Application of the model output statistics method to seasonal streamflow forecasting

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von
Simon Andreas Schick
von Marbach SG

Leiter der Arbeit:
Prof. Dr. R. Weingartner
Dr. O. Rössler
Geographisches Institut der Universität Bern
Application of the model output statistics method to seasonal streamflow forecasting

Inauguraldissertation der Philosophisch-naturwissenschaftlichen Fakultät der Universität Bern

vorgelegt von
Simon Andreas Schick von Marbach SG

Leiter der Arbeit:
Prof. Dr. R. Weingartner
Dr. O. Rössler

Geographisches Institut der Universität Bern

Von der Philosophisch-naturwissenschaftlichen Fakultät angenommen.

Bern, 22. November 2018  Der Dekan:
Prof. Dr. Z. Balogh
Summary

Seasonal streamflow forecasting aims at estimating the river streamflow volumes a few months ahead. Such predictions can be used in the context of water resources management to optimise for example reservoir operation. To improve the forecast quality, one major line of research is devoted to the utilisation of numerical seasonal climate predictions for seasonal streamflow forecasting.

Numerical seasonal climate predictions rely on coupled atmosphere-ocean-land models. These models implement diverse physical, chemical, and biological processes and steadily advance towards earth system models (ESMs). ESMs, which are used today for seasonal forecasting of environmental conditions, simulate land surface and subsurface runoff per grid cell. River streamflow, however, is an unresolved variable.

To resolve river streamflow by means of ESM simulations, the present thesis applies the model output statistics (MOS) method. By definition, the MOS method fits a statistical model to observations of the target variable and past forecasts of the numerical model. Subsequently, the statistical model is applied to new forecasts of the numerical model.

Here, the predictand is defined as mean streamflow of a monthly time window. Occasionally, the length of this time window is varied to test the MOS method at shorter and longer time scales. The pool of candidate predictors contains ESM simulations of runoff, precipitation, and surface air temperature, which are provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Hindcast data from both the old (S4) as well as current (SEAS5) seasonal forecast systems are used. In addition, the pool includes observations of precipitation, temperature, and streamflow, which temporally lag behind the predictand.

The MOS method is developed and tested with three experiments covering the period 1981–2010 and including around 150 European catchments. These experiments focus on five issues: (1) the assumption of linear predictor-predictand relationships as a working hypothesis; (2) model variance as introduced by small sample sizes and noisy predictors; (3) potential effects of a scale mismatch between the ESM grid resolution and the area of the target catchment; (4) the comparison of different predictor combinations; and (5) predictive skill.
Regarding these five issues, the main findings are:

1. The assumption of linearity provides a fair compromise between validity of the model formulation and size of the training set. However, this requires that the predictand does not have a higher temporal resolution than about 20 days.

2. Model variance can be reduced by using a procedure called “bootstrap aggregating”. For the present prediction problem, it is estimated that bootstrap aggregating reduces the mean squared error of prediction on average by about 7%.

3. The present implementation of the MOS method works at a range of spatial scales. This range includes catchments that fall below the ESM grid scale as well as catchments that span large arrays of grid cells.

4. In general a simple linear regression using ESM-simulated runoff as predictor already provides a skill level that is hard to beat with more complex model formulations. An exception to this rule seems to exist for catchments that feature lakes, which extend to a substantial fraction of the catchment area. Additional predictors are needed in this case.

5. For the present set of European catchments and the monthly time scale, skill with respect to climatology restricts on average to the first month ahead. For one-month-ahead forecasts, the mean absolute error of the climatology is decreased by about 25%. Large parts of this skill can be attributed to the initial hydrological conditions.

Being a data-driven approach, the MOS method tries to tackle the prediction problem in a pragmatic and computationally efficient way. However, this comes at the cost of predictions that exhibit a low temporal resolution and are restricted to observing sites along the river network. Thus, I argue that the MOS method could be useful in an operational context by complementing existing approaches.
## Contents

Summary i

Figures v

Tables vi

Acronyms vii

Units viii

1 Introduction 1
   1.1 Problem 1
   1.2 Objective 2
   1.3 Outline 2

2 Comparison of cross-validation and bootstrap aggregating for building a seasonal streamflow forecast model 4
   2.1 Abstract 4
   2.2 Introduction 4
   2.3 Data 5
   2.4 Method 6
   2.5 Results 8
   2.6 Discussion 10
   2.7 Conclusion 12

3 Monthly streamflow forecasting at varying spatial scales in the Rhine basin 14
   3.1 Abstract 14
   3.2 Introduction 15
   3.3 Study region 18
   3.4 Data 19
   3.5 Method 20
   3.6 Results 25
   3.7 Discussion 32
   3.8 Conclusion 35
   3.9 Supplement 36
4 An evaluation of model output statistics for monthly streamflow forecasting in Europe 37
  4.1 Abstract ........................................ 37
  4.2 Introduction ................................... 37
  4.3 Data ............................................ 39
  4.4 Method ......................................... 42
  4.5 Results ......................................... 47
  4.6 Discussion ..................................... 52
  4.7 Conclusion ..................................... 58
  4.8 Supplement .................................... 59

5 Synthesis 61
  5.1 Methodological remarks ...................... 61
  5.2 Main findings .................................. 63
  5.3 Conclusion ..................................... 65
  5.4 Outlook ........................................ 65

6 Appendix 67
  6.1 Data sources ................................... 67
  6.2 Comparison of OOB and LOO .................. 68
  6.3 Computer code .................................. 70

Bibliography 71

Acknowledgements 80
Figures

2.1 Study region .................................................... 6
2.2 $E_{MSP}$ outer cross-validation ............................... 9
2.3 $E_{MSP}$ estimates of LOO and OOB ......................... 11

3.1 Monthly climatology of the Rhine River ................... 18
3.2 Taylor diagram ................................................. 26
3.3 MAESS conditional on the date of prediction ............... 27
3.4 MAESS per subcatchment .................................... 29
3.5 MAESS versus subcatchment attributes ...................... 31
3.6 Correlation of 5-day mean streamflow (45-day lead time) 32
3.7 Correlation of 5-day mean streamflow (175-day lead time) 33

4.1 Study region .................................................... 41
4.2 MAESS, MSESS, and correlation ............................. 48
4.3 PIT value distribution (0-day lead time) .................... 50
4.4 CRPSS with benchmark climatology (SEAS5 hindcast) ... 51
4.5 CRPSS with benchmark preMet model (SEAS5 hindcast) 51
4.6 Selected examples (S4 hindcast) ............................. 53
4.7 CRPSS with benchmark climatology (S4 hindcast) ......... 60
Tables

2.1 Statistical significance of $E_{MSP}$ differences .............. 9
2.2 $E_{MSP}$ of BGS and BAG .................................. 10

3.1 Geography of the Rhine River ................................. 19
3.2 Predictor combinations ....................................... 21
3.3 Subcatchment division ....................................... 23
3.4 MAE (0-month lead time) ................................... 28
3.5 MAESS (0-month lead time) .................................. 28

4.1 Catchment attributes ......................................... 40
4.2 Predictor combinations ....................................... 44
### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAG</td>
<td>BAGged model</td>
</tr>
<tr>
<td>BGS</td>
<td>Best Guess Model</td>
</tr>
<tr>
<td>CRPS</td>
<td>Continuous Ranked Probability Score</td>
</tr>
<tr>
<td>CRPSS</td>
<td>Continuous Ranked Probability Skill Score</td>
</tr>
<tr>
<td>$E_{MSP}$</td>
<td>Mean Squared Error of Prediction</td>
</tr>
<tr>
<td>ECMWF</td>
<td>European Centre for Medium-Range Weather Forecasts</td>
</tr>
<tr>
<td>E-OBS</td>
<td>ENSEMBLES daily gridded OBServational dataset in Europe</td>
</tr>
<tr>
<td>ESM</td>
<td>Earth System Model</td>
</tr>
<tr>
<td>ESP</td>
<td>Ensemble Streamflow Prediction</td>
</tr>
<tr>
<td>GRDC</td>
<td>Global Runoff Data Centre</td>
</tr>
<tr>
<td>H-TESSEL</td>
<td>Hydrology-revised TESSEL</td>
</tr>
<tr>
<td>LOO</td>
<td>Leave One Out</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAESS</td>
<td>Mean Absolute Error Skill Score</td>
</tr>
<tr>
<td>MOS</td>
<td>Model Output Statistics</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>MSESS</td>
<td>Mean Squared Error Skill Score</td>
</tr>
<tr>
<td>OOB</td>
<td>Out Of Bag</td>
</tr>
<tr>
<td>PIT</td>
<td>Probability Integral Transform</td>
</tr>
<tr>
<td>PP</td>
<td>Perfect Prognosis</td>
</tr>
<tr>
<td>revESP</td>
<td>reverse Ensemble Streamflow Prediction</td>
</tr>
<tr>
<td>S4</td>
<td>ECMWF’s seasonal forecast System 4</td>
</tr>
<tr>
<td>SEAS5 (S5)</td>
<td>ECMWF’s SEAsonal forecast System 5</td>
</tr>
<tr>
<td>SRG</td>
<td>Seasonal ReGime</td>
</tr>
<tr>
<td>TESSEL</td>
<td>Tiled ECMWF Scheme for Surface Exchanges over Land</td>
</tr>
</tbody>
</table>
## Units

- `°` degree (plane angle)
- `°C` degree Celsius
- `d` day
- `hPa` hecto pascal
- `km²` square kilometre
- `m` metre
- `m a.s.l.` metre above sea level
- `min` minute
- `mm` millimetre
- `m³ s⁻¹` cubic metre per second
- `s` second
1 Introduction

River streamflow forecasting at the seasonal time scale aims at providing planning criteria for hydro power production (Arsenault, Latraverse, and Duchesne 2016), water allocation (Chiew, Zhou, and McMahon 2003), or inland navigation (Meißner, Klein, and Ionita 2017). In a few countries seasonal streamflow forecasts are an integral part of the public service, including Australia (Schepen and Wang 2015), the United Kingdom (Prudhomme et al. 2017), and the United States (Pagano et al. 2014).

In Switzerland, efforts to monitor and predict droughts centre around the drought.ch platform, which also addresses seasonal streamflow forecasting (Kruse and Seidl 2013; Zappa et al. 2014). Besides activities on the national level, existing continental or global flood forecast systems have been recently adapted to the seasonal time scale (Arnal et al. 2018; Emerton et al. 2018).

In this thesis, seasonal forecasts are pragmatically defined as forecasts in the range of 2 weeks up to 1 year ahead. In addition, subseasonal forecasts occasionally denote forecasts in the range of 2 weeks up to 3 months ahead. These definitions imply that subseasonal and seasonal forecasts fill the gap between short term forecasts and annual or decadal forecasts (Doblas-Reyes et al. 2013; National Academies 2016).

1.1 Problem

Approaches to implement a seasonal streamflow forecast system cover a broad spectrum (Yuan, Wood, and Ma 2015): Traditional approaches regress streamflow against hydrometeorological variables known at the date of prediction. Today, state-of-the-art approaches revolve around hydrological simulation models forced with numerical seasonal climate predictions.

Since the 1990s, operational seasonal climate predictions are increasingly based on global, coupled atmosphere-ocean general circulation models (Carson 1998; Stockdale et al. 2010). These coupled models implement diverse physical, chemical, and biological processes and progressively develop towards earth system models (ESMs).

Land surface schemes of ESMs generally simulate runoff per grid cell. However, they simplify or do not implement hydrological processes such as groundwater dynamics or the river routing (Clark et al. 2015). Thus, river streamflow is an unresolved variable in current generation ESMs used for seasonal forecasting of environmental conditions.

Arguing with computational efficiency, Yuan, Wood, and Ma (2015) suggest as a potential direction for future research the statistical postprocessing
of ESM-simulated runoff. The postprocessing is needed to compensate for hydrological processes not implemented in the ESM, to account for potential biases, to meet the spatial and temporal scale of the forecast problem, or to model the forecast’s probability distribution.

1.2 Objective

The objective of the present thesis is to investigate the potential of statistically postprocessed ESM simulations for seasonal streamflow forecasting. To do so, the model output statistics (MOS) method (Glahn and Lowry 1972; Klein and Glahn 1974) is applied and the predictive skill of the postprocessed simulations estimated in several hindcast experiments.

The MOS method builds the statistical model using past forecasts of the dynamical model and corresponding observations of the target variable. This definition does not refer to a particular regression or classification technique, hence the MOS method can be considered as a conceptual framework. An advantage of the MOS method is that it tailors the statistical postprocessing to the dynamical model and its features, including the spatial resolution and potential biases.

In the following, the predictand is defined as mean streamflow of a given time window, whose length is often set to 1 month. Besides ESM-predicted runoff, the pool of candidate predictors contains lagged (i.e. observed) and concurrent (i.e. ESM-predicted) surface air temperature and precipitation. In one experiment, also lagged streamflow is added to the pool of candidate predictors.

This selection is based on three considerations:

1. The MOS method is constrained to data that are easily available from hydrometeorological measurement networks.

2. For precipitation and surface air temperature observation-based gridded products exist.

3. It is not required that a catchment of interest features specific geographical attributes as it would be for variables like lake levels or snow courses.

1.3 Outline

The MOS method is developed and tested with hindcasts spanning approximately the period 1981–2010. These hindcasts share the assumption of linear predictor-predictand relationships as a working hypothesis and are bundled in three experiments:
• In Sect. 2, a prototype of the MOS method is tested. It conditions seasonal streamflow on hydrometeorological variables preceding the date of prediction and does not yet include ESM simulations. The focus is put on how to deal with model variance induced by small sample sizes and noisy predictors. This experiment is conducted using subcatchments of the Swiss Rhine.

• In Sect. 3, the statistical model actually becomes a MOS method by applying it to ESM simulations. These simulations are from the seasonal forecast system S4 of the European Centre for Medium-Range Weather Forecasts (ECMWF). The focus lies on how the performance of the MOS method varies at different spatial scales and on the comparison of different predictor combinations. Here, the study region comprises the Rhine catchment.

• In Sect. 4, the MOS method is tested in 16 European catchments featuring different climatic and geographical conditions. Input to the MOS method is from ECMWF’s forecast systems S4 and SEAS5. The focus is again on the comparison of different predictor combinations. In addition, emphasis is placed on a comparison with studies that use similar hindcast configurations.

Section 5 then makes some methodological remarks, lists the main findings, draws a conclusion, and suggests some potential directions for future research.
2 Comparison of cross-validation and bootstrap aggregating for building a seasonal streamflow forecast model


Note: The published article contains a wrong formulation of the bootstrap null hypothesis; this error is corrected below. Section 6.2 contrasts the leave-one-out and out-of-bag prediction error estimates in more detail.

2.1 Abstract

Based on a hindcast experiment for the period 1982–2013 in 66 subcatchments of the Swiss Rhine, the present study compares two approaches of building a regression model for seasonal streamflow forecasting. The first approach selects a single “best guess” model, which is tested by leave-one-out cross-validation. The second approach implements the idea of bootstrap aggregating, where bootstrap replicates are employed to select several models, and out-of-bag predictions provide model testing. The target value is mean streamflow for durations of 30, 60, and 90 d, starting with the 1st and 16th day of every month. Compared to the best guess model, bootstrap aggregating reduces the mean squared error of the streamflow forecast by 7% on average. Thus, if resampling is anyway part of the model building procedure, bootstrap aggregating seems to be a useful strategy in statistical seasonal streamflow forecasting. Since the improved accuracy comes at the cost of a less interpretable model, the approach might be best suited for pure prediction tasks, e.g. as in operational applications.

2.2 Introduction

Small sample sizes challenge the application of statistical models for seasonal streamflow forecasting. For example, a daily hydrometeorological time series of length 30 years can be considered as a long record. However, at seasonal time scales the series provides 30 cases (e.g. summer means). Following the nomenclature described by Hastie, Tibshirani, and Friedman (2009), the model building procedure then has to cope with these 30 cases for:
1. model training, i.e. fit models with varying complexity or different
   predictors;

2. model selection, i.e. validate the models and choose the best one(s);
   and

3. model testing, i.e. estimate the final model’s prediction error (possibly
   by combining several models).

To overcome small sample sizes, resampling is commonly used for model
selection and testing. In addition, seasonal streamflow forecasting often
encounters weak predictor-predictand relationships, introduced by missing
or noisy predictors – e.g. precipitation and temperature of the target season.
Models out of any resampling thus can differ markedly, which leads us to
the following question: Are there any benefits if the models resulting from
resampling are combined in a systematic way? To address this question,
we compare (1) the selection of a single “best guess” model along with
leave-one-out cross-validation against (2) bootstrap aggregating along with
out-of-bag prediction error estimates. Bootstrap aggregating was introduced
by Breiman (1996a, “bagging” or “bagged model” for short) and aims to
reduce the variance of a statistical model by applying it to bootstrap
replicates of the data set and combining the corresponding predictions
afterwards.

Below Sect. 2.3 briefly presents the data set, Sect. 2.4 outlines the
methodology, and in Sect. 2.5 and 2.6 the results are presented and discussed,
respectively.

2.3 Data

The hindcast experiment is applied to 66 subcatchments (no nesting) of the
Swiss Rhine at Basel, ranging in mean elevation from 500 to 2300 m and in
area from 20 to 900 km² (Fig. 2.1). Streamflow is regulated and routed for
the purpose of hydro power, flood protection, water supply, and ecological
conservation. Up to ten catchments can be considered as heavily regulated;
for the remaining catchments we assume that anthropogenic effects on the
catchment’s hydrology do not have any impacts at seasonal time scales.

Daily mean streamflow in m³ s⁻¹ for the period 1982–2013 is provided
by public authorities of Germany, Austria, and Switzerland (Sect. 6.1),
whereas daily precipitation and temperature series are catchment averages
derived from the E-OBS gridded data set version 12.0 in 0.25° resolution
(approximately 19 and 28 km in longitude and latitude; Haylock et al. 2008).
2.4 Method

The comparison of the two model building procedures relies itself on the principle of cross-validation, i.e. some cases are in turn excluded from the data set and the complete model building procedures are conducted by using the remaining cases. Section 2.4.1 first introduces the regression model, which is common to both model building procedures. Section 2.4.2 then describes the two resampling approaches; in case of the best guess model, resampling is solely used to estimate the prediction error, whereas in case of the bagged model resampling is at the heart of the model building procedure. Section 2.4.3 finally states the cross-validation implementation and the statistical test in order to contrast the two procedures.

2.4.1 Regression model

The regression model follows closely the approach of Garen (1992), i.e. initial conditions are considered only. The predictand $y_{i,j}$ is in turn mean streamflow of duration $i = 30, 60, 90$ d, starting at the 1st and 16th day of every month (date of prediction $j = 1, \ldots, 24$). For a particular choice of $i$ and $j$, the regression equation is given by

$$y_{i,j} = b_0 \mathbf{1} + X \mathbf{b} + \varepsilon$$

(2.1)

where $b_0$ denotes the intercept, $\mathbf{1}$ a vector of ones, $\mathbf{b}$ the vector of regression coefficients, and $\varepsilon$ the errors. The $n \times p$ matrix $X$ has in its $p = 3$ columns antecedent streamflow, antecedent precipitation, and antecedent temperature as predictors; $n$ equals the number of years. The time aggregation
is individually selected for each predictor according to Spearman’s rank correlation, but has to be one of 10, 20, …, 720 d.

Since these predictors can be highly correlated, the regression coefficients \( b \) are estimated using partial least squares (Mevik and Wehrens 2007). Partial least squares is related to principal components regression, but decomposes the cross-covariance matrix \( X^T y \) instead of the predictor’s covariance matrix. Regarding model selection, we decide to select at least the first partial least squares direction, as otherwise the regression model shrinks to \( b_0 \) in Eq. (2.1). Please note that we do not make any distributional assumptions about \( \varepsilon \).

### 2.4.2 Resampling approaches

The regression model from Eq. 2.1 is applied twofold for a particular catchment and predictand \( y_{i,j} \):

1. A single best guess model is selected and the mean squared error of prediction \( E_{MSP} \) is estimated according to leave-one-out cross-validation.

2. For each of 100 bootstrap replicates of the data set, one model is selected. These models are then combined by simply averaging their predictions (bootstrap aggregating; Breiman 1996a). Here, out-of-bag predictions (Breiman 1996b) are used to estimate \( E_{MSP} \).

Breiman (1996a) showed that the aggregation of unstable models can help to decrease the prediction error. Here, instability (or high model variance) refers to the case when small changes in the data set lead to large changes in the final estimated model. A simple linear model fitted by ordinary least squares can be considered as an example for a stable model whereas neural networks or regression trees generally are examples for unstable models. The present model from Sect. 2.4.1 is in our view a stable model – it is linear, consists of three predictors (where only the time aggregations are allowed to vary), and partial least squares further tries to reduce the dimensionality of the predictor space.

The out-of-bag approach is closely related to the leave-one-out procedure in that one case is left out at a time, i.e. the model averaging considers only those models for which the “left-out” case was not included in the corresponding replicates. Since prediction error estimates of out-of-bag and leave-one-out approximately converge with an increasing number of bootstrap replicates, the out-of-bag estimate provides a convenient alternative – testing a bagged model via leave-one-out can be computationally expensive due to the involved bootstrap. Finally, the choice of 100 replicates is based
on the recommendation by Hastie, Tibshirani, and Friedman (2009) that model training can be stopped as soon as the out-of-bag error has stabilised.

Hereafter, the two approaches are named BGS/LOO (best guess model BGS in combination with leave-one-out LOO) and BAG/OOB (bagged model BAG in combination with out-of-bag OOB), respectively.

2.4.3 Hindcast experiment

In order to contrast BGS/LOO with BAG/OOB, the 32 year period of investigation is used for an additional leave-one-out cross-validation. Doing so, we get an estimate of $E_{MSP}$ independently of LOO and OOB. Here, also the 3 adjacent years of the left out case are omitted to avoid spurious skill due to catchment memory (hence $n = 25$ in Eq. 2.1). Since BGS/LOO and BAG/OOB are nested inside this “buffered” leave-one-out cross-validation, we refer to the latter as the outer cross-validation. Considering a particular catchment and predictand $y_{i,j}$, three steps are applied:

1. $y_{i,j}$ is centred to mean 0 and scaled to standard deviation 1 with respect to the period 1982–2013.

2. Each year (together with its 3 adjacent years) is left out once, while the remaining years are used for the application of the model building procedures BGS/LOO and BAG/OOB.

3. The mean value of $y_{i,j}$ serves as a competing model (hereafter named the seasonal regime, SRG). Analogue to BGS, $E_{MSP}$ is estimated by LOO as well as the outer cross-validation.

Paired differences of $E_{MSP}$ are then used for inference. Here, paired differences are calculated such that $E_{MSP}$ of the more complex model is subtracted from $E_{MSP}$ of the less complex model (always per catchment). The mean difference $\mu$ is used for a right-sided $t$-test (alternative hypothesis $\mu > 0$). Also a nonparametric bootstrap is applied to estimate the probability $P(\mu \leq 0)$, since the differences do not necessarily follow a Gaussian distribution.

2.5 Results

The results are arranged in three sections: Firstly, we contrast BGS with BAG in order to see whether bagging improves the predictions. Secondly, model skill is evaluated by comparing BGS and BAG against SRG. Thirdly, we analyse the accuracy of $E_{MSP}$ estimates from LOO and OOB. In the following subscript $j$ is dropped when error statistics are averaged over $j$. 

8
Figure 2.2: $E_{MSP}$ for all catchments and predictands as obtained from the outer cross-validation; $n = 66$.

Table 2.1: $p$ values for the null hypothesis “the simple model outperforms the complex model”, estimated by a right-sided $t$-test (paired differences with mean difference $\mu$ and alternative hypothesis $\mu > 0$). Paired differences here follow the rule that $E_{MSP}$ of the more complex model is subtracted from $E_{MSP}$ of the less complex model, as specified in the first column. In parentheses also the probabilities $P(\mu \leq 0)$ according to a nonparametric bootstrap with 10 000 replicates are listed. Only $E_{MSP}$ from the outer cross-validation is considered; $n = 66$.

<table>
<thead>
<tr>
<th></th>
<th>$y_{30}$</th>
<th>$y_{60}$</th>
<th>$y_{90}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRG–BGS</td>
<td>0.10 (0.09)</td>
<td>0.99 (0.99)</td>
<td>0.99 (0.99)</td>
</tr>
<tr>
<td>SRG–BAG</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.44 (0.45)</td>
</tr>
<tr>
<td>BGS–BAG</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
</tr>
</tbody>
</table>

2.5.1 Comparison of prediction error

The comparison of BGS against BAG focuses on $E_{MSP}$ from the outer cross-validation: Fig. 2.2 suggests that BAG scores on average the smaller $E_{MSP}$. Also the $p$ values indicate that BAG is most likely able to reduce $E_{MSP}$ (third row in Tab. 2.1). Table 2.2 lists additionally $E_{MSP}$ of BGS and BAG for $y_{30}$, $y_{60}$, and $y_{90}$, averaged over all catchments (i.e. the mean value of the corresponding whisker boxes in Fig. 2.2). Independently of the predictand, the reduction of $E_{MSP}$ by using BAG instead of BGS amounts to 7 to 8%.
Table 2.2: $E_{MSP}$ of BGS and BAG from the outer cross-validation, averaged over all catchments; $E_{MSP}$ is based on centred and standardised $y_{i,j}$. The last row indicates the reduction in $E_{MSP}$ when BAG is used instead of BGS.

<table>
<thead>
<tr>
<th></th>
<th>$y_{30}$</th>
<th>$y_{60}$</th>
<th>$y_{90}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGS</td>
<td>1.02</td>
<td>1.09</td>
<td>1.13</td>
</tr>
<tr>
<td>BAG</td>
<td>0.95</td>
<td>1.00</td>
<td>1.04</td>
</tr>
<tr>
<td>1−BAG/BGS</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
</tr>
</tbody>
</table>

2.5.2 Comparison of model skill

For the evaluation of model skill we focus again on the $E_{MSP}$ estimates from the outer cross-validation (Fig. 2.2). Due to standardisation of $y_{i,j}$, the benchmark model SRG shows an $E_{MSP}$ near 1 for all catchments (a perfectly estimated mean value would yield an $E_{MSP}$ of 1). On average SRG is a serious competitor and outperforms BGS and BAG in several catchments. Reduction of $E_{MSP}$ by using BGS and BAG instead of SRG is strongest for $y_{30}$ and weakest for $y_{90}$. These findings are also supported by Tab. 2.1, which reports the $p$ values of the $t$-test and the bootstrap: It is questionable to unlikely that BGS reduces $E_{MSP}$ on average, whereas BAG very likely does so for $y_{30}$ and $y_{60}$, but not for $y_{90}$.

2.5.3 Comparison of prediction error estimation

Figure 2.3 shows the differences in $E_{MSP}$, when LOO and OOB estimates are subtracted from the estimates obtained in the outer cross-validation, which are here considered to be the reference. Thus, a positive difference can be attributed as an underestimation and a negative difference as an overestimation of the prediction error. Apart from a few outliers, the differences lie in the interval $[-0.1, 0.1]$ and are symmetrically centred around 0 – on average neither LOO nor OOB tend to optimism or pessimism. The heavy negative outliers correspond to the same catchment, which turns out to be regulated due to hydro power.

2.6 Discussion

In the present study, a hindcast experiment was conducted that mimics the operational use of a simple forecasting system. The objective was the comparison of two model model building procedures, which both rely on the same regression model, but use different resampling strategies: A single best
Figure 2.3: $E_{MSP}$ estimates of LOO (in case of SRG and BGS) and OOB (in case of BAG) subtracted from $E_{MSP}$ estimates of the outer cross-validation for all catchments and predictands; $n = 66$.

guess model, which is tested by leave-one-out cross-validation, and a bagged model, which employs the bootstrap technique in order to build an ensemble of models. An useful byproduct of bagging is the out-of-bag prediction error estimate, which in theory can replace an additional resampling. Regarding the methodology, several points need some attention:

- Catchments were selected without a priori reasoning about their adequacy for seasonal streamflow forecasting. Strictly speaking, none of these catchments exhibits natural streamflow, though some anthropogenic effects might be averaged out due to the seasonal time scale. However, most of these effects are hardly quantifiable and it is not clear whether or not they favour model skill.

- The standardisation of $y_{i,j}$ attaches all seasons and catchments equal weights for the analysis. Doing so, model skill in seasons/catchments with large streamflow variability is masked, e.g. in spring when snow melting occurs and the models perform best (not shown).

- In order to compare the model building procedures, $E_{MSP}$ estimates from the outer cross-validation are considered as the “true” values. This assumption is indeed critical, but unavoidable in the present context – otherwise the real-world data set has to be replaced with a synthetic one.

- The residual analysis (not shown) reveals that the prediction errors are not independent and identically distributed. High flow is commonly underestimated, whereas low flows are often overestimated. Technically, the model can be considered as misspecified, since it lacks
relevant predictors (most likely precipitation and temperature during the season to predict). Therefore, common techniques to estimate prediction intervals are not applicable. It remains to be tested whether a substitution of the missing predictors by climate indices or seasonal climate predictions mitigates model misspecification.

2.7 Conclusion

The results are valid only for the present data set, though the sample size of 66 catchments in combination with 72 predictands might permit more general conclusions:

- BAG scores on average the lower $E_{MSP}$ than BGS. Bagging is useful if the model is unstable (Breiman 1996a). Since we consider the applied model as rather simple and stable, we argue that instability is introduced by weak predictor-predictand relationships in combination with small sample sizes. These weak (and sometimes spurious) relationships propagate through the screening of the time aggregation, the selection of partial least squares directions, and the final regression coefficients. Small changes in the data set thus often cause that completely different models are identified as the correct one.

- For 30 and 60 d mean streamflow, BAG outperforms in the majority of catchments a naive forecasting strategy, which relies on long-term averages only (SRG). Otherwise it is either questionable (30 d mean streamflow in case of BGS and 90 d mean streamflow in case of BAG) or very unlikely that BGS and BAG provide on average a smaller $E_{MSP}$ than SRG.

- LOO and OOB estimates of $E_{MSP}$ are for most catchments close to $E_{MSP}$ from the outer cross-validation. Neither LOO nor OOB tend to optimistic or pessimistic estimates. Thus, instead of testing the bagged model via the outer cross-validation, also the OOB estimates had been quite accurate.

In practice, statistical seasonal streamflow forecasting is commonly confronted with small sample sizes and weak predictor-predictand relationships due to missing or noisy predictors. The results of the present study indicate that bagging is also able to reduce a pseudo model variance, introduced by weak relationships and intensified by small sample sizes. If resampling is anyway part of the model building procedure and weak relationships come along with small sample sizes, we propose to prefer bagging to the best guess model approach – the computational costs are nearly the same,
out-of-bag predictions provide model testing, and prediction errors are likely to decrease. This benefit however comes at the cost of a hardly interpretable model. We thus argue that bagging is most useful when prediction alone is the goal, i.e. in operational forecasting, be it seasonal streamflow or another environmental variable.
3 Monthly streamflow forecasting at varying spatial scales in the Rhine basin


Note: For formal consistency throughout the thesis, the notation regarding the lead time has been standardised and differs from the original publication.

3.1 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 doing so, the MOS method is tested for catchments areas covering 4 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, reduce the mean absolute error of the former in the range of 5 to 12%, which is consistently reproduced at the subcatchment scale. An additional experiment conducted for 5-day mean streamflow indicates that the dynamical predictions help to reduce uncertainties up to about 20 d ahead, but it also reveals some shortcomings of the present MOS method.
3.2 Introduction

Environmental forecasting at the subseasonal to seasonal timescale 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 the 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 0 up to several months (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, Wood, and Ma 2015). Predictions from an ESM can be used threefold towards 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. (2015) 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 ESM involved, but can take systematic errors of the ESM into account (Brunet, Verret, and Yacowar 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. 2017), generalised 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.

16
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) are entered into the regression equation and the resulting forecast skill is estimated with respect to the ESP-inspired regression model.

The variation of the catchment area is borrowed 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 the 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 1 and 2 weeks, which increases to about 1 month when focusing on low flows (Fundel, Jörg-Hess, and Zappa 2013; Jörg-Hess, Griessinger, and Zappa 2015). Also for daily low flow Demirel, Booij, and Hoekstra (2015) report a sharp decrease in skill after 30 d for the Moselle River. For a set of French catchments Crochemore, Ramos, and Pappenberger (2016) show that weekly streamflow forecasts are improved for lead times up to about 1 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 0, 1, and 2 months. In order to resolve the monthly timescale and to test the MOS method at shorter time intervals, an experiment is further conducted for 5-day mean streamflow. Here, 0 lead time refers to forecasting one time interval ahead, while e.g. a 1-month lead time denotes a temporal gap of 1 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. 3.3 introduces the
study region, Sect. 3.4 describes the data set, Sect. 3.5 exposes the methodology in more detail, and in Sects. 3.6 and 3.7 the results are presented and discussed, respectively.

### 3.3 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, hydropower, 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 3.1 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 grassland, wasteland, surface water, and glacier.

Concerning the climatology of the period 1981–2011 (Fig. 3.1), 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 21st century (Bosshard et al. 2014).
Table 3.1: Geography of the Rhine River at Basel and Lobith according to CORINE (2016), EU-DEM (2016), and GRDC (2016).

<table>
<thead>
<tr>
<th></th>
<th>Lobith</th>
<th>Basel</th>
</tr>
</thead>
<tbody>
<tr>
<td>area (km²)</td>
<td>159 700</td>
<td>36 000</td>
</tr>
<tr>
<td>gauging station (m a.s.l.)</td>
<td>20</td>
<td>250</td>
</tr>
<tr>
<td>elevation min (m a.s.l.)</td>
<td>−230</td>
<td>250</td>
</tr>
<tr>
<td>elevation max (m a.s.l.)</td>
<td>4060</td>
<td>4060</td>
</tr>
<tr>
<td>elevation mean (m a.s.l.)</td>
<td>490</td>
<td>1050</td>
</tr>
<tr>
<td>farmed area (%)</td>
<td>47.7</td>
<td>36.8</td>
</tr>
<tr>
<td>forest (%)</td>
<td>35.8</td>
<td>31.6</td>
</tr>
<tr>
<td>grassland (%)</td>
<td>3.4</td>
<td>11.4</td>
</tr>
<tr>
<td>urban area (%)</td>
<td>9.6</td>
<td>7.0</td>
</tr>
<tr>
<td>wasteland (%)</td>
<td>1.8</td>
<td>8.2</td>
</tr>
<tr>
<td>surface water (%)</td>
<td>1.4</td>
<td>4.0</td>
</tr>
<tr>
<td>glacier (%)</td>
<td>0.3</td>
<td>1.0</td>
</tr>
</tbody>
</table>

3.4 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.

3.4.1 Observations

The streamflow observations consist of a set of 135 time series in m³ 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 and Sect. 6.1) and belong to catchments with nearly natural to heavily regulated streamflow.

The ENSEMBLES gridded observational data set in Europe (E-OBS, version 16.0) provides precipitation and surface air temperature on a 0.25° regular grid (Haylock et al. 2008; E-OBS 2017). These fields are based 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 to a high correlation (Hofstra et al. 2009). In the case of the Rhine basin an E-OBS tile approximately covers an area
of 500 km\(^2\).

### 3.4.2 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 the case of the Rhine basin to a tile area of about 4500 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 7 months. Up to 2010, initial conditions are from 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 5 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 2017a).

### 3.5 Method

The following subsections outline the experiment, which is individually conducted for both the Rhine at Lobith and Basel. Section 3.5.1 details the predictor combinations and the regression strategy, Sect. 3.5.2 introduces the variation of the catchment area, and Sect. 3.5.3 illustrates the validation of the resulting hindcasts.

#### 3.5.1 Model building

The predictand \( y_{i,j} \) denotes observations of mean streamflow at a specific gauging site in 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.
Table 3.2: 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 from the E-OBS gridded data set or ECMWF’s S4 hindcast archive.

<table>
<thead>
<tr>
<th>model</th>
<th>$p^\text{pre}$</th>
<th>$t^\text{pre}$</th>
<th>$p^\text{sub}$</th>
<th>$t^\text{sub}$</th>
<th>$q^\text{sub}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>refRun</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>-</td>
</tr>
<tr>
<td>preMet</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>subMet</td>
<td>-</td>
<td>-</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>-</td>
</tr>
<tr>
<td>S4P</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>S4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S4T</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>-</td>
<td>S4</td>
<td>-</td>
</tr>
<tr>
<td>S4PT</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>S4</td>
<td>S4</td>
<td>-</td>
</tr>
<tr>
<td>S4Q</td>
<td>E-OBS</td>
<td>E-OBS</td>
<td>-</td>
<td>-</td>
<td>S4</td>
</tr>
</tbody>
</table>

**Predictor combinations**

The set of predictors consists of variables that either precede or succeed the date of prediction $i$ (Tab. 3.2). The first model refRun (reference run) aims 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 from 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.

**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

$$a_{i,j} = \arg \max_k | \text{cor}(y_{i,j}, x_{i,k}) |$$  \hspace{1cm} (3.1)

where $x_{i,k}$ is one of the predictors from Tab. 3.2 and $k = -10, -20, \ldots, -720$ d in the case of $p^\text{pre}$ and $t^\text{pre}$ (backward in time relative to the date of prediction) and $k = 5, 10, \ldots, j$ d in the case of $p^\text{sub}$, $t^\text{sub}$, and $q^\text{sub}$ (forward in
time relative to the date of prediction). The limit of 720 d is chosen since larger values rarely get selected.

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. 3.2 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. 3.5.1), 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 1996a; Schick, Rössler, and Weingartner 2016).

**Cross-validation**

Each year with a buffer of 2 years (i.e. the 2 preceding and subsequent years) is left out and the regression outlined in Sect. 3.5.1 is applied to the remaining years. The fitted models then predict the central years that have been left out. Buffering is used to avoid artificial forecast quality due to hydrometeorological persistence (Michaelsen 1987).

**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 1-month lead time according to

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

(3.2)

where $b = 24 \cdot 60 \cdot 60$ s equals the number of seconds of 1 d and both $\hat{y}$ and $\hat{z}$ have unit m$^3$s$^{-1}$. For the 0-month 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. 3.5.3) since the length of the streamflow series prevents e.g. January 1981, with a lead time of 1 month, from being forecast.

**3.5.2 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 grid points of the ESM involved), the orography, or the degree to which streamflow is regulated. In order that these factors are held while screening through a range of catchment areas, we propose to vary the working scale within a particular target catchment.
Table 3.3: Subcatchment division of the Rhine at Lobith and Basel. The median area covers 4 orders of magnitude.

<table>
<thead>
<tr>
<th>Subcatchments</th>
<th>number of subcatchments</th>
<th>area km²</th>
<th>min</th>
<th>median</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobith level 1</td>
<td>1</td>
<td>-</td>
<td>159700</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Lobith level 2</td>
<td>5</td>
<td>19690</td>
<td>33220</td>
<td>43550</td>
<td></td>
</tr>
<tr>
<td>Lobith level 3</td>
<td>12</td>
<td>8284</td>
<td>13040</td>
<td>17610</td>
<td></td>
</tr>
<tr>
<td>Basel level 1</td>
<td>1</td>
<td>-</td>
<td>36000</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Basel level 2</td>
<td>10</td>
<td>1871</td>
<td>2946</td>
<td>6346</td>
<td></td>
</tr>
<tr>
<td>Basel level 3</td>
<td>124</td>
<td>6</td>
<td>187</td>
<td>2654</td>
<td></td>
</tr>
</tbody>
</table>

Following this line of argumentation we apply the model building procedure from Sect. 3.5.1 to three distinct sets of subcatchments, which we term “spatial levels” (Tab. 3.3). 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 from 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.

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 it 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 are unknown, we only can argue that the results provide a lower bound of the forecast quality due to this methodological constraint.

3.5.3 Validation

The forecast quality of the regression models is analysed using 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 ($\rho$).

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. 3.5.3. Finally, the validation of the 5-day mean streamflow forecasts (Sect. 3.5.3) complements the monthly analysis.

**Benchmarks**

Climatology and runoff simulated by H-TESSEL serve as benchmarks. The climatology is estimated using 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. 3.5.1 is applied.

**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 with 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 $\mathbb{E}(\hat{Z} - Z)$ removed.

**Statistical significance**

In the 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 $r$

$$d = |\hat{z}^r - z| - |\hat{z}^{S4*} - z|$$

(3.3)

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 the 0-month lead time and $r$ being the preMet model are included in the Supplement 3.9.
Skill
To evaluate whether a particular model $m$ has skill with respect to a reference model $r$ the MAE ratio

$$s = 1 - \frac{\text{MAE}_m}{\text{MAE}_r}$$

(3.4)

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%.

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. 3.4) 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, Giorgi, and Coppola 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 (2016), which has a horizontal resolution of 25 m.

Five-day mean streamflow
In order to predict 5-day mean streamflow, Eq. (3.2) is used with a step size of 5 d. However, the monthly dates of prediction 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. 3.5.3 are not valid. The results of the 5-day mean streamflow experiment thus are restricted to a qualitative interpretation.

3.6 Results
The experiment spans several dimensions (i.e. Lobith versus Basel, dates of prediction, lead times, predictor combinations, spatial levels), so we frequently need to collapse one or several dimensions. The Supplement 3.9 aims to complete the results as presented below.
3.6.1 Taylor diagram

Figure 3.2 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 the 0-month 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.

For Lobith and the 0-month lead time we observe an elongated cluster, which comprises all models except the climatology and the refRun model. Some models score a higher correlation – a closer look would reveal that these are the S4P, S4PT, and S4Q models with H-TESSEL standing at the forefront.

3.6.2 Date of prediction versus lead time

Figure 3.3 takes a closer look at the clusters in Fig. 3.2 with the example 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
Figure 3.3: 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 target 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”; $n = 30$.

In general, the patterns repeat more or less along the spatial levels and the S4PT model only beats the reference models in the denominator of Eq. (3.4) at the 0-month lead time. An exception can be observed at Lobith for the month of June, for which the S4PT model most likely outperforms the climatology at the 1-month lead time.

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 1 month. The comparison against the bias-corrected H-TESSEL runoff shows that the S4PT model might provide more accurate
Table 3.4: Mean absolute error at the 0-month lead time of the benchmarks climatology and H-TESSEL and the predictor combinations from Tab. 3.2, rounded to integers. clim: climatology, H-T: H-TESSEL, rR: refRun, pM: preMet, sM: subMet. All values have unit m$^3$s$^{-1}$; $n = 360$.

<table>
<thead>
<tr>
<th></th>
<th>clim</th>
<th>H-T</th>
<th>rR</th>
<th>pM</th>
<th>sM</th>
<th>S4P</th>
<th>S4T</th>
<th>S4PT</th>
<th>S4Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobith</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>level 1</td>
<td>633</td>
<td>421</td>
<td>336</td>
<td>501</td>
<td>502</td>
<td>455</td>
<td>500</td>
<td>453</td>
<td>451</td>
</tr>
<tr>
<td>level 2</td>
<td>633</td>
<td>415</td>
<td>299</td>
<td>484</td>
<td>498</td>
<td>429</td>
<td>482</td>
<td>435</td>
<td>443</td>
</tr>
<tr>
<td>level 3</td>
<td>633</td>
<td>415</td>
<td>285</td>
<td>480</td>
<td>498</td>
<td>424</td>
<td>476</td>
<td>428</td>
<td>444</td>
</tr>
<tr>
<td>Basel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>level 1</td>
<td>239</td>
<td>190</td>
<td>131</td>
<td>197</td>
<td>195</td>
<td>183</td>
<td>195</td>
<td>183</td>
<td>191</td>
</tr>
<tr>
<td>level 2</td>
<td>239</td>
<td>185</td>
<td>115</td>
<td>197</td>
<td>193</td>
<td>179</td>
<td>192</td>
<td>178</td>
<td>187</td>
</tr>
<tr>
<td>level 3</td>
<td>239</td>
<td>182</td>
<td>110</td>
<td>198</td>
<td>193</td>
<td>179</td>
<td>194</td>
<td>178</td>
<td>188</td>
</tr>
</tbody>
</table>

Table 3.5: MAE skill score of the S4* models relative to the preMet model (Eq. 3.4, expressed in %) at the 0-month lead time. The $p$ values for the null hypothesis “the preMet and S4* models score an equal mean absolute error” are enclosed in brackets; $n = 360$.

<table>
<thead>
<tr>
<th></th>
<th>S4P</th>
<th>S4T</th>
<th>S4PT</th>
<th>S4Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobith</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>level 1</td>
<td>9</td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>level 2</td>
<td>11</td>
<td>0</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>level 3</td>
<td>12</td>
<td>0</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Basel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>level 1</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>level 2</td>
<td>9</td>
<td>3</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>level 3</td>
<td>10</td>
<td>2</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

predictions for early summer, but not otherwise.

### 3.6.3 Mean absolute error

In order to conclude the analysis of the monthly predictions at Lobith and Basel, Tab. 3.4 reports the MAE at the 0-month lead time. Reading Tab. 3.4 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, S4PT, and S4Q models lower the MAE by about 40 to 50 m$^3$s$^{-1}$ for Lobith and by about 10 to 20 m$^3$s$^{-1}$ for Basel; and H-TESSEL outperforms the S4* models in the case of Lobith, but not Basel. When reading Tab. 3.4
along the columns, we generally note a decreasing MAE when going from spatial level 1 to spatial level 3 at Lobith. In the case of Basel, the MAE remains more or less constant except for the refRun model.

Focusing on the MOS method, Tab. 3.5 lists the corresponding MAE skill score (Eq. 3.4) 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 12\%. In the case of the S4T model an error reduction is either not existent (Lobith) or small (Basel), which comes along with high \( p \) values.

### 3.6.4 Subcatchments

Figure 3.4 depicts the MAE skill score (Eq. 3.4) for the S4PT model relative to the preMet model for each subcatchment at the 0-month lead time. If
the MAE difference does not exhibit a p value smaller than 0.05 (Eq. 3.3), the subcatchment is coloured in white. We observe that the MAE skill score takes values in the range of about −0.05 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. 3.4 are contrasted in Fig. 3.5 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. 3.5.

While the first two attributes concern the geography of the subcatchments, 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.

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.

### 3.6.5 Five day mean streamflow

Figure 3.6 shows the correlation coefficient of the 5-day mean streamflow observations and corresponding predictions for all models and benchmarks up to a lead time of 45 d. 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 3 weeks; and the subMet model comes close to the refRun model in about 3 weeks.

In addition, we see that the bias-corrected H-TESSEL runoff starts rather cautiously, but it 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 d (Lobith) and 15 d (Basel).

For the full range of lead times, the spatial levels introduce some clear differences (Fig. 3.7): The refRun and subMet models are improved at longer lead times along the spatial levels. For lead times longer than about
Figure 3.5: MAE skill score of the S4PT model with respect to the preMet model for each subcatchment and the 0-month lead time, plotted against subcatchment attributes (see Sect. 3.5.3 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; $n = 360$. 
50 d, the bias-corrected H-TESSEL runoff stays in close harmony with the climatology, while the S4* and preMet models start to score a smaller correlation instead. This effect seems to be mitigated at spatial levels 2 and 3.

### 3.7 Discussion

#### 3.7.1 Model building

In the 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. 3.2). Part of this correlation is also the annual cycle (Fig. 3.1), 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, which 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 5-day mean streamflow the refRun model gets degraded. At short lead times the correlation amounts to about 0.8, while for longer lead
The spatial levels can affect the forecast quality in two ways:

- via the ignorance of the water travel time (Sect. 3.5.2)
- or by the aggregation of the E-OBS and S4 fields at the catchment scale not being 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 5-day streamflow predictions, where at spatial levels 2 and 3 the forecast quality is improved. Using local information of precipitation, surface
air temperature, or runoff appears to compensate for the ignorance of the water travel time.

### 3.7.3 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 only outperforms the climatology at the 0-month lead time; the revESP simulation is outperformed at the 0-month 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 5-day mean streamflow forecasts (Sect. 3.6.5) 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 the case of the 5-day mean streamflow.

### 3.7.4 MOS method

In the case of the monthly mean streamflow forecasts at the 0-month lead time, the MOS method based on precipitation or runoff provides a smaller mean absolute error than the preMet model (Tab. 3.5). Figure 3.6 suggests that this error reduction at the monthly timescale arises from the predictions of the first 15 to 20d. 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 Supplement 3.9). Thus, a rejection of the S4T model does not allow any inference about the forecast quality of surface air temperature itself.

Figure 3.5 indicates that the subcatchment area is most likely 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. 3.4 shows that for small subcatchments in the Alpine region positive skill is rarely 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.

### 3.7.5 H-TESSEL

Within ECMWF’s seasonal forecasting system S4, H-TESSEL aims to provide a lower boundary condition for the simulation of the atmosphere and consequently neither implements streamflow routing nor groundwater storage (Balsamo et al. 2009; ECMWF 2017a). However, H-TESSEL in combination with a linear bias correction often performs best (Tab. 3.4).

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 the case of Lobith (Tab. 3.4). This is most likely related to overfitting, which is not sufficiently smoothed by the model averaging (Sect. 3.5.1).

### 3.8 Conclusion

The present study tests a model output statistics (MOS) method for monthly and 5-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 the 0-month 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 12% compared to the model that is constrained to the initial conditions. For lead times of 1 and 2 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 for 5-day mean streamflow finally 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 timescale, catchments with areas up to $160\,000\,\text{km}^2$, 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 further investigating runoff simulations from land surface components of earth system models for subseasonal to seasonal streamflow forecasting.

### 3.9 Supplement

For the supplementary materials please refer to the online publication or directly download https://doi.org/10.5194/hess-22-929-2018-supplement.
4 An evaluation of model output statistics for monthly streamflow forecasting in Europe


4.1 Abstract

Subseasonal and seasonal forecasts of the atmosphere, oceans, sea ice, or land surfaces often rely on earth system model (ESM) simulations. While the most recent generation of ESMs simulates runoff per land surface grid cell, it does not typically simulate river streamflow. Here, we apply the model output statistics (MOS) method to postprocess ESM forecasts with the goal to predict river streamflow. To do so, the seasonal hindcast archive of the European Centre for Medium-Range Weather Forecasts (ECMWF) is used to test the predictive power of several linear regression models. These models relate observed river streamflow to surface and subsurface runoff, precipitation, and surface air temperature simulated by ECMWF’s forecast systems S4 and SEAS5. In addition, the pool of candidate predictors contains observed precipitation and temperature preceding the date of prediction. The experiment is conducted for 16 European catchments in the period 1981–2006 and focuses on monthly average streamflow at lead times of 0 and 20 d. The results show that a simple linear bias correction of total runoff performs on a level of skill that is hard to beat with more complex model formulations. An exception of this finding is observed for a catchment that features lakes, which extend to about 14 % of the catchment area. On average, the MOS-based forecasts reduce the mean absolute error of the climatology by about 25 % at the 0-day lead time. At the 20-day lead time, skill is present in some rare cases of persistent initial hydrological conditions.

4.2 Introduction

Subseasonal and seasonal forecasts of environmental conditions are increasingly based on numerically coupled models of the various earth system components. These include general circulation models of the atmosphere and
oceans and dynamical land surface or sea ice models (National Academies 2016).

Such global forecast systems represent an increasing number of physical, chemical, and biological processes and continuously progress towards earth system models (ESMs). However, not all environmental variables of interest are resolved. For example, current generation ESM land surface components simulate runoff per grid cell, but in general they do not simulate river streamflow (Clark et al. 2015; Yuan, Wood, and Ma 2015). To the best of our knowledge, ESM runoff simulations have been virtually ignored so far for seasonal streamflow forecasting.

Only recently, Emerton et al. (2018) introduce the GloFAS seasonal forecasting system. This system builds upon the seasonal forecasting capabilities of the European Centre for Medium-Range Weather Forecasts (ECMWF) and feeds runoff simulated by the ESM land surface scheme to the Lisflood model: Subsurface runoff enters a groundwater module and streamflow is routed according to the kinematic wave equations. While this system is run operationally, a comprehensive verification of its forecasts is not published yet.

A complementary approach to predict river streamflow with ESM-based runoff simulations exists in the application of the model output statistics (MOS) method. The MOS method emerged in the context of weather prediction (Glahn and Lowry 1972; Klein and Glahn 1974), where it regressed the variable of interest against the output of a numerical weather model. Today, the MOS method usually has to deal with an ensemble of model integrations. The ensemble accounts for uncertainties regarding the initial conditions and implementation of the physical simulation model (e.g. Schefzik, Thorarinsdottir, and Gneiting 2013).

The MOS method can be broadly understood as the attempt to statistically model the correlation of dynamical forecasts and observations. Besides the prediction of variables not resolved by the dynamical model, the MOS method also can target bias correction (Barnston and Tippett 2017), model combination (Slater, Villarini, and Bradley 2017), and the modelling of the forecast’s probability distribution (Zhao et al. 2017). Often, several of these targets are addressed at the same time.

The MOS method is also sporadically used to predict river streamflow at the subseasonal and seasonal time scales. Early examples include Landman and Goddard (2002) and Foster and Uvo (2010), while more recently the approaches of Sahu et al. (2017), Lehner et al. (2017), and Slater and Villarini (2018) fall within the realm of the MOS method.

In most of these studies, the predictand consists of (sub)seasonal streamflow volumes and the model formulation is based on the assumption of linear predictor-predictand relationships. However, the predictors included in the regression equations vary considerably and include ESM-simulated precipi-
tation, wind velocity, surface air temperature, the geopotential height of atmospheric pressure levels, or time series of land use cover and population density.

Here, we test the application of the MOS method to ESM-based subseasonal forecasts of surface and subsurface runoff. In addition, models are formulated that include precipitation and surface air temperature as predictors. The present implementation of the MOS method relies on the linear regression model and is prototyped in Schick, Rössler, and Weingartner (2018). To mature the prototype we add an error model and conduct a validation in 16 European river systems featuring a range of climatic and geographical conditions.

The hindcast experiment uses data from both the old (S4) as well as current (SEAS5, or S5 in short) seasonal forecast systems of ECMWF for the period 1981–2006. In order to separate the skill originating from the traditional weather forecasting time scale and the potential skill at the subseasonal time scale, the predictand is defined as mean streamflow of a time window of 30 d with lead times of 0 and 20 d.

Below, Sect. 4.3 first introduces the data set, Sect. 4.4 details the MOS method and the hindcast verification, Sect. 4.5 and Sect. 4.6 present and discuss the results, and Sect. 4.7 concludes the study.

4.3 Data

For the following, the hydrometeorological data cover the time period 1981–2006 and have a daily temporal resolution. Spatial fields get aggregated by taking catchment area averages based on the percental grid cell coverage of the catchment polygon. In addition, each grid cell is weighted by the cosine of its latitude to account for the meridional variation of the grid cell area.

4.3.1 Catchments

The catchment selection targets to sample the climatic and geographical conditions across Europe and applies the following criteria:

- Missing values in the streamflow time series are not allowed.
- Large catchments are preferred, as we assume that large areas favour the detection of skill in seasonal climate predictions.
- The number of catchments should be suitable to analyse and visualise the results of individual catchments.

Table 4.1 and Fig. 4.1 show the set of the selected 16 catchments, which includes lowlands and mountainous regions as well as subarctic, temperate,
Table 4.1: Selected catchments and corresponding sites of gauging stations, data providers, catchment areas, and average streamflows in the period 1981–2006.

<table>
<thead>
<tr>
<th>name</th>
<th>site</th>
<th>provider</th>
<th>km²</th>
<th>m³ s⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Angerman</td>
<td>Sollefteå</td>
<td>GRDC</td>
<td>30 900</td>
<td>532</td>
</tr>
<tr>
<td>2 Danube</td>
<td>Bratislava</td>
<td>GRDC</td>
<td>131 400</td>
<td>2072</td>
</tr>
<tr>
<td>3 Duero</td>
<td>Herrera de Duero</td>
<td>MAFFE</td>
<td>12 900</td>
<td>24</td>
</tr>
<tr>
<td>4 Elbe</td>
<td>Neu-Darchau</td>
<td>GRDC</td>
<td>131 100</td>
<td>691</td>
</tr>
<tr>
<td>5 Garonne</td>
<td>Lamagistère</td>
<td>MEST</td>
<td>32 300</td>
<td>386</td>
</tr>
<tr>
<td>6 Glama</td>
<td>Langnes</td>
<td>GRDC</td>
<td>39 900</td>
<td>702</td>
</tr>
<tr>
<td>7 Kemijoki</td>
<td>Isohaara</td>
<td>GRDC</td>
<td>53 900</td>
<td>577</td>
</tr>
<tr>
<td>8 Lower Bann</td>
<td>Movanagher</td>
<td>GRDC</td>
<td>5000</td>
<td>92</td>
</tr>
<tr>
<td>9 Neva</td>
<td>Novosaratovka</td>
<td>GRDC</td>
<td>285 600</td>
<td>2479</td>
</tr>
<tr>
<td>10 Oder</td>
<td>Hohensaaten-Finow</td>
<td>GRDC</td>
<td>110 500</td>
<td>492</td>
</tr>
<tr>
<td>11 Rhine</td>
<td>Lobith</td>
<td>GRDC</td>
<td>159 400</td>
<td>2335</td>
</tr>
<tr>
<td>12 Rhone</td>
<td>Beaucaire</td>
<td>MEST</td>
<td>95 800</td>
<td>1707</td>
</tr>
<tr>
<td>13 Seine</td>
<td>Paris</td>
<td>MEST</td>
<td>43 600</td>
<td>315</td>
</tr>
<tr>
<td>14 Tisza</td>
<td>Senta</td>
<td>GRDC</td>
<td>141 100</td>
<td>795</td>
</tr>
<tr>
<td>15 Torne</td>
<td>Kukkolankoski</td>
<td>GRDC</td>
<td>40 500</td>
<td>423</td>
</tr>
<tr>
<td>16 Trent</td>
<td>Colwick</td>
<td>GRDC</td>
<td>7500</td>
<td>84</td>
</tr>
</tbody>
</table>

These river systems are subject to streamflow regulation of varying degree. For example, Nilsson et al. (2005) classify the streamflow of several of the catchments in Tab. 4.1 as impacted by damming. However, human activities affecting river streamflow can hardly be avoided when it comes to streamflow forecasting in large European river systems. Whether these human activities lead to a pattern that can be learned by the MOS method or instead act as a source of noise will be discussed later in more detail.

4.3.2 Observations

Daily mean streamflow observations in m³ s⁻¹ are provided by the Global Runoff Data Centre (GRDC 2016), the Spanish Ministry of Agriculture and Fisheries, Food and Environment (MAFFE 2017), and the French Ministry for an Ecological and Solidary Transition (MEST 2017). Catchment
Figure 4.1: The hindcast experiment is conducted for 16 catchments situated in Europe. Black crosses on yellow background indicate the sites of the gauging stations, light blue lines show some large rivers, and the numbers refer to the entries in Tab. 4.1. Map produced with data from Natural Earth (2018).

polygons are either retrieved from the GRDC (2016) or derived from the European catchments and Rivers network system (ECRINS 2012).

Daily observations of surface air temperature and precipitation are taken from the ENSEMBLES gridded observational data set in Europe, version 16.0 (E-OBS in short). This data set is based on a statistical interpolation of weather station observations and is available on a 0.25° regular grid (Haylock et al. 2008; E-OBS 2017).

4.3.3 Hindcast archive

ECMWF’s former (S4) and current (S5) seasonal forecast systems consist of fully coupled atmosphere-ocean-land models; in addition, S5 includes the dynamical LIM2 sea ice model. The net horizontal resolution of the atmosphere model equals about 80 km (S4) and 35 km (S5), whereas the NEMO ocean model approximately operates on a 1° (S4) and 0.25° (S5) grid (ECMWF 2017b).

The H-TESSEL land surface model, which is part of both S4 and S5, dynamically divides each grid cell into fractions of bare ground, low and high vegetation, intercepted water, snow, and snow under high vegetation. The partitioning into infiltration and surface runoff happens according to the Arno scheme. Vertical water flow in the soil, which is discretised into
four layers with a total depth of about 3 m, follows the Richards equation. Total runoff finally equals the sum of surface runoff and open drainage at the soil bottom (Balsamo et al. 2009; ECMWF 2018).

For both S4 and S5 the hindcast spans back to 1981 with initial conditions taken from ERA-Interim. Reforecasts are initialised on the first day of each month and simulate the subsequent 7 months. The number of hindcast ensemble members equals 15 (S4) and 25 (S5), respectively. Please note that this describes ECMWF’s standard hindcast configuration, i.e. for certain dates of prediction more ensemble members and a longer lead time are available.

We downloaded the following variables on a regular 0.75° (S4) and 0.4° (S5) grid in daily resolution: Precipitation, air temperature 2 m above ground, total runoff, and surface and subsurface runoff (S5 only). After taking catchment area averages as described above, the ensemble gets compressed to its mean value.

4.4 Method

The predictand $y_{w,l}$ denotes mean streamflow in m$^3$s$^{-1}$ of a time window with length $w = 30$ d and lead time $l = 0, 20$ d. Here, lead time is defined as the time difference between the date of prediction and the onset of the actual prediction window $w$. The date of prediction is set to the first day of the month in the period 1981–2006.

In order to predict with a 20-day lead time, we do not regress $y_{30,20}$, but instead predict $\hat{y}_{20,0}$ and $\hat{y}_{50,0}$, followed by integration in time and taking differences, i.e.

$$\hat{y}_{30,20} = (\hat{y}_{50,0} \cdot 50 - \hat{y}_{20,0} \cdot 20)/30$$

Doing so allows to formulate simple models that separate the variables representing the initial hydrological conditions and meteorological forcings. The rationale behind this procedure will become more clear below and deals with the temporal gap introduced by the lead time $l$. Thus, for the regression we effectively use $w = 20, 30, 50$ d and $l = 0$ d.

4.4.1 Regression model

The modelling procedure is individually applied for each prediction window $w$ and date of prediction within the calendar year, leaving 26 years to perform the regression. Having said this, we drop for the following the subscripts $w$ and $l$. 42
Time aggregation screening

The time aggregation of a particular predictor is defined with respect to the date of prediction and involves summation (e.g. in the case of precipitation) or averaging (e.g. in the case of surface air temperature) in time. The time aggregation period is not fixed in advance, but is individually selected for each predictor based on the linear correlation with $y$. It is constrained to the sets

- $A_{\text{pre}} = \{10, 20, \ldots, 720\} \text{d}$ for predictors that carry information preceding the date of prediction (backward in time), and
- $A_{\text{sub}} = \{5, 10, \ldots, 200\} \text{d}$ for predictors that carry information subsequent to the date of prediction (forward in time).

For the refRun model (explained below) we set $A_{\text{sub}} = \{5, 10, \ldots, w + l\} \text{d}$.

In doing so, the time window of the ESM-based predictors can differ from the actual forecast window. This allows to account for a delayed catchment response to the atmospheric forcings or could help to better detect skillfully predicted climate anomalies.

Predictor combinations

The regression equation is given by

$$y = \psi(x, D) + \varepsilon = x^T \hat{\beta} + \varepsilon$$

(4.2)

with $x^T = [1 \ x_1 \ x_2 \ x_3 \ x_4]$ being the predictor vector and $\beta$ the coefficient vector. Both the time aggregation periods of the entries in $x$ as well as the ordinary least squares estimate of $\beta$ are based on the training set $D$. Please note that we do not make any distributional assumption about the error term $\varepsilon$.

Table 4.2 shows the different predictor combinations that make up $x$. The models named refRun (reference run) and preMet (preceding meteorology) are intended to provide an upper and lower boundary of prediction skill. Conceptually, the precipitation and surface air temperature variables mimic the meteorological forcings; if preceding the date of prediction, they implicitly approximate the initial hydrological conditions (Schick, Rössler, and Weingartner 2018). Thus, the preMet model is constrained to the initial hydrological conditions.

The remaining models contain predictors from the S4 and S5 hindcast archives: Besides precipitation and surface air temperature, we test total runoff as well as surface and subsurface runoff as individual predictors. Please see the Supplement 4.8 for a technical note concerning the S5sro+ssro model.
Table 4.2: The predictor combinations consider the variables $p$: precipitation, $t$: surface air temperature, $ro$: total runoff, $sro$: surface runoff, and $ssro$: subsurface runoff. Predictors get aggregated in time either preceding or subsequent to the date of prediction; the subscripts indicate the data source, i.e. the E-OBS data set and the S4 and S5 hindcast archives.

<table>
<thead>
<tr>
<th>name</th>
<th>preceding</th>
<th>subsequent</th>
</tr>
</thead>
<tbody>
<tr>
<td>refRun</td>
<td>$p_{E-OBS}$</td>
<td>$t_{E-OBS}$</td>
</tr>
<tr>
<td>preMet</td>
<td>$p_{E-OBS}$</td>
<td>$t_{E-OBS}$</td>
</tr>
<tr>
<td>S4PT</td>
<td>$p_{E-OBS}$</td>
<td>$t_{E-OBS}$</td>
</tr>
<tr>
<td>S5PT</td>
<td>$p_{E-OBS}$</td>
<td>$t_{E-OBS}$</td>
</tr>
<tr>
<td>S4ro</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S5ro</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S5ro+ssro</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Bootstrap aggregating

Bootstrap aggregating ("bagging" in machine learning parlance) dates back to Breiman (1996a) and is a technique to reduce model variance. For the present prediction problem and modelling strategy, bagging helps to stabilise model variance as introduced by the small sample size and the sometimes weak relationships (Schick, Rössler, and Weingartner 2016). The “bagged” prediction follows

$$\hat{y} = \frac{1}{b} \sum_{j=1}^{b} \psi(x, D_j)$$  \hspace{1cm} (4.3)

where the subscript $j$ indicates the $j$-th non-parametric bootstrap replicate of $D$. Please note that the number of bootstrap replicates $b$ should not be regarded as a tuning parameter, but is set to a value such that the prediction error stabilises. In order to guarantee the robustness of the analysis we set $b = 100$, which can be considered as rather high (e.g. Breiman 1996a recommends $b \in \{25, \ldots, 50\}$).

4.4.2 Error model

The proposed method to build the error model employs the so called “out-of-bag” prediction error estimate (Breiman 1996b), which avoids an additional cross-validation. In each of the $b$ bootstrap replicates we (most likely) miss some of the cases contained in the full training set. Thus, for the $i$-th case
\[ \hat{\epsilon}_i = y_i - \frac{1}{\sum_{j=1}^{b} \mathbb{1}(y_i \notin D_j)} \sum_{j=1}^{b} \psi(x_i, D_j) \cdot \mathbb{1}(y_i \notin D_j) \]  

with \( \mathbb{1}(\cdot) \) denoting the indicator function that returns one if its argument evaluates to true and zero otherwise. For the 20-day lead time, Eq. 4.4 needs to be adapted according to Eq. 4.1.

Having estimated the prediction error for each case in the training set, we then use a kernel density estimate to specify the probability density function \( f \) of a future prediction \( \hat{y} \)

\[ \hat{f}(y) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K \left( \frac{y - \hat{y} - \hat{\epsilon}_i}{h} \right) \]  

with \( n \) being the sample size of the training set \( D \) and the kernel \( K(z) \) the standard Gaussian density function

\[ K(z) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} z^2 \right) \]  

The bandwidth parameter \( h \) is automatically selected according to the method of Sheather and Jones (1991) as implemented in the statistical software R (R Core Team 2018). This method seems to work well for a variety of density shapes as it uses a separate pilot estimate for the second derivative of the unknown density function (Jones, Marron, and Sheather 1996).

### 4.4.3 Hindcast verification

The modelling procedure outlined in Sect. 4.4.1 and 4.4.2 is subject to a buffered leave-one-out scheme. A buffer of 2 years to the right and left of the left-out year is used in order to avoid artificial skill due to hydrometeorological persistence (Michaelsen 1987).

**Reliability**

Reliable forecast distributions reproduce the observations’ frequency, i.e. the forecast distribution is neither too narrow (overconfident) nor too wide (underconfident). Here, we follow Laio and Tamea (2007), who propose to evaluate the probability integral transform (PIT) values graphically via the empirical cumulative distribution function. The PIT value of a forecasted cumulative distribution function \( \hat{F}_i(y) \) and corresponding observation \( y_i \) is defined as the probability

\[ \text{PIT} = \hat{F}_i(y_i) \]
If \( y \) is continuous and the forecasts are reliable, the PIT values follow the uniform distribution \( U(0, 1) \).

**Scoring rules**

To verify the accuracy of the predictions, three scoring rules are used, namely the mean absolute error (MAE), the mean squared error (MSE), and the continuous ranked probability score (CRPS), the latter being defined as

\[
\text{CRPS} = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left( \hat{F}_i(y) - 1(y \geq y_i) \right)^2 dy
\]

with \( n \) denoting the sample size.

**Skill**

Climatology (i.e. mean streamflow of the predictand’s time window in the period 1981-2006, calculated for each date of prediction within the calendar year) and the preMet model from Tab. 4.2 are used as benchmark models. In addition, we also use total runoff predicted by the S4 and S5 systems in combination with a linear bias correction, named S4ro.lm and S5ro.lm in the following (lm: linear model).

The S4ro.lm and S5ro.lm benchmarks can be considered as simpler versions of the S4ro and S5ro models (Tab. 4.2) in that they neither employ Eq. 4.1 nor the time aggregation screening and bagging; this resembles the approach of Balsamo et al. (2015) to verify the ERA-Interim/Land simulation with respect to river streamflow observations.

Identical to the other models, the benchmark models undergo the cross-validation and the forecast distribution is based on the kernel density estimate from Eq. 4.5. In the case of the climatology and the S4ro.lm and S5ro.lm models, the residuals \( \varepsilon_i \) are the in-sample prediction errors of the training set.

Having a model of interest \( m_1 \) and benchmark model \( m_2 \), the mean absolute error skill score (MAESS) is then defined as

\[
\text{MAESS} = 1 - \frac{\text{MAE}^{m_1}}{\text{MAE}^{m_2}}
\]

The mean squared error skill score (MSESS) and the continuous ranked probability skill score (CRPSS) are defined analogously.

**Statistical significance**

Statistical tests are conducted conditional on the date of prediction within the calendar year (thus \( n = 26 \)): 
• To test the PIT values for uniformity, we report the number of null hypothesis rejections of the Pearson’s chi-squared test using four, five, and six equally sized bins. In addition, we use confidence bands based on the Kolmogorov-Smirnov test. Since in both cases the null hypothesis assumes a uniform distribution, we set the significance level to 0.25 in order to have more control on the type II error (that is not rejecting the null hypothesis when it is in fact false). The value of the Kolmogorov-Smirnov test statistic at the 0.25 level is taken from D’Agostino and Stephens (1986).

• In order to test whether a model $m_1$ and a benchmark $m_2$ differ in terms of the CRPS, we use paired differences of the individual CRPS values. The null hypothesis “the mean difference equals zero” is then tested with the two-sided $t$-test. It must be noted that the paired differences do not always follow a Gaussian distribution. However, a comparison with a non-parametric bootstrap and the Wilcoxon test showed that the $t$-test leads to the most conservative results, for what reason we only report the $p$ values of the $t$-test.

### 4.5 Results

First, a general overview is presented that focuses on the complete time series of observations and predictions. Second, the reliability of the forecast probability distribution is validated. Third, the CRPSS is analysed conditional on the date of prediction within the calendar year. Finally, some examples are highlighted to enable a better comparison with existing studies.

#### 4.5.1 Overview

Figure 4.2 shows the overall MAESS and MSESS with the climatology as benchmark per catchment and model. In addition, Pearson’s correlation between predictions and observations is shown to the right. Here, we do not test for statistical significance, but aim to summarise the hindcast results.

We observe that the correlation of the climatology can go up to about 0.8 with the median being around 0.6, showing that several streamflow time series exhibit a pronounced periodic component. The refRun model decreases on average the MAE of the climatology by about 35% (i.e MAESS of 0.35) and the MSE of the climatology by about 55% (i.e. MSESS of 0.55). While the preMet model beats climatology at the 0-day lead time, it does so at the 20-day lead time only for a few catchments.
Figure 4.2: MAESS, MSESS, and Pearson’s correlation between predictions and observations per model and catchment. Top row: 0-day lead time, bottom row: 20-day lead time. In the case of MAESS and MSESS the benchmark model is climatology; \( n = 16 \).

On average, the models containing the ESM-based predictors reduce the MAE (MSE) of the climatology by about 25\% (40\%) at the 0-day lead time. At the 20-day lead time, the corresponding reduction amounts to 9\% (15\%). The few outliers belong either to the Oder (negative outliers at the 0-day lead time), Neva (positive outliers at the 20-day lead time), and Lower Bann or Duero (negative outliers at the 20-day lead time).

Concerning differences in skill between the predictor combinations, it seems that the runoff-based models outperform the S4PT and S5PT models. This holds in particular at the 20-day lead time, where the S4PT and S5PT models are hardly distinguishable from the preMet model. The S4ro and S4ro.lm models score on average the same skill as their S5 counterparts, but the MAESS and MSESS values show a slightly larger spread.

In the following, we focus on the S5-based models since S5 is the current operational system and differences in Fig. 4.2 between S4 and S5 are small, if present at all. However, in order to compare the skill of the present MOS method with other studies, we also pick up later some results of the S4-based models.
4.5.2 Reliability

Figure 4.3 shows the empirical cumulative distribution function of the PIT values at the 0-day lead time, accompanied by the Kolmogorov confidence bands at the 0.25 significance level. The numbers in the top left corner report the number of rejected null hypotheses of the chi-squared test based on four, five, and six bins, again at the 0.25 level.

We observe a few erratic deviations from the 1:1 diagonal and PIT values larger than 0.5 tend to be above the bisector. The latter indicates underconfidence in the upper tail of the forecast distribution, i.e. the right tail of the distribution is too heavy. Otherwise no consistent pattern can be found.

Concerning the statistical significance, the PIT values are almost never outside the Kolmogorov confidence band. The chi-squared test rejects on average about one out of 16 distributions from being reliable. For the 20-day lead time (not shown), the overall picture remains the same.

4.5.3 CRPSS

Figure 4.4 shows conditional on the date of prediction within the calendar year the CRPSS with the climatology as benchmark. If the paired differences in the CRPS can be assumed to differ from zero according to the \( t \)-test, a large (small) cross is drawn in the case of the 0.01 (0.05) significance level. The top row corresponds to the 0-day lead time and the bottom row to the 20-day lead time. Figure 4.7 in the Supplement 4.8 completes Fig. 4.4 for the refRun, preMet, S4PT, and S4ro models.

While we observe in most cases positive skill at the 0-day lead time, the statistical significance is frequently absent. Significant positive skill tends to cluster in spring, however, a clear overall pattern does not emerge. Instead skill varies between catchments, dates of prediction, and models. For some catchments we observe a pronounced seasonality: For example, the runoff-based models score well in autumn and winter for the Oder, whereas for the Rhine they are most skillful in spring and early summer.

At the 20-day lead time (bottom row in Fig. 4.4) skill gets drastically reduced. Exceptions are the Neva, for which the S5PT model still scores positive skill, and the Angerman and the Oder, where the runoff-based models get not so much degraded compared to the 0-day lead time. Otherwise, the sparse positive skill tends again to loosely cluster in spring.

Figure 4.5 equals Fig. 4.4, but employs the preMet model as benchmark to calculate the CRPSS. Thus, for the S5PT model skill solely originates from the S5-predicted precipitation and surface air temperature.

Starting with the 0-day lead time (top row), we observe for the S5PT model some positive skill scattered among the catchments and dates of
Figure 4.3: Empirical cumulative distribution of the PIT values \((n = 26)\) obtained for the 0-day lead time and each catchment. The number of null hypothesis rejections of the chi-squared test are reported in the top left corner (corresponding to four, five, and six bins at the 0.25 significance level). The dashed red lines indicate the Kolmogorov 0.25 confidence band.
**Figure 4.4:** CRPSS at the 0-day (top row) and 20-day lead time (bottom row) with the climatology as benchmark. The months refer to the date of prediction. $p$ values smaller than 0.01 (0.05) for the null hypothesis “no difference in the mean CRPS value” are indicated with a large (small) cross; $n = 26$.

**Figure 4.5:** As Fig. 4.4, but with the preMet model as benchmark.
prediction. The runoff-based models in general do a better job, which is most evident for the Rhine. On the other hand, these models score some large negative CRPSS values. At the 20-day lead time, skill of the S5PT model virtually drops to zero, while the runoff-based models are still able to outperform the preMet model, most notably in the case of the Oder.

4.5.4 Selected examples

In the following we have a closer look at three specific examples. These examples are selected to enable a comparison with recent studies that employ the S4 hindcast archive for seasonal streamflow forecasting in the European domain.

Figure 4.6 shows for the S4-based models conditional on the date of prediction within the calendar year: the linear correlation of predictions and corresponding observations (top panel); the CRPSS with the preMet model as benchmark (middle panel); and for the Elbe River (catchment number 4 in Tab. 4.1 and Fig. 4.1) the MSESS with respect to the climatology as well as the CRPSS with respect to the preMet model (bottom panel). All results refer to the 0-day lead time.

The correlation of the S4-based predictions scatters around 0.6 with rather small seasonal variations (top panel). The S4PT model tends to score the lowest correlation, which is most pronounced for March to June. Otherwise, differences between the models mostly are small.

Regarding the CRPSS with respect to the preMet model (middle panel), we observe that the median CRPSS of the S4PT model stays close to zero. In addition, the S4PT model is often outperformed by the S4ro and S4ro.lm models. Overall, the pattern throughout the calendar year resembles the pattern obtained for the correlation in the top panel.

For the Elbe River (bottom panel), we observe a marked drop of the MSESS in August, which is more pronounced for the preMet model; otherwise the MSESS varies around 0.5. Regarding the CRPSS, the S4ro and S4ro.lm models tend to positive skill at the beginning and end of the calendar year. The CRPSS of the S4PT model instead is close to zero.

4.6 Discussion

The discussion starts with some technical aspects regarding the regression and error models. Subsequently, we discuss the different predictor combinations and compare the present hindcast results with results reported in other studies. Finally, we discuss the potential effects of streamflow regulation and the sources of forecast skill.
Figure 4.6: Correlation of predictions and corresponding observations, CRPSS, and MSESS of the S4-based models conditional on the date of prediction within the calendar year. All scores belong to the 0-day lead time and the benchmark model is noted in the legend. The top and middle panels refer to all catchments \((n = 16)\), whereas the bottom panel focuses on the Elbe River (catchment number 4 in Tab. 4.1 and Fig. 4.1).
4.6.1 Regression model

Temporal resolution

The MOS method aims to model the correlation between dynamical forecasts and observations of the target variable. Apart from long term trends and seasonal patterns, this correlation emerges at the (sub)seasonal time scale only at a low temporal resolution, if present at all (Troccoli 2010). The MOS method thus depends on a suitable time averaging applied to the involved variables and inevitably operates at a low temporal resolution.

Time aggregation screening

The S4ro.lm and S5ro.lm benchmark models do not apply a time aggregation screening, but instead regress the predictand against total runoff based on the same time window. The results show that these benchmarks compete well against their counterparts (i.e. S4ro, S5ro, and S5sro+ssro). Thus, for the predictors that carry the runoff simulations the additional effort of the time aggregation screening only leads to small improvements (Fig. 4.2).

Linearity

The model formulation strictly assumes a linear relationship between the predictors and the predictand. From both an empirical as well as theoretical point of view, the assumption of linearity gains validity with an increasing time aggregation window length (Yuval and Hsieh 2002; Hsieh et al. 2003).

However, the residual analysis (not shown) reveals that low flows tend to be overpredicted and high flows tend to be underpredicted, often leading to skewed residual distributions. In addition, the pooled time series of the residuals sometimes exhibit autocorrelation. These issues could be related to missing predictors or imply that the time averaging windows of 20, 30, and 50 d is too short to completely linearise the predictor-predictand relationship.

While it is not surprising to observe model deficiencies for the preMet model or the models containing the ESM-based predictors, they are also present for the refRun model, though to a smaller degree. Experimentation with interactions, higher order terms, and transformations of the predictand (not shown) revealed that overfitting rapidly becomes an issue. Extrapolation beyond the domain covered by the training set then sometimes leads to disastrous predictions.
4.6.2 Error model

While the kernel density estimator is able to deal with skewed residual distributions, it assumes otherwise independent and identically distributed errors. The verification in terms of the PIT diagram (Fig. 4.3) shows a few erratic departures from the 1:1 diagonal. In addition, the error model tends to be underconfident in the upper part of the distribution.

Given the model misspecifications reported above, the cross-validation in combination with a rather small sample size, and the conservative significance level, we judge the reliability of the forecast probability distribution as reasonable.

4.6.3 Predictor combinations

The runoff-based models tend to outperform the models containing precipitation and surface air temperature (Fig. 4.2). The fact that the S4PT and S5PT models theoretically work on an acceptable level in many instances (i.e. the refRun model in Fig. 4.2 and 4.7) further underlines the predictive skill of the runoff-based models.

A notable exception that contrasts the runoff-based models with the models based on meteorological predictors is provided by the Oder and the Neva River: For the Oder, the models based on meteorological predictors completely fail, but the runoff-based models score above average, and vice versa for the Neva (Fig. 4.4). These two cases are shortly discussed now.

The Oder catchment

The Oder catchment differs from the other catchments particularly in one feature: According to the International Hydrogeological Map of Europe (IHME 2014) the lithology of the Oder catchment is dominated by coarse and fine sediments and the aquifer productivity is classified as low to moderate for nearly the entire catchment. In addition, the runoff efficiency (streamflow divided by precipitation, equals for the Oder about 0.28) and total annual precipitation (about 500 mm) belong to the lowest values contained in the present set of catchments.

The combination of high evapotranspiration and the presumably low contribution of groundwater to streamflow might imply that the soil is the controlling factor for the generation of streamflow. If so, the model formulation based on the meteorological predictors is too simplistic to account for temporal variations of the soil moisture content.
The Neva catchment

The preMet and refRun models score similar for the Neva catchment both at the 0-day as well as the 20-day lead time (Fig. 4.7). This indicates that the initial hydrological conditions strongly control the generation of streamflow. While the S4 and S5 runoff simulations carry the information of the soil moisture content and snow pack at the date of prediction, preceding precipitation and temperature aim to account for the sum of all hydrological storages.

The Neva differs from the other catchments in the presence of several large lakes (e.g. the Lake Ladoga and the Lake Onega, see Fig. 4.1). According to the Global Lakes and Wetlands Database (GLWD, Lehner and Döll 2004), about 14% (39 000 km$^2$) of the catchment area is covered by lakes. Thus, we speculate that H-TESSEL-based runoff is not a suitable predictor if lakes represent a substantial fraction of the catchment area. This would be in line with the study of Huziy and Sushama (2017), who increase the accuracy of streamflow simulations by including a lake-river routing model into the CLASS land surface scheme.

4.6.4 Selected Examples

Below, we put the results from Sect. 4.5.4 into context with the results from other studies. In these studies hydrological models are forced with seasonal climate predictions in the European domain. While the hindcast configurations of these studies do not match exactly those of the present study, it should still be valid to compare the magnitude and seasonal variations of skill.

The below referenced studies frequently use the Ensemble Streamflow Prediction framework (ESP, Wood and Lettenmaier 2008) for benchmarking. ESP model runs derive predictive skill exclusively from the initial hydrological conditions, what conceptually corresponds to the preMet model of the present study.

Greuell et al. (2018) use the S4 hindcast archive in combination with the VIC model. For monthly mean streamflow forecasts validated against observations of about 700 gauging stations, they report on average a correlation between 0.6 and 0.7 at the 0-day lead time. This is reproduced by the S4ro and S4ro.lm models (top panel in Fig. 4.6).

In Arnal et al. (2018) and Wetterhall and Di Giuseppe (2018), the Lisflood model is forced with the output from ECMWF’s S4 and ENS-Extended systems. In terms of the CRPSS, the ESP run is outperformed on average within the first month, but not beyond. For monthly mean streamflow at the 0-day lead time, the median CRPSS reaches in winter its maximum at about 0.2 (Arnal et al. 2018). Thus, the present study agrees
with the skillful lead time, but fails to reproduce a median CRPSS in the range of 0.2 (S4PT model in the middle panel of Fig. 4.6).

Monthly mean streamflow of the Elbe River at Neu-Darchau is predicted in Meißner, Klein, and Ionita (2017) with the LARSIM model and the S4 hindcast archive. At the 0-day lead time, the MSESS with respect to climatology of the ESP run is for most months in the range of 0.4 to 0.7; for August, the MSESS is close to zero. Thus, both the magnitude and seasonal variations are approximately reproduced by the preMet model (bottom panel of Fig. 4.6).

Benchmarking the LARSIM-S4 run with the ESP run in terms of the CRPS leads to a CRPSS of 0.16 in May and a CRPSS of 0.22 for June; otherwise the CRPSS stays close to zero at the 0-day lead time. In the present study, such high values for May and June are not reproduced (S4PT model in the bottom panel of Fig. 4.6).

In summary, the skill obtained in the present study mostly corresponds to the skill reported in the above referenced studies. Differences mainly concern the skill estimates that use the ESP run (here: preMet model) as benchmark. Since such skill estimates are aimed to number the added skill when using seasonal climate predictions, they could be overly sensitive to the underlying study period.

4.6.5 Streamflow regulation

As already noted, the streamflow time series may contain numerous anthropogenic artefacts introduced by e.g. damming and regulation, water consumption, and diversions. While the temporal aggregation most likely cancels some of these anthropogenic signals, the potentially remaining human “noise” ends up in the predictand. Subsequently, it is theoretically possible that the MOS method learns anthropogenic patterns in the streamflow series.

A visual inspection of the daily streamflow series (not shown) reveals that obvious anthropogenic artefacts are mainly present for the Angerman, Glama, and Kemijoki Rivers. For these catchments the time series show some rectangular-like fluctuations at a frequency of a few days, most likely induced by streamflow regulation and hydro power production. However, the refRun model, which is aimed at estimating the potential skill, performs poorly mainly for the Duero and Oder River (Fig. 4.7). This indicates that human “noise” per se is not a problem.

4.6.6 Sources of skill

Skill with respect to climatology restricts for most catchments and dates of prediction to the first month ahead (Fig. 4.2, 4.4, and 4.7). At the 20-day
lead time the remaining skill tends to roughly cluster in spring. This most likely can be attributed to snow accumulation in winter and subsequent melting in spring. The high skill observed for the Neva might be enabled by the presence of several large lakes.

In any case, the results indicate that skill originates mostly from the initial hydrological conditions rather than from the predictions of precipitation and surface air temperature (S5PT model in Fig. 4.5 and S4PT model in the middle panel of Fig. 4.6). The initial conditions relevant for (sub)seasonal streamflow forecasting include hydrological storages such as soils, aquifers, surface water bodies, and snow (e.g. Dijk et al. 2013, Shukla et al. 2013, or Yossef et al. 2013).

The rather low contribution of ESM-simulated precipitation and surface air temperature to streamflow forecast skill is not surprising. (Sub)seasonal climate predictions show limited skill for the European continent: Besides the prediction of long term trends, some skill is on average present within the first month ahead, but not beyond (Slater, Villarini, and Bradley 2017; Monhart et al. 2018; Rodrigues, Doblas-Reyes, and Coelho 2018).

4.7 Conclusion

Dynamical earth system models (ESMs) used today for subseasonal and seasonal forecasting of environmental conditions in general simulate runoff at the surface and at the bottom of the soil column. River streamflow, however, remains an unresolved variable and requires an additional modelling effort to forecast. The present study does so by an application of the model output statistics (MOS) method.

The test bed of this MOS application consists of 16 European catchments and monthly average streamflow at the 0-day and 20-day lead time in the period 1981–2006. Input to the MOS method is provided by the seasonal hindcast archive of the European Centre for Medium-Range Weather Forecasts (ECMWF). Hindcasts of both the S4 and SEAS5 forecast systems are employed.

The present implementation of the MOS method tries to establish a statistical link between streamflow observations and variables simulated by S4 and SEAS5, namely surface and subsurface runoff, precipitation, surface air temperature, and combinations thereof. In addition, the pool of candidate predictors contains observed precipitation and temperature preceding the date of prediction. Technically, the approach relies on the linear regression model in combination with a time aggregation screening and bootstrap aggregating. Probabilistic forecasts are obtained by modelling the residuals with a kernel density estimator.

At the 0-day lead time the MOS method decreases the mean absolute
error of the climatology by about 25% on average; at the 20-day lead time, the decrease drops to about 9%. This result holds for both the S4 and SEAS5 forecast systems.

However, skill varies considerably between models, catchments, and dates of prediction within the calendar year. In addition, skill is also frequently absent, especially at the 20-day lead time. The results further indicate that skill originates mainly from the initial hydrological conditions. In particular, skill at the 20-day lead time is most likely enabled by snow accumulation in winter and subsequent melting in spring or by the presence of large lakes.

The results show that predictor combinations comprising runoff simulations often perform best. Even an univariate regression of monthly streamflow against total runoff scores on a level close to the more complex models. Additional predictors seem to be required if the catchment features lakes whose areas are large compared to the catchment area.

A comparison against studies that force hydrological models with seasonal climate predictions in the European domain shows that skill estimates often agree. Differences mainly concern the estimates of skill provided by seasonal climate predictions, which in the present study probably are too low.

Runoff simulated by the land surface component of ESMs provides a yet virtually untapped source for (sub)seasonal river streamflow forecasting. To advance river streamflow forecasting at the subseasonal and seasonal time scale, it could be worth to further investigate ESM-based runoff simulations, be it in combination with the MOS method or other approaches.

### 4.8 Supplement

#### 4.8.1 Technical note

After aggregation in time, winter surface runoff \( \text{sro} \), contained in the S5sro+ssro model) can include years with zero and near-zero values as well as years with larger values. This is in particular the case for the Angerman, Kemijoki, and Torne catchments. Selecting in the bootstrap by chance only years with zero and near-zero values results in large regression coefficients and subsequently leads to disastrous overpredictions when applied to the out-of-sample cases.

As an empirical rule, we set all surface runoff values (after aggregation in time) smaller than 1 m\(^3\)s\(^{-1}\) to 0 m\(^3\)s\(^{-1}\). These 0 m\(^3\)s\(^{-1}\) surface runoff values frequently introduce singular covariance matrices. The QR-algorithm implemented in the statistical software R (R Core Team 2018) sets either of the coefficients of collinear variables to zero, which we retain for the
regression model.

4.8.2 Additional figures

Figure 4.7: CRPSS at the 0-day (top row) and 20-day lead time (bottom row) with the climatology as benchmark. The months refer to the date of prediction. $p$ values smaller than 0.01 (0.05) for the null hypothesis “no difference in the mean CRPS value” are indicated with a large (small) cross; $n = 26$. 
5 Synthesis

The synthesis is organised in four parts: Firstly, some methodological remarks are made. These remarks concern issues that frequently show up in Sect. 2 to 4, namely artificial skill, anthropogenic signals in the streamflow series, and the pros and cons of the MOS method. Secondly, the main findings are summarised. Thirdly, the conclusion is presented and fourthly, some potential directions for future research are suggested.

5.1 Methodological remarks

5.1.1 Artificial skill

All three hindcast experiments base on a buffered leave-one-out validation (Michaelsen 1987), spanning approximately the period 1981–2010. Besides the variance of the cross-validation estimator, two potential sources of artificial skill exist:

- Weisheimer et al. (2017) show that the seasonal predictability of the atmosphere over the Euro-Atlantic sector varies considerably in time. For example the period 1981–2010 is attributed an above average predictability of the winter North Atlantic Oscillation.

- Pronounced trends in the streamflow time series can be misused by the MOS method in that it selects predictors showing a trend, too. If the benchmark model is not able to account for potential trends (e.g. climatology), skill gets artificially inflated whenever trends in the predictand and predictors are not physically related.

While the first source must be accepted, the second source could be eliminated by applying a detrending to both the predictand as well as the predictors. A detrending, however, is not applied in Sect. 2 to 4 for the following two reasons:

1. The number of variables requires an automatic detrending procedure, which is not a trivial task. If not applied carefully, a detrending can introduce artefacts.

2. I believe that the pool of candidate predictors is physically plausible and a reasonable model should reproduce the low frequency variability in the predictand.
5.1.2 Anthropogenic effects

The streamflow observations used in Sect. 2 to 4 contain to varying degrees anthropogenic signals, which result from e.g. reservoir operation, hydro power production, or water consumption. The temporal averaging most likely smooths out parts of these anthropogenic signals, especially for long time aggregation periods. The remaining signals can either act as a source of noise or result in a pattern that is amenable to statistical learning.

In Sect. 2, a catchment with outstanding poor skill turns out to be heavily regulated. For the set of catchments present in Sect. 4, however, it is argued that anthropogenic effects do not necessarily degrade forecast skill. Therefore, the relationship between anthropogenic effects and predictive skill of the MOS method remains unclear, though it can be stated that an application of the MOS method does not a priori exclude catchments with strong anthropogenic impacts.

5.1.3 Pros and cons of the MOS method

The MOS method features some generic advantages and disadvantages. Some of these are inherent to the data-driven approach, others are specific to the present prediction problem (see also Sect. 3.7 and 4.6).

Advantages include:

- The ESM simulations do not need to be bias corrected.
- The predictor-predictand mapping can theoretically bridge different spatial scales or implicitly account for anthropogenic effects in the streamflow time series.
- Let aside overfitted models, the MOS method should in principal fall back to climatology if the predictors are not correlated with the observations (Zhao et al. 2017).
- Compared to forecast approaches that use the ESM output to force hydrological simulation models, the MOS method could save computational costs.

Disadvantages include:

- The temporal resolution of the predictand inevitably is low (Sect. 4.6.1).
- Predictions at locations along the river network without observations are not feasible.
- The model fitting needs a sufficiently large training set, what makes it impossible to rapidly integrate new observational data sources.
• The model fitting requires past forecasts of the involved dynamical model.

5.2 Main findings

5.2.1 Assumption of linearity

On the one hand, the assumption of linear predictor-predictand relationships is supported by the time averaging, which starts to linearise the predictor-predictand relationships (Yuval and Hsieh 2002; Hsieh et al. 2003). On the other hand, the assumption of linearity is a practical constraint as the present sample size makes it hard to model potential nonlinearities (Sect. 4.6.1). The results show that the assumption of linearity is valid as a first approximation if the time averaging window of the predictand spans at least about 20 d (Sect. 3 and 4).

5.2.2 Model variance

The model fitting conditions on the date of prediction within the calendar year to account for seasonally varying model parameters. This reduces the sample size, which in turn controls the feasible model complexity. The combination of small sample sizes and noisy predictors provides the motivation to test bootstrap aggregating.

For the present regression problem, it is estimated that bootstrap aggregating decreases on average the squared error of prediction by about 7% (Sect. 2). However, bootstrap aggregating can fail to smooth out model variance when interactions or higher order terms are included in the regression equation (Sect. 4.6.1).

5.2.3 Spatial scales

Predictors are defined as catchment area averages and consequently any spatial information is ignored. In Sect. 3 the MOS method is tested with sets of catchments ranging in median area from about 200 to 160,000 km². For this range of catchment areas effects of a spatial scale mismatch are not observed. Thus, within its own limitations, the regression succeeds for catchments falling below the grid scale as well as catchments covering large arrays of grid cells.
5.2.4 Predictor combinations

Sections 3 and 4 indicate that the most simple approach, i.e. a linear regression of observed streamflow against ESM-simulated runoff, is hard to beat with more complex formulations of the MOS method. An exception to this rule most likely are catchments that feature lakes, which constitute a substantial fraction of the catchment area. In this case additional predictors need to be included in the regression equation (Sect. 4).

The ESM runoff simulations used in Sect. 3 and 4 base on ECMWF’s H-TESSEL land surface model. Major hydrological processes not implemented in H-TESSEL include groundwater dynamics and the river and lake routing. While it remains unclear to which degree the regression of observed streamflow against simulated runoff mimics the river routing and groundwater dynamics, it is most likely not able to account for the lake routing (Sect. 4).

5.2.5 Predictive skill

In the hindcast experiments of Sect. 2 to 4, only statistical models are used to benchmark the MOS method. In order to better classify the forecast quality of the MOS method, some results are compared with studies that use hydrological simulation models in similar hindcast configurations. The comparison covers only a subset of the present hindcast results, however, the MOS method generally reproduces the results reported in the referenced studies (Sect. 3.7.3 and 4.6.4).

Overall, with respect to forecasts of monthly mean streamflow, verified as a continuous variable, and the present set of European catchments, the main findings are:

- Skill against climatology is on average restricted to the first month ahead and corresponds in the case of the mean absolute error to a reduction of about 25%. Conditional on the date of prediction within the calendar year, however, skill is also frequently absent. Longer lead times seem to be feasible if the catchment features large lakes or a pronounced snow melting in spring (Sect. 3 and 4).

- Skill can mostly be attributed to the initial hydrological conditions (Sect. 3 and 4).

- Besides the initial hydrological conditions, ESM-based predictions of precipitation (and eventually surface air temperature) improve the streamflow forecast accuracy within the first month ahead (Sect. 3 and 4). In the case of the Rhine catchment and the period 1981–2011, it is estimated that the integration of seasonal climate predictions
reduces the mean absolute error in the range of 5 to 12% compared to forecasts that are based on the initial hydrological conditions only (Sect. 3).

5.3 Conclusion

ESM-based runoff simulations only recently appear in the scientific literature dealing with seasonal streamflow forecasting (Sect. 4.2). Thus, the potential of ESM-based runoff simulations for seasonal streamflow forecasting is largely unexplored, be it in combination with the MOS method or other approaches such as river routing algorithms.

The present application of the MOS method focuses on ECMWF’s seasonal hindcast archive and European catchments. Furthermore, only a small number of candidate predictors are considered. Within this setting, MOS-processed ESM simulations seem to provide seasonal streamflow forecasts on a level of skill similar to hydrological models forced with seasonal climate predictions.

Given the main findings above and the pros and cons listed in Sect. 5.1.3, I argue that the MOS method could be useful in an operational context by complementing existing approaches: The MOS method suffers from other uncertainties than e.g. hydrological models forced by seasonal climate predictions, which is beneficial in the context of multi model approaches.

Furthermore, the application of the MOS method requires only a few parameters to be estimated and is computationally cheap since the ESM simulations are performed by organisations like ECMWF. An operational implementation therefore is straightforward and mainly concerns a stable data flow. This, however, comes at the cost of predictions that are constrained to a low temporal resolution and observing sites along the river network.

5.4 Outlook

To further investigate the potentials and limitations of MOS-processed ESM runoff simulations at the seasonal time scale, directions for future research could include:

- A comprehensive benchmarking of the MOS method with other approaches used in seasonal streamflow forecasting. This should focus on how and why the MOS method can complement these approaches.

- Testing the MOS method with data sets not used in the present thesis. This could include catchments outside of Europe or seasonal forecast
data bases such as the NMME (Kirtman et al. 2014) or S2S (Vitart et al. 2017), which contain other ESMs than ECMWF’s S4 and SEAS5 systems.

- Testing predictors that explicitly represent hydrological storages not implemented in ESM land surface schemes (e.g. glacier volumes or lake and groundwater levels). This could benefit from using ESM reanalysis data sets in addition to seasonal hindcast archives.

- Applying the MOS method to catchments whose hydrological processes are heavily controlled by (at best known) human activities. This could also include – as proposed in the work of Slater and Villarini (2018) – the consideration of predictors representing human activities, e.g. time series of land use statistics.

- Refining the statistical implementation. This could focus on how to formulate a “global” model instead of separately fitting models for each season, which would enlarge the size of the training set.
6 Appendix

6.1 Data sources

I gratefully acknowledge the following providers of data:

- the State Institute for the Environment, Measurements and Conservation Baden Wuerttemberg,
- the Bavarian Environmental Agency,
- the German Federal Institute for Geosciences and Natural Resources,
- the State of Vorarlberg,
- the Austrian Federal Ministry of Agriculture, Forestry, Environment and Water,
- the Swiss Federal Office for the Environment,
- the French Ministry for an Ecological and Solidary Transition,
- the Spanish Ministry of Agriculture and Fisheries, Food and Environment,
- the Global Runoff Data Centre at Koblenz,
- the EU-FP6 project ENSEMBLES and ECA&D project,
- the Copernicus data service and European Environment Agency,
- the European Centre for Medium-Range Weather Forecasts,
- the Natural Earth map data repository, and
- the World Wildlife Fund.

In addition, Bernhard Wehren provided streamflow data for the Kander River at Hondrich.
6.2 Comparison of OOB and LOO

The comparison of the prediction error estimate of the leave-one-out (LOO) and out-of-bag (OOB) procedures is suggested in Hastie, Tibshirani, and Friedman (2009, Exercise 15.2). The question is whether these estimates converge for an increasing number of bootstrap replicates.\footnote{The present comparison is based on suggestions by and discussions with Tobias Fissler. However, the formulation and the correctness of the result are within my own responsibility.}

**Notation:** Given a set $D$ consisting of $n$ independent and identically distributed (iid) random vectors $Z = [Y, X]^T$

$$D = \{Z_1, \ldots, Z_n\} \tag{6.1}$$

then $D_{-i}$ specifies the set $D$ with $Z_i$ removed. Furthermore $D_j$ is the $j$-th replicate of $D$ when sampling $n$ times with replacement (i.e. the $j$-th non-parametric bootstrap replicate). Please note that $D_{-i}$ amounts to sampling $n - 1$ times with replacement.

In order to predict $Y$ by means of $X$, bagging employs $b$ replicates of $D$. For each replicate one model is fitted and the resulting predictions are combined by unweighted averaging. The prediction for $Y$ based on the training set $D$ is denoted with $\hat{Y}(X, D)$.

**Preliminaries:** The comparison relies on the strong law of large numbers\footnote{Theorem 3.30 in Breiman, L. (1992): Probability. SIAM Classics In Applied Mathematics 7.} that states for iid random variables $U_1, \ldots, U_n$

$$\frac{1}{n} \sum_{i=1}^{n} U_i \xrightarrow{n \to \infty} \mathbb{E}(U_1) \quad \tag{6.2}$$

In addition, the expectation of the indicator function $\mathbb{1}(.)$ for a discrete event $H$ is\footnote{Page 16 in Keener, R. W. (2010): Theoretical Statistics. Springer Texts in Statistics.}

$$\mathbb{E}(\mathbb{1}(H = h)) = \mathbb{P}(H = h) \quad \tag{6.3}$$

and the conditional expectation of $U$ given $\mathbb{P}(H = h) > 0$ is\footnote{Formula 1.1, Chapter 4 in Çınlar, E. (2011): Probability and Stochastics. Springer Graduate Texts in Mathematics 261.}

$$\mathbb{E}(U \mid H = h) = \mathbb{E}(U \cdot \mathbb{1}(H = h)) / \mathbb{P}(H = h) \quad \tag{6.4}$$

For the following, we silently assume that all expectations exist and smoothness, where necessary, takes place.
**LOO:** The mean squared error $E_{MSP}$ (or $E$ for short) of the bagged model can be estimated with the leave-one-out procedure according to

$$
\hat{E}_{n,b}^{LOO} = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \frac{1}{b} \sum_{j=1}^{b} \hat{Y}(X_i, D_{-i}) \right|^2
$$

(6.5)

As $b \to \infty$ we get (using Eq. 6.2)

$$
\hat{E}_{n,\infty}^{LOO} = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \mathbb{E}\left(\hat{Y}(X_i, D_{-i})\right) \right|^2
$$

(6.6)

**OOB:** The out-of-bag estimate for the mean squared error $E$ instead is defined as

$$
\hat{E}_{n,b}^{OOB} = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \frac{1}{b} \sum_{j=1}^{b} \mathbb{1}(Z_i \notin D^j) \sum_{j=1}^{b} \left( \hat{Y}(X_i, D^j) \cdot \mathbb{1}(Z_i \notin D^j) \right) \right|^2
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \frac{1}{b} \sum_{j=1}^{b} \mathbb{1}(Z_i \notin D^j) \cdot \frac{1}{b} \sum_{j=1}^{b} \left( \hat{Y}(X_i, D^j) \cdot \mathbb{1}(Z_i \notin D^j) \right) \right|^2
$$

(6.7)

For $b \to \infty$ we approach (using Eq. 6.2 to 6.4)

$$
\hat{E}_{n,\infty}^{OOB} = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \mathbb{E}\left(\hat{Y}(X_i, D^1) \cdot \mathbb{1}(Z_i \notin D^1) \right) \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \mathbb{E}\left(\hat{Y}(X_i, D^1) \cdot \mathbb{1}(Z_i \notin D^1) \right) \right|^2
$$

(6.8)

**Conclusion:** Approximately, we can state that

$$
\hat{E}_{n,\infty}^{OOB} = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \mathbb{E}\left(\hat{Y}(X_i, D^1) \cdot \mathbb{1}(Z_i \notin D^1) \right) \right|^2
$$

$$
\approx \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \mathbb{E}\left(\hat{Y}(X_i, D_{-i}) \right) \right|^2
$$

(6.9)

The equality does not hold since $D_{-i}^1$ amounts to sampling $n - 1$ times with replacement. Put the other way round: The equality would hold if we sample $n$ times with replacement in the LOO validation. However, one could argue that for large $n$ the difference is small.
6.3 Computer code

On https://github.com/schiggo some computer code to reproduce the experiments is hosted: The `hydroBE` R-package contains helper functions to e.g. calculate the continuous ranked probability score or to import data from various providers. The `SSOmod` R-package implements the present MOS method.
Bibliography


The reader might have noticed that the present thesis excessively uses listings, enumerations, and a small English vocabulary. Following this line, I allow myself to finally list here all the people that supported this document:

- Rolf Weingartner gave me the opportunity to work on seasonal stream-flow forecasting and supervised me in an open-minded manner.
- Ole Rössler contributed many ideas and discussions.
- Tom Reist helped whenever administrative issues needed to be clarified.
- Massimiliano Zappa agreed to review the present thesis.
- Tobias Fissler enabled the mathematical treatment of the out-of-bag prediction error estimate.
- Regula, Candace, Yannick, and Johannes spent time to fix bad writing styles.
- Over the years, the Institute of Geography was a pleasant and helpful environment in many scientific and non-scientific aspects – you are too many to name you all.
- Last but not least, my family and friends in and around Bern enriched my study years in various flavours.

Thank you.
Erklärung
gemäß Art. 28 Abs. 2 RSL 05

Name/Vorname: Schick Simon
Matrikelnummer: 07-911-639
Studiengang: Doktorat Geographie - Doktorat PHIL NAT, Reglement 2008
  Bachelor ☐ Master ☐ Dissertation ☑

Titel der Arbeit: Application of the model output statistics method to seasonal streamflow forecasting

LeiterIn der Arbeit: Prof. Dr. Rolf Weingartner
                     Dr. Ole Rössler

Ich gewähre hiermit Einsicht in diese Arbeit.

Ort/Datum

Unterschrift
Akademischer Lebenslauf

Studium

seit 2014 Doktorand Geographisches Institut
2011–2014 Master of Science in Geographie
2008–2011 Bachelor of Science in Geographie

Publikationen


Konferenzen

2015 Comparison of cross-validation and bagging for building a seasonal runoff forecast model. Vortrag, Konferenz Water Resources Assessment and Seasonal Prediction, Koblenz.

Sonstiges

2016 Teilnahme an der Sommerschule Extreme value modeling and water resources. Universität Lyon, Institut Camille Jordan.
2014 Auszeichnung Nachwuchspreis für junge Hydrologen. IHP-HWRP Deutschland, Koblenz.