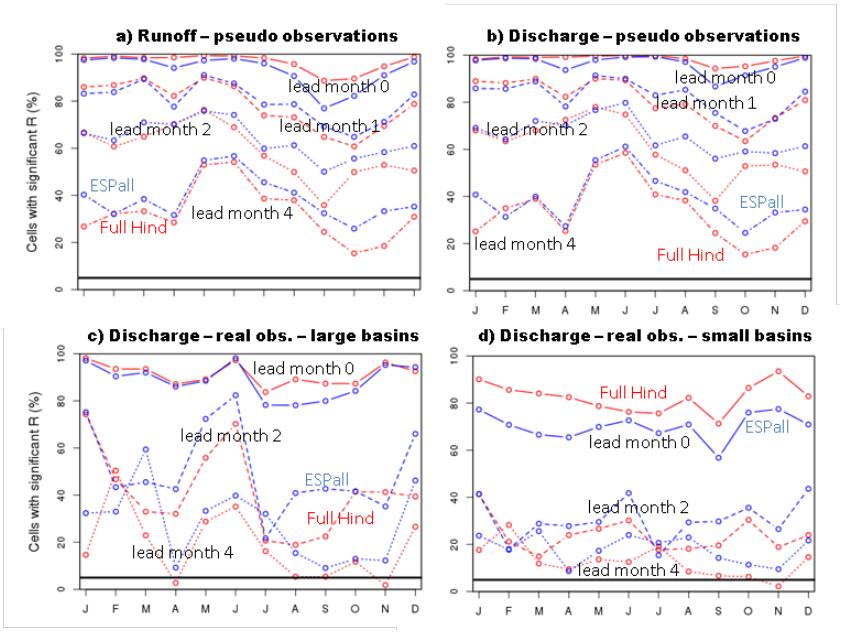

Figure 4: As Fig. 2c but here the annual cycle of skill of the ESPall simulations (blue) is compared with that of the Full Hindcasts (red). The first two panels show theoretical skill obtained with the pseudo-observations for runoff (a) and discharge (b). The other two panels compare actual skill of discharge for large (c) and small (d) basins. Different line types correspond to different lead months.

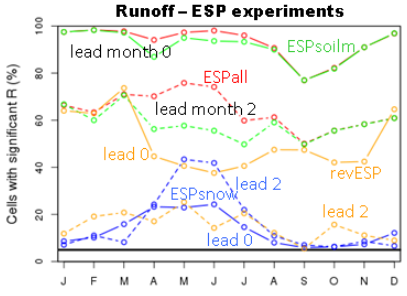

Figure 5: As Fig. 4 but for the annual cycle of the skill in the runoff hindcasts of the four ESP experiments for lead months 0 and 2. Different colours correspond to different experiments and different line types to different lead months.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions

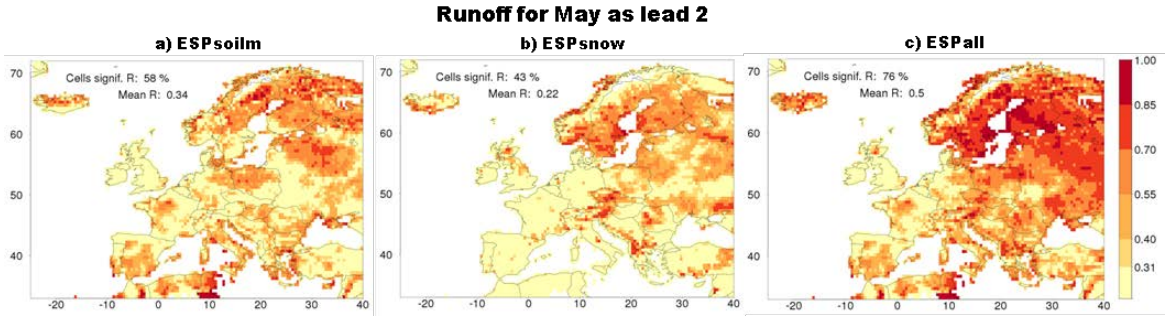

Figure 6: Example that compares the skill in the runoff hindcasts of three ESP experiments, for target month May as lead month 2. For more explanation, see Fig. 1a. White, terrestrial cells correspond to cells where observations or hindcasts consist for more than one third of zeros or one sixth of ties.

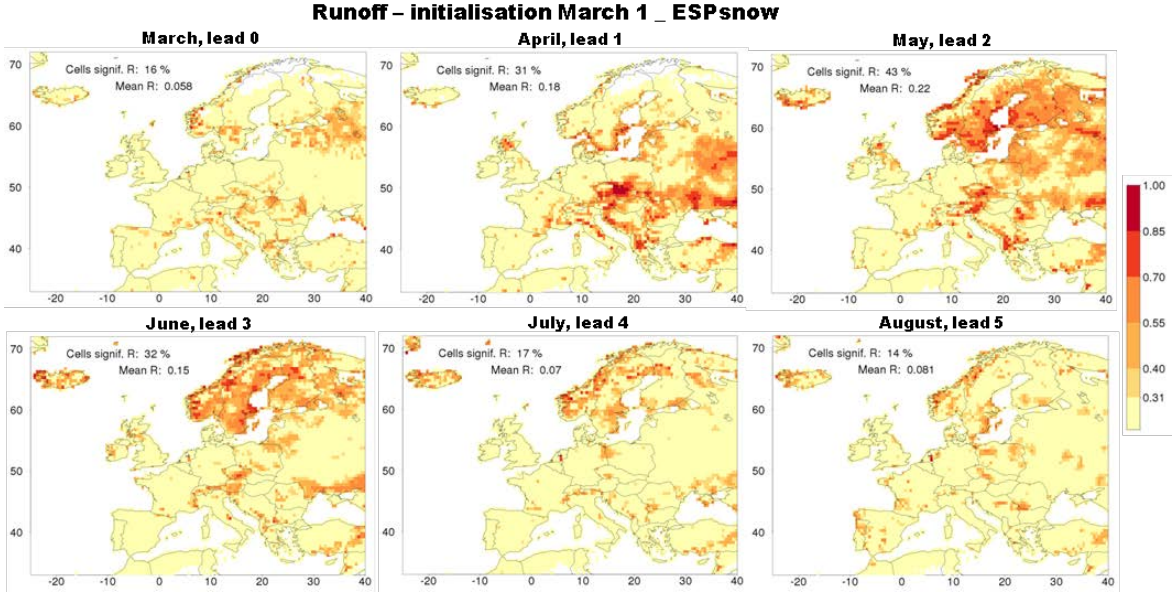

Figure 7: Example showing the variation of skill in runoff as a function of lead time in the ESPsnow experiment, for initialisation on March 1. For more explanation, see Figs. 1a and 6.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions

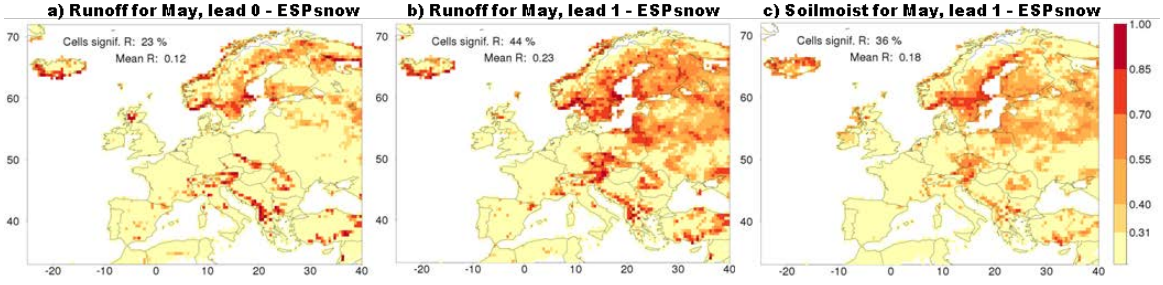

Figure 8: Example illustrating that skill in runoff for a target month may increase with lead time, for May as target month 0 (a) and 1 (b) in the ESPsnow experiment. The explanation resides in the indirect effect of skill in the hindcasts of soil moisture for May, after initialisation on April 1 (c). For more explanation, see Figs. 1a and 6.

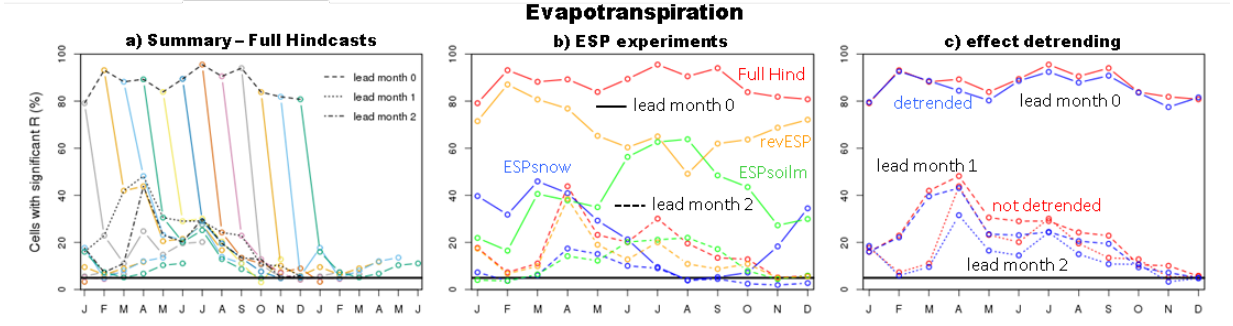

Figure 9: Summary plots of the skill of the hindcasts of evapotranspiration. Panel a summarises the Full Hindcasts (for more explanation, see Fig. 1b), panel b depicts the annual cycles of skill for the Full Hindcasts and three ESP-experiments for lead months 0 and 2, and panel c compares the annual cycles of skill of the undetrended and the detrended Full Hindcasts for the first three lead months.

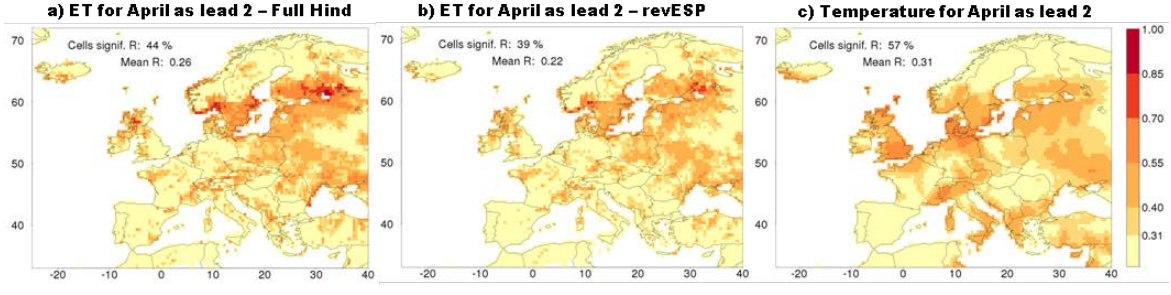

Figure 10: Explanation of the skill in the hindcasts of evapotranspiration for target month April as lead month 2. The panels map the skill of the Full Hindcasts (a), of the reverse ESP experiment (b) and of the hindcasts of temperature (c). For more explanation, see Fig. 1a.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

Hydrology and Earth System Sciences
Discussions

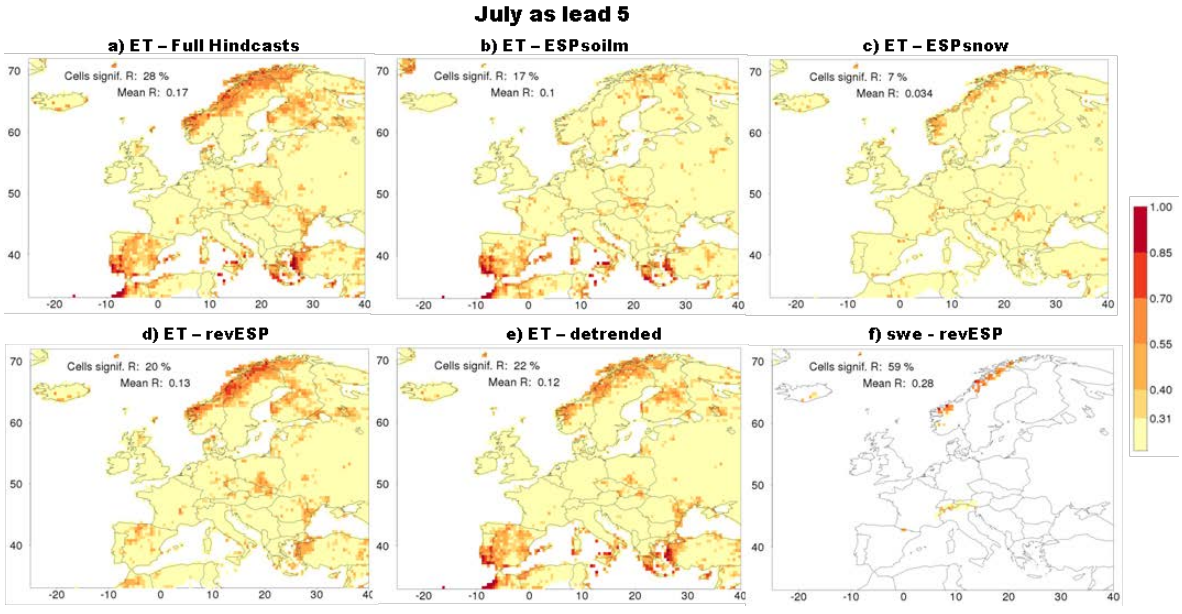

Figure 11: Explanation of the skill in the hindcasts of evapotranspiration for July by taking lead month 5 as an example. The panels map the skill in evapotranspiration of the Full Hindcasts (a), ESPsoilm (b), ESPsnow (c), revESP (d) and the Full Hindcasts after detrending (e). The final is panel (f) depicts skill of the hindcasts of snow water equivalent in the reverse ESP experiment. For more explanation, see Fig. 1a. Note that statistics in the legends of the panels refer only to that part of the domain for which R was computed, which consists of all coloured cells.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 30 November 2016

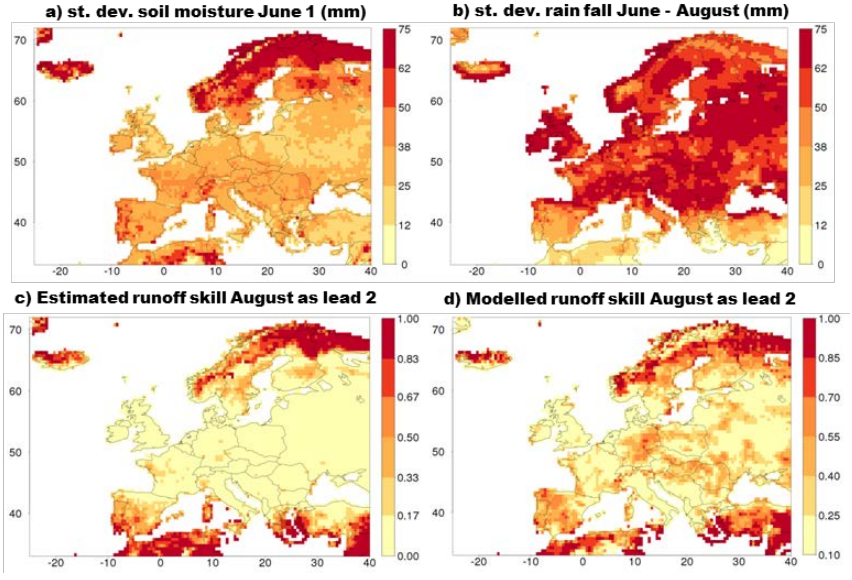

Figure 12: Illustration of a simple method that separates skill in runoff due to initial soil moisture into two components, exemplified for target month August as lead month 2. Skill is caused by variability in initial soil moisture (standard deviation in panel a) and eliminated by variability in rain fall during the course of the simulations (standard deviation in panel b). The two components are combined into an estimate of the skill (Eq. 1) in panel c, which is compared with the skill of the Full Hindcasts (panel d). For more explanation, see Fig.s 1.

Table 1 Sources of skill for hotspot regions and periods of skill. SM is soil moisture.

Region	period	sources of skill
Fennoscandia	Jan – Mar	SM
	Apr - Jul	SM and snow
	Aug – Oct	SM
Poland and North Germany	Oct - Mar	SM
	Apr - May	SM and snow
West France	Dec – May	SM
Romania and Bulgaria	Oct - Mar	SM
	Apr - May	SM and snow
South Mediterranean	Jun – Aug	SM