Monthly streamflow forecasting at varying spatial scales in the Rhine basin

Simon Schick[1,2], Ole Rössler[1,2], and Rolf Weingartner[1,2]

[1]Institute of Geography, University of Bern, Bern, Switzerland
[2]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

Correspondence to: Simon Schick (simon.schick@giub.unibe.ch)

Abstract. Model output statistics (MOS) methods can be used to empirically relate an environmental variable of interest to predictions from earth system models (ESMs). This variable often belongs to a spatial scale not resolved by the ESM. Here, using the linear model fitted by least squares, we regress monthly mean streamflow of the Rhine River at Lobith and Basel against seasonal predictions of precipitation, surface air temperature, and runoff from the European Centre for Medium-Range Weather Forecasts. To address potential effects of a scale mismatch between the ESM's horizontal grid resolution and the hydrological application, the MOS method is further tested with an experiment conducted at the subcatchment scale. This experiment applies the MOS method to 133 additional gauging stations located within the Rhine basin and combines the forecasts from the subcatchments to predict streamflow at Lobith and Basel. In so doing, the MOS method is tested for catchments areas covering four orders of magnitude. Using data from the period 1981-2011, the results show that skill, with respect to climatology, is restricted on average to the first month ahead. This result holds for both the predictor combination that mimics the initial conditions and the predictor combinations that additionally include the dynamical seasonal predictions. The latter, however, reduces the mean absolute error of the former in the range of 5 to 11 percent, which is consistently reproduced at the subcatchment scale. An additional experiment conducted for five day mean streamflow indicates that the dynamical predictions help to reduce uncertainties up to about 20 days ahead, but also reveals some shortcomings of the present MOS method.

1 Introduction

Environmental forecasting at the subseasonal to seasonal time scale promises a basis for planning in e.g. energy production, agriculture, shipping, or water resources management. While the uncertainties of these forecasts are inherently large, they can be reduced when the quantity of interest is controlled by slowly-varying and predictable phenomena. For example, the El Niño-Southern Oscillation plays an important role in predicting the atmosphere, and snow accumulation and melting often forms the backbone in predicting hydrological variables of the land surface (National Academies, 2016).

In case of streamflow forecasting the ESP-revESP experiment proposed by Wood and Lettenmaier (2008) provides a methodological framework to disentangle forecast uncertainty with respect to the initial conditions and the meteorological forcings. Being a retrospective simulation, the experiment consists of model runs where the initial conditions are assumed to be known

and the meteorological forcing series are randomly drawn (ESP, Ensemble Streamflow Prediction) and vice versa (revESP, reverse Ensemble Streamflow Prediction). In this context the initial conditions refer to the spatial distribution, volume, and phase of water in the catchment at the date of prediction.

The framework allows for the estimation of the time range at which the initial conditions control the generation of streamflow: When the prediction error of the ESP simulation exceeds that of the revESP simulation, the meteorological forcings start to dominate the streamflow generation. Similarly, when the prediction error of the ESP simulation approaches the prediction error of the climatology (i.e. average streamflow used as naive prediction strategy), the initial conditions no longer control the streamflow generation.

In both cases this time range depends on the interplay between climatological features (e.g. transitions between wet and dry or cold and warm seasons) and catchment specific hydrological storages (e.g. surface water bodies, soils, aquifers, and snow) and can vary from zero up to several months (van Dijk et al., 2013; Shukla et al., 2013; Yossef et al., 2013). Indeed, this source of predictability is the rationale behind the application of the ESP approach in operational forecast settings, and it can be further exploited by conditioning on climate precursors (e.g. Beckers et al., 2016).

An emerging option for streamflow forecasting is the integration of seasonal predictions from earth system models (ESMs), i.e. coupled atmosphere-ocean-land general circulation models (Yuan et al., 2015b). Predictions from an ESM can be used threefold to the aim of streamflow forecasting by

1. forcing a hydrological model with the predicted evolution of the atmosphere;
2. employing runoff simulated by the land surface model;
3. using the predicted states of the atmosphere, ocean, or land surface in a perfect prognosis or model output statistics context with the streamflow as the predictand.

The first approach requires a calibrated hydrological model for the region of interest. In order to correct a potential bias and to match the spatial and temporal resolution of the hydrological model, it further involves a postprocessing of the atmospheric fields. A postprocessing might also be applied to the streamflow forecasts to account for deficiencies of the hydrological model. See e.g. Yuan et al. (2015a) or Bennett et al. (2016) for recent implementations of such a model chain.

In the second approach the land surface model takes the hydrological model's place with the difference that the atmosphere and land surface are fully coupled. Since the land surface component of ESMs often represents groundwater dynamics and the river routing in a simplified way (Clark et al., 2015), the simulated runoff might be fed to a routing model as e.g. in Pappenberger et al. (2010). To the best of our knowledge, this approach has not yet been tested with a specific focus on subseasonal or seasonal streamflow forecasting.

The third approach deals with developing an empirical prediction rule for streamflow. If the model building procedure is based on observations only, the approach is commonly referred to as perfect prognosis (PP). On the other hand, the model might be built using the hindcast archive of a particular ESM (model output statistics, MOS). In both cases the final prediction rule is applied to the actual ESM outcome to forecast the quantity of interest. Therefore, MOS methods require the presence of a hindcast archive of the involved ESM, but can take systematic errors of the ESM into account (Brunet et al., 1988).

Studies that map ESM output to streamflow with PP or MOS methods include multiple linear regression (Marcos et al., 2017), principal components regression and canonical correlation analysis (Foster and Uvo, 2010; Sahu et al., 2016), generalized linear models (Slater et al., 2017), or artificial neural networks (Humphrey et al., 2016). Whatever the selected predictors, PP and MOS methods generally conduct the mapping across spatial scales. For example, if the catchment of interest falls below the grid scale of the ESM, PP and MOS methods implicitly perform a downscaling step. If the catchment covers several grid points, the method implicitly performs an upscaling.

The present study aims to take up this scale bridging and to test a MOS-based approach for monthly mean streamflow forecasting and a range of catchment areas. To analyse the limits of predictability and to aid interpretation, we first define predictor combinations motivated by the ESP-revESP framework. Next, seasonal predictions of precipitation, surface air temperature, and runoff from the European Centre for Medium-Range Weather Forecasts (ECMWF) enter the regression equation and the resulting forecast skill is estimated with respect to the ESP-inspired regression model.

The variation of the catchment area borrows from the concept of the 'working scale' (Blöschl and Sivapalan, 1995): Given a particular target catchment, the regression models are applied at the catchment scale as well as at two levels of subcatchment scales. In case of the latter, the resulting forecasts are combined in order to get a forecast at the outlet of the target catchment. By validating the combined forecasts of the subcatchments at the main outlet, any differences in the forecast quality can be attributed to the working scales.

This experiment is conducted for the Rhine River at Lobith and Basel in Western Europe. Studies using subseasonal or seasonal climate predictions indicate for several parts of the Rhine basin moderate skill beyond the lead time of traditional weather forecasts. These studies apply the model chain as outlined above in approach number one: Concerning catchments of the Alpine and High Rhine, Orth and Seneviratne (2013) estimate the skillful lead time for daily mean streamflow to lie between one and two weeks, which increases to about one month when focusing on low flows (Fundel et al., 2013; Jörg-Hess et al., 2015). Also for daily low flow Demirel et al. (2015) report for the Moselle River a sharp decrease in skill after 30 days. For a set of French catchments Crochemore et al. (2016) show that weekly streamflow forecasts are improved for lead times up to about one month when using postprocessed precipitation predictions. Singla et al. (2012) advance spring mean streamflow forecasts for the French part of the Rhine basin with seasonal predictions of precipitation and surface air temperature.

As a compromise between skillful lead time and temporal resolution, we decide to focus on monthly mean streamflow at lead times of zero, one, and two months. In order to resolve the monthly time scale and to test the MOS method at shorter time intervals, an experiment is further conducted for five day mean streamflow. Here, zero lead time refers to forecasting one time interval ahead, while e.g. a one month lead time denotes a temporal gap of one month between the release of a forecast and its time of validity.

Strictly speaking, the present study deals with hindcasts or retrospective forecasts. However, for the sake of readability we use the terms forecast, hindcast, and prediction interchangeably. Below, Sect. 2 introduces the study region, Sect. 3 describes the data set, Sect. 4 exposes the methodology in more detail, and in Sect. 5 and 6 the results are presented and discussed, respectively.

Table 1. Geography of the Rhine River at Basel and Lobith according to CORINE (2013), EU-DEM (2013), and GRDC (2016).

	Lobith	Basel
area (km^2)	159700	36000
gauging station ($\mathrm{m\,a.\,s.}$)	20	250
elevation min ($\mathrm{m\,a.\,s.}$)	-230	250
elevation max ($\mathrm{m\,a.\,s.}$)	4060	4060
elevation mean ($\mathrm{m\,a.\,s.}$)	490	1050
farmed area (%)	47.7	36.8
forest (%)	35.8	31.6
grass land (%)	3.4	11.4
urban area (%)	9.6	7.0
wasteland (%)	1.8	8.2
surface water (%)	1.4	4.0
glacier (%)	0.3	1.0

2 Study region

The Rhine River is situated in Western Europe and discharges into the North Sea; in the south its basin is defined by the Alps. About 58 million people use the Rhine water for the purpose of navigation, hydro power, industry, agriculture, drinking water supply, and leisure (ICPR, 2009). The present study focuses on two gauging stations: The first is located in Lobith near the Dutch-German border, the second in Basel in the tri-border region of France, Germany, and Switzerland.

Table 1 lists some geographical attributes. The Rhine at Basel covers an area of approximately one fifth of the Rhine at Lobith whereas the mean elevation halves when going from Basel to Lobith. The negative minimum elevation of the Rhine at Lobith is due to a coal mine. Dominant land use classes are farmed areas and forests, but the Rhine at Basel proportionately includes more grass land, wasteland, surface water, and glacier.

Concerning the climatology of the period 1981-2011 (Fig. 1), we observe that streamflow peaks at Lobith in winter and at Basel in early summer. Streamflow at Basel is dominated by snow accumulation in winter, subsequent snow melting in spring, and high precipitation in summer. At Lobith precipitation exhibits less variability and higher surface air temperature intensifies evaporation. Based on recent climate projections, it is expected that streamflow in the Rhine basin increases in winter, decreases in summer, and slightly decreases in its annual mean in the last third of the 21th century (Bosshard et al., 2014).

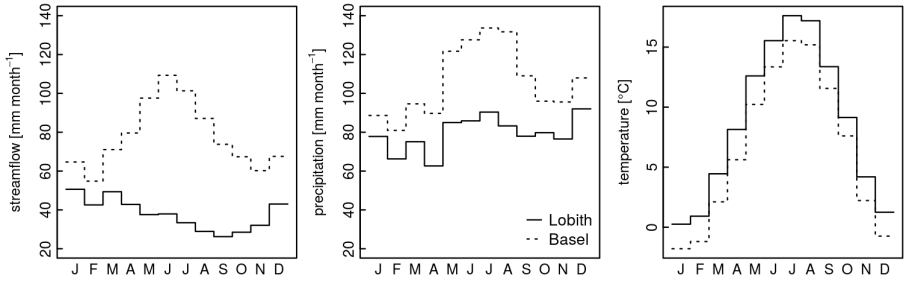

Figure 1. Monthly area averages of streamflow, precipitation, and surface air temperature for the Rhine at Lobith and Basel with respect to the period 1981-2011 (GRDC, 2016; E-OBS, 2016).

3 Data

Observations of river streamflow and gridded precipitation, surface air temperature, and runoff of the period 1981-2011 in daily resolution constitute the data set. Throughout the study gridded quantities get aggregated to (sub)catchment area averages.

3.1 Observations

The streamflow observations consist of a set of 135 time series in $\mathrm{m^3\,s^{-1}}$. These series as well as the corresponding catchment boundaries are provided by several public authorities and the Global Runoff Data Centre (GRDC (2016); see also Sect. 9), and belong to catchments with nearly natural to heavily regulated streamflow.

The ENSEMBLES gridded observational data set in Europe (E-OBS, version 14.0) provides precipitation and surface air temperature on a 0.25° regular grid (Haylock et al., 2008; E-OBS, 2016). These fields base upon the interpolation of station data and are subject to inhomogeneities and biases. However, a comparison against meteorological fields derived from denser station networks attests a high correlation (Hofstra et al., 2009). In case of the Rhine basin an E-OBS tile approximately covers an area of 500 $\mathrm{km^2}$.

3.2 Dynamical seasonal predictions

Precipitation, surface air temperature, and runoff from ECMWF's seasonal forecast system 4 (S4) archive are on a 0.75° regular grid, amounting in case of the Rhine basin to a tile area of about 4500 $\mathrm{km^2}$. The hindcast set consists of 15 members of which we take the ensemble mean. Runs of the coupled atmosphere-ocean-land model are initialised on the first day of each month and simulate the subsequent seven months. Up to 2010, initial conditions are out of ERA Interim, and the year 2011 is based on the operational analysis.

The atmospheric model (IFS cycle 36r4) consists of 91 vertical levels with the top level at 0.01 hPa in the mesosphere. The horizontal resolution is truncated at TL255 and the temporal discretisation equals 45 min. The NEMO ocean model has 42 levels with a horizontal resolution of about 1°. Sea ice is considered by using its actual extent from the analysis and relaxing it towards the climatology of the past five years (Molteni et al., 2011).

The H-TESSEL land surface model implements four soil layers with an additional snow layer on the top. Interception, infiltration, surface runoff, and evapotranspiration are dealt with by dynamically separating a grid cell into fractions of bare ground, low and high vegetation, intercepted water, and shaded and exposed snow. In contrast, the soil properties of a particular layer are uniformly distributed within one grid cell. Vertical water movement in the soil follows Richards's equation with an additional sink term to allow for water uptake by plants. Runoff per grid cell equals the sum of surface runoff and open drainage at the soil bottom (Balsamo et al., 2009; ECMWF, 2016).

4 Method

The following subsections outline the experiment, which is individually conducted for both the Rhine at Lobith and Basel. Section 4.1 details the predictor combinations and the regression strategy, Sect. 4.2 introduces the variation of the catchment area, and Sect. 4.3 illustrates the validation of the resulting hindcasts.

4.1 Model building

The predictand $y_{i,j}$ denotes observations of mean streamflow at a specific gauging site in $\mathrm{m^3\,s^{-1}}$ for $j = 5, 10, \ldots, 180$ d, starting the first day of each calendar month $i = 1, \ldots, 12$ in the period 1981-2011.

4.1.1 Predictor combinations

The set of predictors consists of variables that either precede or succeed the date of prediction i (Tab. 2). The first model refRun (reference run) is aimed to estimate how well the regression works given the best available input data. The combinations named preMet (preceding meteorology) and subMet (subsequent meteorology) are constrained to precipitation and surface air temperature preceding and subsequent to the date of forecast, respectively.

The S4* combinations constitute the MOS method and consider the seasonal predictions out of the S4 hindcast archive, where we use the asterisk as wildcard to refer to any of the S4P, S4T, S4PT, and S4Q models. The S4P and S4T models are used to separate the forecast quality with respect to precipitation and temperature. The S4Q model is tested as H-TESSEL does not implement groundwater dynamics and preceding precipitation and temperature might tap this source of predictability.

4.1.2 Regression

For a particular $y_{i,j}$ we first apply a correlation screening to select the optimal aggregation time $a_{i,j}$ for each predictor according to

$$a_{i,j} = argmax_k \mid cor(y_{i,j}, x_{i,k}) \mid \tag{1}$$

where $x_{i,k}$ is one of the predictors from Tab. 2 and $k = -10, -20, \ldots, -720$ d in case of $p^{\text{pre}}$and t^{pre} (backward in time relative to the date of prediction) and $k = 5, 10, \ldots, j$ d in case of p^{sub}, t^{sub}, and q^{sub} (forward in time relative to the date of prediction). The limit of 720 d is chosen since larger values rarely get selected.

Table 2. Predictor combinations consisting of (with respect to the date of prediction) preceding and subsequent precipitation (p), surface air temperature (t), and runoff (q); the numerical values are either out of the E-OBS gridded data set or ECMWF's S4 hindcast archive.

	preceding		subsequent		
model	p^{pre}	t^{pre}	p^{sub}	t^{sub}	q^{sub}
refRun	E-OBS	E-OBS	E-OBS	E-OBS	-
preMet	E-OBS	E-OBS	-	-	-
subMet	-	-	E-OBS	E-OBS	-
S4P	E-OBS	E-OBS	S4	-	-
S4T	E-OBS	E-OBS	-	S4	-
S4PT	E-OBS	E-OBS	S4	S4	-
S4Q	E-OBS	E-OBS	-	-	S4

The ordinary least squares hyperplane is then used for prediction without any transformation, basis expansion, or interaction. However, model variance can be an issue: Specifically for the preMet model from Tab. 2 we expect the signal-to-noise ratio to be low for most of the predictands. In combination with the moderate sample size $n = 26$ for model fitting (with respect to the cross-validation, see Sect. 4.1.3), perturbations in the training set can lead to large changes in the predictor's time lengths $a_{i,j}$ and regression coefficients. In order to stabilise model variance, we draw 100 non-parametric bootstrap replicates of the training set, fit the model to these replicates, and combine the predictions by unweighted averaging (Breiman, 1996; Schick et al., 2016).

4.1.3 Cross-validation

Each year with a buffer of two years (i.e. the two preceding and subsequent years) is left out and the regression outlined in Sect. 4.1.2 is applied to the remaining years. The fitted models then predict the central left-out years. Buffering is used to avoid artificial forecast quality due to hydrometeorological persistence (Michaelsen, 1987).

4.1.4 Lead time

Lead time is introduced by integrating the predicted $\hat{y}_{i,j}$ in time and taking differences with respect to j. For example monthly mean streamflow z_i in July ($i = 7$) is predicted with a lead time of one month according to

$$\hat{z}_7 = (\hat{y}_{6,60} \cdot (30 + 31) \cdot b - \hat{y}_{6,30} \cdot 30 \cdot b)/(31 \cdot b) \tag{2}$$

where $b = 24 \cdot 60 \cdot 60$ s equals the number of seconds of one day and both $\hat{y}$ and $\hat{z}$ have unit $\mathrm{m}^3\,\mathrm{s}^{-1}$. For zero lead time, we set $\hat{z}_i = \hat{y}_{i,30}$. Please note that the year 1981 needs to be dropped from the validation (Sect. 4.3) since the length of the streamflow series prevents to forecast e.g. January 1981 with a lead time of one month.

Table 3. Subcatchment division of the Rhine at Lobith and Basel. The median area covers four orders of magnitude.

	number of	area km^2		
	subcatchments	min	median	max
Lobith level 1	1	-	159700	-
Lobith level 2	5	19690	33220	43550
Lobith level 3	12	8284	13040	17610
Basel level 1	1	-	36000	-
Basel level 2	10	1871	2946	6346
Basel level 3	124	6	187	2654

4.2 Spatial levels

Contrasting the forecast quality of a given model for catchments separated in space inevitably implies a large number of factors, e.g. the geographic location (and thus the involved ESM grid points), the orography, or the degree to which streamflow is regulated. In order to hold these factors while screening through a range of catchment areas, we propose to vary the working scale within a particular target catchment.

Following this line of argumentation we apply the model building procedure from Sect. 4.1 to three distinct sets of subcatchments, which we term 'spatial levels' (Tab. 3). Spatial level 1 simply consists of the target catchment itself, i.e. the Rhine at Lobith and Basel. At spatial levels 2 and 3 we take additional gauging stations from within the Rhine basin, which naturally divide the basin into subcatchments.

For these subcatchments we have streamflow observations belonging to the entire upstream area, but not the actual subcatchment area itself. To arrive at an estimate of the water volume generated by the subcatchment, we equate the predictand $y_{i,j}$ to the difference of outflow and inflow of that subcatchment. For a particular date of prediction and spatial level, the sum of the resulting subcatchment forecasts $\hat{z}_i$ then constitutes the final forecast for the Rhine at Lobith and Basel, respectively.

This procedure implies that we ignore the water travel time: First when taking the differences of outflows and inflows and second when summing up the subcatchment forecasts. While the former increases the observational noise, the latter does not affect the regression itself, but adds a noise term to the final forecast at Lobith and Basel. As the statistical properties of the noise introduced by the water travel time is unknown, we only can argue that the results provide a lower bound of the forecast quality due to this methodological constraint.

4.3 Validation

The forecast quality of the regression models is analysed with the pairs of cross-validated monthly mean streamflow forecasts and observations $(\hat{z}, z)$. These series cover the period 1982-2011 and have a sample size of $n = 360$. In general the validation is based on the mean absolute error (MAE) and Pearson's correlation coefficient (ρ).

The first validation steps focus on the forecasts at Lobith and Basel and thus consider the sum of the subcatchment forecasts $\hat{z}$ per spatial level. The forecasts in the subcatchments itself are addressed in Sect. 4.3.5. Finally, the validation of the five day mean streamflow forecasts (Sect. 4.3.6) complements the monthly analysis.

4.3.1 Benchmarks

Climatology and runoff simulated by H-TESSEL serve as benchmarks. The climatology is estimated with the arithmetic mean from the daily streamflow observations. After averaging in time, runoff from H-TESSEL gets post-calibrated via linear regression against the streamflow observations per spatial level. For both benchmarks the cross-validation scheme from Sect. 4.1.3 is applied.

4.3.2 Taylor diagram

Taylor diagrams (Taylor, 2001) provide an instrument to contrast model performances. The plotting position of a particular model has a distance from the origin equal to the standard deviation of its forecasts $\hat{z}$ and is located on the line having an angle of incline $\phi = arccos(\rho)$. The plotting position of the observations z has a distance from the origin equal to the standard deviation of z and is located on the abscissa. The distance between these two plotting positions equals the root mean squared error with the unconditional bias $E(\hat{Z} - Z)$ removed.

4.3.3 Statistical significance

In case of the monthly analysis it turns out that the paired differences of absolute errors for a given lead time, spatial level, and reference model r

$$d = \mid \hat{z}^{\mathrm{r}} - z \mid - \mid \hat{z}^{\mathrm{S4*}} - z \mid \tag{3}$$

no longer exhibit serial correlation and approximately follow a Gaussian distribution. Using the mean difference $\bar{d}$, we then report the p-values of the two-sided t-test with null hypothesis $\bar{d} = 0$ and alternative hypothesis $\bar{d} \neq 0$. The sample autocorrelation functions and quantile plots against the Gaussian distribution of d for zero lead time and r being the preMet model are included in the additional materials (Sect. 10).

4.3.4 Skill

To evaluate whether a particular model m has skill with respect to a reference model r the MAE ratio

$$s = 1 - \frac{\mathrm{MAE}_m}{\mathrm{MAE}_r} \tag{4}$$

is employed. For example, m could be a S4* model and r the preMet model. $s = 0.1$ means that the model m lowers the MAE of model r by 10 $\%$.

4.3.5 Subcatchments

To help in the interpretation of the forecast quality of the MOS method regarding the spatial levels at Lobith and Basel, we plot in a qualitative manner the MAE skill score (Eq. 4) of the S4* and preMet models in space as well as against the subcatchment area, the median of the terrain roughness, the MAE skill score of the subMet with the preMet model as reference, and the MAE skill score of the refRun model with the climatology as reference.

The terrain roughness is included since the atmospheric flow in complex terrain is challenging to simulate and atmospheric general circulation models need to filter the topography according to their spatial resolution (Maraun and Widmann, 2015; Torma et al., 2015). The terrain roughness is defined as the difference of the maximum and minimum elevation value within a 3 times 3 pixel window (Wilson et al., 2007). It is derived here from the digital elevation model EU-DEM (2013), which has a horizontal resolution of 25 m.

4.3.6 Five day mean streamflow

In order to predict five day mean streamflow, Eq. 2 is used with a step size of five days. However, the monthly date of predictions impose some restrictions to the validation: First, it is not possible to derive regular time series at different lead times as in the monthly analysis. Furthermore, the distributional assumptions required for the statistical test from Sect. 4.3.3 are not valid. The results of the five day mean streamflow experiment thus are restricted to a qualitative interpretation.

5 Results

The experiment spans several dimensions (i.e. Lobith versus Basel, date of prediction, lead times, predictor combinations, spatial levels), so we frequently need to collapse one or several dimensions. The additional materials as listed in Sect. 10 try to complete the results as presented bellow.

5.1 Taylor diagram

Figure 2 shows the Taylor diagrams for Lobith and Basel to get a global overview regarding the lead times, predictor combinations, and spatial levels. Accurate forecasts reproduce the standard deviation of the observations (thus lie on the circle with radius equal to the standard deviation of the observations), and also exhibit high correlation (so travel on this circle towards the observations on the abscissa). At a first glimpse the spatial levels do not introduce clear differences and most of the models mass at the same spots.

The benchmark climatology is outperformed at zero lead time by all models. At longer lead times the subMet model pops up besides the refRun model and the remaining models approach climatology. For the refRun model we note a correlation of about 0.9 independently of the lead time while the observation's variability generally is underestimated.

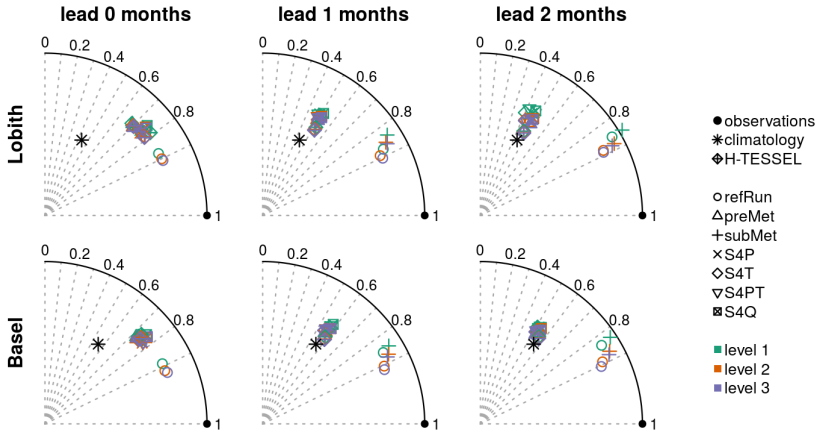

Figure 2. Taylor diagrams for the benchmarks climatology and H-TESSEL and the predictor combinations from Tab. 2 at Lobith (top row) and Basel (bottom row); $n = 360$.

For Lobith and zero lead time we observe an elongated cluster, which comprises all models but the climatology and the refRun model. Some models score a higher correlation – zooming in would reveal that these are the S4P, S4PT, and S4Q models with H-TESSEL standing at the forefront.

5.2 Date of prediction versus lead time

Figure 3 takes a closer look at the clusters in Fig. 2 at hand of the S4PT model and in addition breaks down the prediction skill into the different calendar months. Please note that the ordinate lists the calendar month and not the date of prediction – e.g. the top rows show the skill in predicting January's mean streamflow for lead times of zero up to two months. Crosses indicate p-values smaller than 0.05 when Eq. 3 is applied to the individual calendar months.

In general, the patterns repeat more or less along the spatial levels and the S4PT model beats the reference models in the denominator of Eq. 4 only at zero lead time. An exception can be observed for June, for which the S4PT model most likely outperforms the climatology at one month lead time.

For May, the S4PT model outscores both the preMet and the subMet model. While significant differences between the S4PT and the preMet models are rare, the subMet model starts to outperform the S4PT model already at a lead time of one month. The comparison against the bias corrected H-TESSEL runoff shows that the S4PT model might provide more accurate predictions for late spring and early summer, but not otherwise.

5.3 Mean absolute error

In order to conclude the analysis of the monthly predictions at Lobith and Basel, Tab. 4 reports the MAE at zero lead time. Reading Tab. 4 along the rows reveals a more or less consistent pattern: The refRun model approximately halves the MAE of the climatology; differences between the preMet, subMet, and S4T models are small; compared to the preMet model, the S4P,

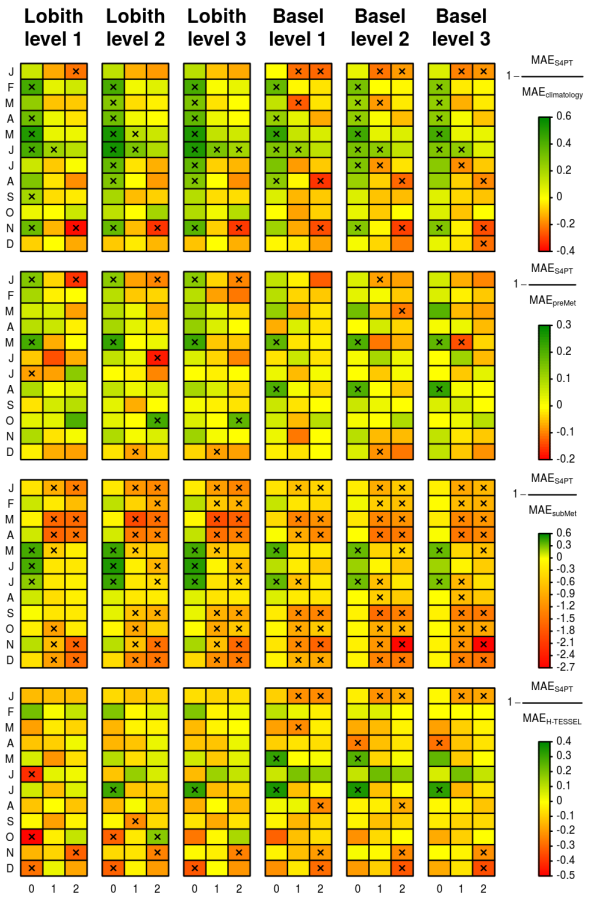

Figure 3. MAE skill score of the S4PT model with respect to the climatology, the preMet and subMet models, and bias corrected H-TESSEL runoff. The ordinate depicts the calendar month and the abscissa the monthly lead time. Crosses indicate p-values smaller than 0.05 for the null hypothesis 'the reference model in the denominator and the S4PT model score an equal mean absolute error'; $n = 30$.

S4PT, and S4Q models lower the MAE by about 40 $\mathrm{m^3\,s^{-1}}$ for Lobith and by about 15 $\mathrm{m^3\,s^{-1}}$ for Basel; and H-TESSEL outperforms the S4* models in case of Lobith, but not Basel. When reading Tab. 4 along the columns, we generally note at Lobith a decreasing MAE when going from spatial level 1 to spatial level 3. In case of Basel, the MAE remains more or less constant except for the refRun model.

Focusing on the MOS method, Tab. 5 lists the corresponding MAE skill score (Eq. 4) of the S4* models using the preMet model as the reference. The p-values for the null hypothesis 'the preMet and S4* models score an equal mean absolute error' are listed in brackets. We see that the S4P, S4PT, and S4Q models score an error reduction ranging from 5 to 11 %. In case of the S4T model an error reduction is either not existent (Lobith) or small (Basel), which comes along with high p-values.

Table 4. Mean absolute error at zero lead time of the benchmarks climatology and H-TESSEL and the predictor combinations from Tab. 2, rounded to integers. All values have unit $\mathrm{m}^3\,\mathrm{s}^{-1}$; $n = 360$.

	climatology	H-TESSEL	refRun	preMet	subMet	S4P	S4T	S4PT	S4Q
Lobith level 1	633	419	334	500	498	459	503	467	445
Lobith level 2	633	417	295	484	494	440	484	445	442
Lobith level 3	633	417	287	482	495	436	481	441	439
Basel level 1	239	191	131	199	195	189	196	188	189
Basel level 2	239	186	117	199	192	184	194	184	186
Basel level 3	239	184	112	199	193	184	195	183	187

Table 5. MAE skill score of the S4* models relative to the preMet model (Eq. 4, expressed in percent) at zero month lead time. p-values for the null hypothesis 'the preMet and S4* models score an equal mean absolute error' are enclosed in brackets; $n = 360$.

	S4P		S4T		S4PT		S4Q	
Lobith level 1	8	(<0.01)	-1	(0.64)	7	(<0.01)	11	(<0.01)
Lobith level 2	9	(<0.01)	0	(0.93)	8	(<0.01)	9	(<0.01)
Lobith level 3	10	(<0.01)	0	(0.88)	9	(<0.01)	9	(<0.01)
Basel level 1	5	(0.01)	1	(0.25)	6	(0.01)	5	(0.01)
Basel level 2	7	(<0.01)	2	(0.04)	8	(<0.01)	6	(<0.01)
Basel level 3	7	(<0.01)	2	(0.11)	8	(<0.01)	6	(<0.01)

5.4 Subcatchments

Figure 4 depicts the MAE skill score (Eq. 4) for the S4PT model relative to the preMet model for each subcatchment at zero lead time. If the MAE difference does not exhibit a p-value smaller than 0.05 (Eq. 3), the subcatchment is coloured in white. We observe that the MAE skill score takes values in the range of about -0.06 to 0.11 and both the lowest and highest scores occur at Basel and spatial level 3. Negative scores can only be found at Basel and spatial level 3, and positive skill tends to cluster in space.

The same skill scores from Fig. 4 are contrasted in Fig. 5 with the subcatchment area, the median of the terrain roughness, the MAE skill score of the subMet model relative to the preMet model, and the MAE skill score of the refRun model relative to the climatology. If the MAE difference of the S4PT and the preMet models does not exhibit a p-value smaller than 0.05, the symbol is drawn with a reduced size. The horizontal lines depict the MAE skill scores from Tab. 5.

While the first two attributes concern the geography of the subcatchment, the third attribute indicates the relevance of the initial conditions for the subsequent generation of streamflow. The fourth attribute shows how well the S4PT model performs relative to the climatology as benchmark, when it has access to the best available input data.

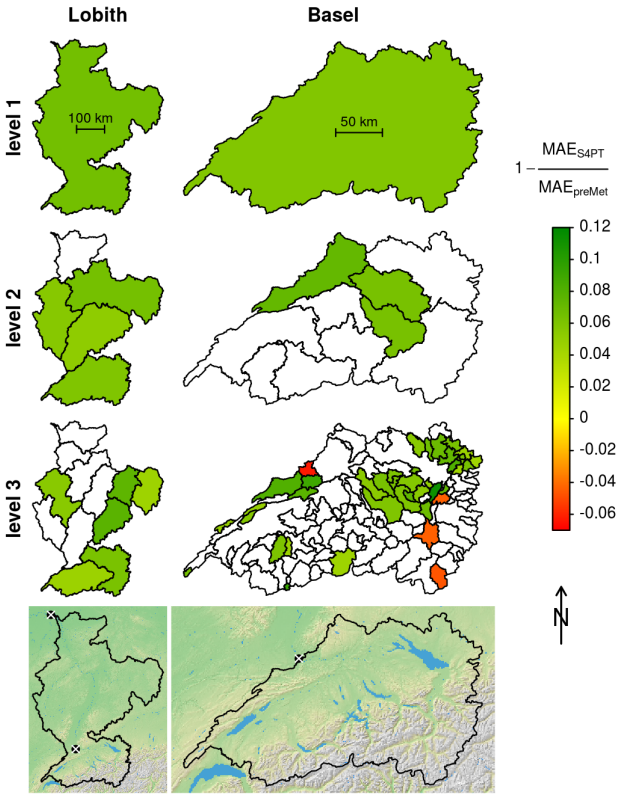

Figure 4. MAE skill score of the S4PT model with respect to the preMet model for each subcatchment and zero lead time. Subcatchments are coloured only when the p-value for the null hypothesis 'the preMet and S4PT models score an equal mean absolute error' is smaller than 0.05. In the bottom maps the main outlets at Lobith and Basel are marked with a white cross and open water surfaces are coloured in blue (CORINE, 2013; EU-DEM, 2013); $n = 360$.

The resulting patterns suggest that positive skill does not depend on the subcatchment area. On the other hand, a low terrain roughness and a weak relevance of the initial conditions seem to favour positive skill. The last row finally indicates that positive skill is restricted to subcatchments where the refRun model outperforms climatology. Roughly, a hypothetical relationship appears to strengthen from the top to the bottom plots.

5.5 Five day mean streamflow

Figure 6 shows the correlation coefficient of the five day mean streamflow observations and corresponding predictions for all models and benchmarks up to a lead time of 45 days. We observe that the refRun model scores a correlation of about 0.8 with a slowly decreasing tendency towards longer lead times. Furthermore, the subMet model crosses the preMet model approximately in the second week; the preMet model approaches climatology within about three weeks; and the subMet model comes close to the refRun model in about three weeks.

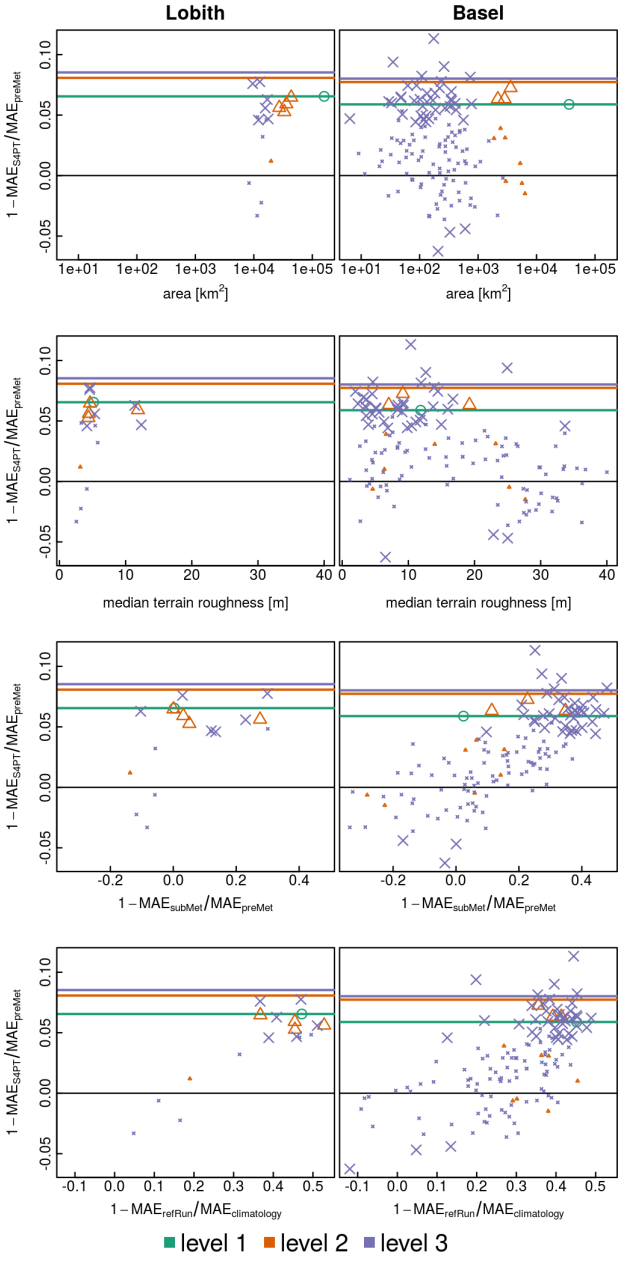

Figure 5. MAE skill score of the S4PT model with respect to the preMet model for each subcatchment and zero lead time, plotted against subcatchment attributes (see Sect. 4.3.5 for details). Large symbols note a p-value smaller than 0.05 for the null hypothesis 'the preMet and S4PT models score an equal mean absolute error'. The horizontal lines indicate the corresponding skill per spatial level at Lobith and Basel; $n = 360$.

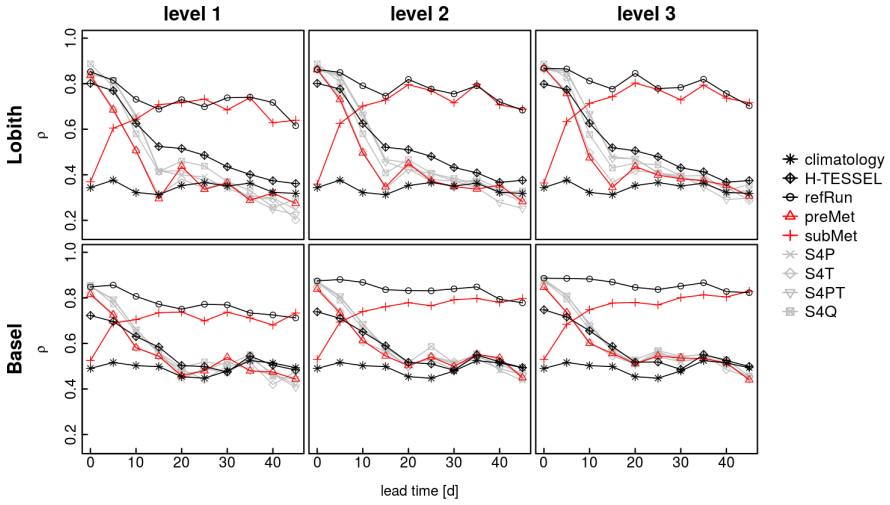

Figure 6. Correlation coefficient of five day mean streamflow observations and predictions for lead times up to 45 days; $n = 360$.

In addition, we see that the bias corrected H-TESSEL runoff starts rather cautious, but seems to slightly outperform the S4* models at longer lead times. While the S4T model is hardly distinguishable from the preMet model, the S4P, S4PT, and S4Q models appear to outperform the preMet model within the first 20 days (Lobith) and 15 days (Basel).

For the full range of lead times, the spatial levels introduce some clear differences (Fig. 7): The refRun and subMet models get improved at longer lead times along the spatial levels. For lead times longer than about 50 days, the bias corrected H-TESSEL runoff stays in close harmony to the climatology, while the S4* and preMet models instead start to score a smaller correlation. This effect seems to be mitigated at spatial levels 2 and 3.

6 Discussion

6.1 Model building

In case of the monthly streamflow, the refRun model ends up with a correlation of about 0.9 for all lead times, spatial levels, and both Lobith and Basel (Fig. 2). Part of this correlation is also the annual cycle (Fig. 1), which already leads to a correlation of about 0.5 when using the climatology as prediction rule. The forecasts from the refRun model do not fully reproduce the observations' variance, what might be improved with a transformation of the predictand (Wang et al., 2012). This option – along with predictors that more explicitly represent the initial conditions, e.g. lake levels, soil moisture content, or snow courses – preferably should be tested in a future study with a small number of catchments and longer time series.

For the five day mean streamflow the refRun model gets degraded. At short lead times the correlation amounts to about 0.8, while for longer lead times the correlation exhibits a decreasing trend. Either the present model formulation is less valid (especially for small values of j, say 5 or 10 days, the assumption of linearity might fail) or the scheme to introduce the lead

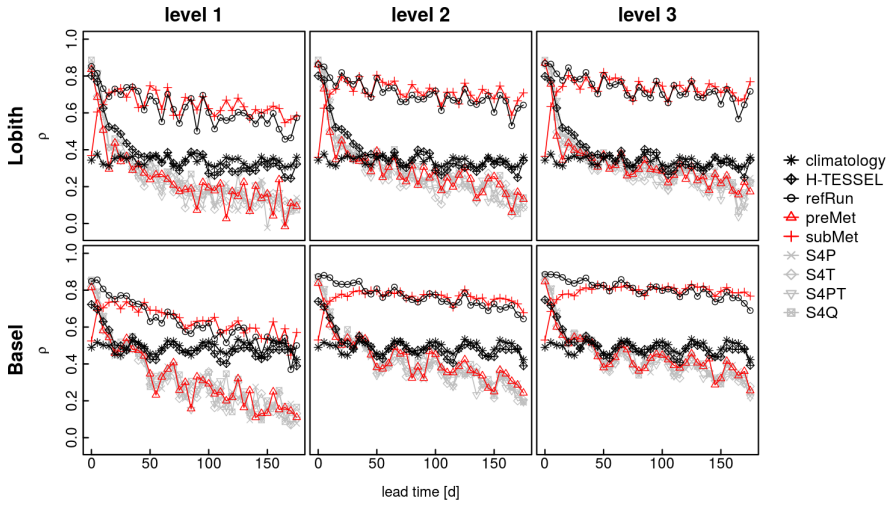

Figure 7. Correlation coefficient of five day mean streamflow observations and predictions for lead times up to 175 days; $n = 360$.

time (Eq. 2) is not appropriate for mean values of small time windows (e.g. the subtraction of streamflow volumes of 155 and 150 days only allows for small prediction errors). Since the final forecast values are not part of the regression equation, it is even possible to perform worse than climatology (Fig. 7).

6.2 Spatial levels

The spatial levels can affect the forecast quality either

- via the ignorance of the water travel time (Sect. 4.2)
- or the aggregation of the E-OBS and S4 fields at the catchment scale is not the appropriate spatial resolution (e.g. large scale grid averages cancel any spatial variability, and for catchment areas below the grid scale a grid point does not necessarily contain information valid at the local scale).

However, clear differences between the spatial levels can only be observed for the five day streamflow predictions, where at spatial levels 2 and 3 the forecast quality gets improved. Using local information of precipitation, surface air temperature, or runoff appears to compensate for the ignorance of the water travel time.

6.3 preMet-subMet

In Yossef et al. (2013) the ESP-revESP framework is applied to the world's largest river basins using the global hydrological model PCRaster Global Water Balance (PCR-GLOBWB). Considering all calendar months and the Rhine at Lobith, the ESP simulation outperforms the climatology only at zero lead time; the revESP simulation is outperformed at zero lead time by

both the ESP simulation and climatology; and at longer lead times the revESP simulation clearly outperforms both the ESP simulation and climatology. Therefore, the results of Yossef et al. (2013) and those of the present study are mostly in line.

The analysis of the five day mean streamflow forecasts (Sect. 5.5) further reveals that the crossover of the preMet and subMet models occurs approximately in the second week. However, this estimate ignores variations within the calendar year and should be considered as a rough guess, since the regression method is far from being perfect in case of the five day mean streamflow.

6.4 MOS method

In case of the monthly mean streamflow forecasts at zero lead time, the MOS method based on precipitation or runoff provides a smaller mean absolute error than the preMet model (Tab. 5). Figure 6 suggests that this error reduction at the monthly time scale arises from the predictions of the first 15 to 20 days. Here, it must be stressed that for the present regression strategy temperature subsequent to the date of prediction often is a weak predictor (regression coefficients of the refRun model at spatial level 1 are included in the additional materials, see Sect. 10). Thus, a rejection of the S4T model does not allow any inference about the forecast quality of surface air temperature itself.

Figure 5 indicates that the subcatchment area most likely is not relevant to score positive skill. Rather the S4PT model outperforms the preMet model in subcatchments where the terrain roughness and the relevance of the initial conditions are low. However, the terrain roughness and the relevance of the initial conditions are not independent attributes: Fig. 4 shows that for small subcatchments in the alpine region positive skill is sparely present (spatial levels 2 and 3 at Basel). These subcatchments generally exhibit a high terrain roughness as well as a high relevance of the initial conditions due to snow accumulation in winter and subsequent melting in spring and summer. A possible explanation could be that errors in the initial condition estimates outweigh the moderate skill contained in the seasonal climate predictions.

6.5 H-TESSEL

Within ECMWF's seasonal forecasting system S4, H-TESSEL is aimed to provide a lower boundary condition for the simulation of the atmosphere and consequently does neither implement streamflow routing nor groundwater storage (Balsamo et al., 2009; ECMWF, 2016). However, H-TESSEL in combination with a linear bias correction often performs best (Tab. 4).

The S4Q model, which has access to the same input data and in addition conditions on preceding precipitation and temperature, scores a lower forecast accuracy than H-TESSEL in case of Lobith (Tab. 4). This most likely is related to overfitting, which is not sufficiently smoothed by the model averaging (Sect. 4.1.2).

7 Conclusions

The present study tests a model output statistics (MOS) method for monthly and five day mean streamflow forecasts in the Rhine basin. The method relies on the linear regression model fitted by least squares and uses predictions of precipitation and surface air temperature from the seasonal forecast system S4 of the European Centre for Medium-Range Weather Forecasts. Observations of precipitation and surface air temperature prior to the date of prediction are employed as a surrogate for the

initial conditions. In addition, runoff simulated by the S4 land surface component, the H-TESSEL land surface model, is evaluated for its predictive power.

MOS methods often bridge the grid resolution of the dynamical model and the spatial scale of the actual predictand. In order to estimate how the forecast quality depends on the catchment area, a hindcast experiment for the period 1981-2011 is conducted that varies the working scale within the Rhine basin at Lobith and Basel. This variation is implemented by applying the MOS method to subcatchments and combining the resulting forecasts to predict streamflow at the main outlets at Lobith and Basel.

On average, the monthly mean streamflow forecasts based on the initial conditions are skillful with respect to the climatology at zero lead time for both the Rhine at Lobith and Basel. The MOS method, which in addition has access to the dynamical seasonal predictions, further reduces the mean absolute error by about 5 to 11 $\%$ compared to the model that is constrained to the initial conditions. For lead times of one and two months the forecasts virtually reduce to climatology. These results hold for the entire range of tested subcatchment scales, meaning that effects of a scale mismatch between the horizontal grid resolution and the catchment area do not emerge. Applying the MOS method finally for five day mean streamflow results in a rather moderate forecast quality.

We conclude that the present model formulation – in particular the assumption of linearity – is valid for the monthly time scale, catchments with areas up to 160000 km^2, and water travel times similar to the Rhine river. However, the results also show that a simple linear bias correction of the runoff predicted by the H-TESSEL land surface model is hard to beat. Given the simplicity of a linear bias correction, we think that it could be worth to further investigate runoff simulations from land surface components of earth system models for subseasonal to seasonal streamflow forecasting.

8 Code availability

The regression approach from Sect. 4.1.2 is implemented in an R package maintained on github.com/schiggo.

9 Data availability

E-OBS (2016), CORINE (2013), and EU-DEM (2013) are public data sets. Access to the ECMWF and GRDC archive must be requested. Data from the various public authorities as listed in the Acknowledgements are partly public.

10 Additional materials

The additional materials include Fig. 3, Fig. 4, and Fig. 5 for the S4* models. Figure 7 shows per spatial level for each S4* model at zero months lead time: The sample autocorrelation function and a quantile plot against the Gaussian distribution (paired differences of absolute residuals with respect to the preMet model, Eq. 3) as well as a scatterplot of predictions and observations. Figure 8 shows for the $y_{i,30}$ predictand the regression coefficients (for predictors standardised to mean zero and

standard deviation one) and the aggregation periods $a_{i,j}$ (Eq. 1) of the refRun model at spatial level one ($n = 100$ due to the bootstrap resampling).

Acknowledgements. Streamflow series and catchment boundaries are provided by the following public authorities: State Institute for the Environment, Measurements and Conservation Baden Wuerttemberg; Bavarian Environmental Agency; State of Voralberg; Austrian Federal Ministry of Agriculture, Forestry, Environment and Water; and Swiss Federal Office for the Environment. Further we acknowledge the E-OBS data set from the EU-FP6 project ENSEMBLES (ensembles-eu.metoffice.com) and the data providers in the ECA&D project (www.ecad.eu) as well as the Copernicus data and information funded by the European Union (EU-DEM and CORINE). We also thank the Global Runoff Data Centre and the European Centre for Medium-Range Weather Forecasts for the access to the data archives. The reviews and comments by Joost Beckers, Kean Foster, and Fredrik Wetterhall substantially improved the manuscript. The study was funded by the Group of Hydrology, which is part of the Institute of Geography at the University of Bern, Bern, Switzerland.

References

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., and Betts, A. K.: A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System, J. Hydrometeorol., 10, 623–643, doi:10.1175/2008JHM1068.1, 2009.

Beckers, J. V. L., Weerts, A. H., Tijdeman, E., and Welles, E.: ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction, Hydrol. Earth. Syst. Sc., 20, 3277–3287, doi:10.5194/hess-20-3277-2016, 2016.

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resour. Res., 52, 8238–8259, doi:10.1002/2016WR019193, 2016.

Blöschl, G. and Sivapalan, M.: Scale issues in hydrological modelling: A review, Hydrol. Process., 9, 251–290, doi:10.1002/hyp.3360090305, 1995.

Bosshard, T., Kotlarski, S., Zappa, M., and Schär, C.: Hydrological Climate-Impact Projections for the Rhine River: GCM–RCM Uncertainty and Separate Temperature and Precipitation Effects, J. Hydrometeorol., 15, 697–713, doi:10.1175/JHM-D-12-098.1, 2014.

Breiman, L.: Bagging Predictors, Mach. Learn., 24, 123–140, doi:10.1023/A:1018054314350, 1996.

Brunet, N., Verret, R., and Yacowar, N.: An Objective Comparison of Model Output Statistics and "Perfect Prog" Systems in Producing Numerical Weather Element Forecasts, Weather Forecast., 3, 273–283, doi:10.1175/1520-0434(1988)003<0273:AOCOMO>2.0.CO;2, 1988.

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., Hooper, R. P., Kumar, M., Leung, L. R., Mackay, D. S., Maxwell, R. M., Shen, C., Swenson, S. C., and Zeng, X.: Improving the representation of hydrologic processes in Earth System Models, Water Resour. Res., 51, 5929–5956, doi:10.1002/2015WR017096, 2015.

CORINE: Corine Land Cover 2006 raster data, http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3, last access: 26. October 2017, 2013.

Crochemore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrol. Earth. Syst. Sc., 20, 3601–3618, doi:10.5194/hess-20-3601-2016, 2016.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, Hydrol. Earth. Syst. Sc., 19, 275–291, doi:10.5194/hess-19-275-2015, 2015.

E-OBS: Daily temperature and precipitation fields in Europe, http://www.ecad.eu/download/ensembles/ensembles.php, last access: 26. October 2017, 2016.

ECMWF: IFS Documentation, http://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation, last access: 26. October 2017, 2016.

EU-DEM: Digital Elevation Model over Europe, http://www.eea.europa.eu/data-and-maps/data/eu-dem, last access: 26. October 2017, 2013.

Foster, K. L. and Uvo, C. B.: Seasonal streamflow forecast: a GCM multi-model downscaling approach, Hydrol. Res., 41, 503–507, doi:10.2166/nh.2010.143, 2010.

Fundel, F., Jörg-Hess, S., and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, Hydrol. Earth. Syst. Sc., 17, 395–407, doi:10.5194/hess-17-395-2013, 2013.

GRDC: The Global Runoff Data Centre, http://www.bafg.de/GRDC/EN/Home/homepage_node.html, last access: 26. October 2017, 2016.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res.-Atmos., 113, doi:10.1029/2008JD010201, 2008.

Hofstra, N., Haylock, M., New, M., and Jones, P. D.: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature, J. Geophys. Res.-Atmos., 114, doi:10.1029/2009JD011799, 2009.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., and Maier, H. R.: A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network, J. Hydrol., 540, 623– 640, doi:10.1016/j.jhydrol.2016.06.026, 2016.

ICPR: Internationally Coordinated Management Plan for the International River Basin District of the Rhine, International Commission for the Protection of the Rhine, 2009.

Jörg-Hess, S., Griessinger, N., and Zappa, M.: Probabilistic Forecasts of Snow Water Equivalent and Runoff in Mountainous Areas, J. Hydrometeorol., 16, 2169–2186, doi:10.1175/JHM-D-14-0193.1, 2015.

Maraun, D. and Widmann, M.: The representation of location by a regional climate model in complex terrain, Hydrol. Earth. Syst. Sc., 19, 3449–3456, doi:10.5194/hess-19-3449-2015, 2015.

Marcos, R., Llasat, M. C., Quintana-Segui, P., and Turco, M.: Seasonal predictability of water resources in a Mediterranean freshwater reservoir and assessment of its utility for end-users, Sci. Total Environ., 575, 681–691, doi:10.1016/j.scitotenv.2016.09.080, 2017.

Michaelsen, J.: Cross-Validation in Statistical Climate Forecast Models, J. Clim. Appl. Meteorol., 26, 1589–1600, doi:10.1175/1520-0450(1987)026<1589:CVISCF>2.0.CO;2, 1987.

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

National Academies: Next Generation Earth System Prediction, The National Academies Press, 1 edn., doi:10.17226/21873, 2016.

Orth, R. and Seneviratne, S. I.: Predictability of soil moisture and streamflow on subseasonal timescales: A case study, J. Geophys. Res.-Atmos., 118, 10,963–10,979, doi:10.1002/jgrd.50846, 2013.

Pappenberger, F., Cloke, H. L., Balsamo, G., Ngo-Duc, T., and Oki, T.: Global runoff routing with the hydrological component of the ECMWF NWP system, Int. J. Climatol., 30, 2155–2174, doi:10.1002/joc.2028, 2010.

Sahu, N., Robertson, A. W., Boer, R., Behera, S., DeWitt, D. G., Takara, K., Kumar, M., and Singh, R. B.: Probabilistic seasonal streamflow forecasts of the Citarum River, Indonesia, based on general circulation models, Stoch. Env. Res. Risk A., pp. 1–12, doi:10.1007/s00477-016-1297-4, 2016.

Schick, S., Rössler, O., and Weingartner, R.: Comparison of cross-validation and bootstrap aggregating for building a seasonal streamflow forecast model, Proceedings of the International Association of Hydrological Sciences, 374, 159–163, doi:10.5194/piahs-374-159-2016, 2016.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrol. Earth. Syst. Sc., 17, 2781–2796, doi:10.5194/hess-17-2781-2013, 2013.

Singla, S., Céron, J.-P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., and Vidal, J.-P.: Predictability of soil moisture and river flows over France for the spring season, Hydrol. Earth. Syst. Sc., 16, 201–216, doi:10.5194/hess-16-201-2012, 2012.

Slater, L. J., Villarini, G., Bradley, A. A., and Vecchi, G. A.: A dynamical statistical framework for seasonal streamflow forecasting in an agricultural watershed, Clim. Dynam., doi:10.1007/s00382-017-3794-7, 2017.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, doi:10.1029/2000JD900719, 2001.

Torma, C., Giorgi, F., and Coppola, E.: Added value of regional climate modeling over areas characterized by complex terrain–Precipitation over the Alps, J. Geophys. Res.-Atmos., 120, 3957–3972, doi:10.1002/2014JD022781, 2015.

van Dijk, A. I. J. M., Peña Arancibia, J. L., Wood, E. F., Sheffield, J., and Beck, H. E.: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resour. Res., 49, 2729–2746, doi:10.1002/wrcr.20251, 2013.

Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, Water Resour. Res., 48, doi:10.1029/2011WR010973, 2012.

Wilson, M. F. J., O'Connell, B., Brown, C., Guinan, J. C., and Grehan, A. J.: Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope, Mar. Geod., 30, 3–35, doi:10.1080/01490410701295962, 2007.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35, doi:10.1029/2008GL034648, 2008.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resour. Res., 49, 4687–4699, doi:10.1002/wrcr.20350, 2013.

Yuan, X., Roundy, J. K., Wood, E. F., and Sheffield, J.: Seasonal Forecasting of Global Hydrologic Extremes: System Development and Evaluation over GEWEX Basins, B. Am. Meteorol. Soc., 96, 1895–1912, doi:10.1175/BAMS-D-14-00003.1, 2015a.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, Wiley Interdisciplinary Reviews: Water, 2, 523–536, doi:10.1002/wat2.1088, 2015b.