

Seasonal Streamflow Forecasting for Fresh Water Reservoir Management in the Netherlands: An Assessment of Multiple Prediction Systems

RUUD T. W. L. HURKMANS,^{a,b} BART VAN DEN HURK,^c MAURICE SCHMEITS,^{a,d} FREDRIK WETTERHALL,^e
AND ILIAS G. PECHLIVANIDIS^f

^a *Royal Netherlands Meteorological Institute, De Bilt, Netherlands*

^b *HKV, Lelystad, Netherlands*

^c *Deltares, Delft, Netherlands*

^d *Institute for Environmental Studies, Vrije Universiteit, Amsterdam, Netherlands*

^e *ECMWF, Reading, United Kingdom*

^f *Swedish Meteorological and Hydrological Institute, Norrköping, Sweden*

(Manuscript received 29 June 2022, in final form 13 April 2023, accepted 17 April 2023)

ABSTRACT: For efficient management of the Dutch surface water reservoir Lake IJssel, (sub)seasonal forecasts of the water volumes going in and out of the reservoir are potentially of great interest. Here, streamflow forecasts were analyzed for the river Rhine at Lobith, which is partly routed through the river IJssel, the main influx into the reservoir. We analyzed seasonal forecast datasets derived from the European Flood Awareness System (EFAS), the Swedish Meteorological and Hydrological Institute (SMHI) European Hydrological Predictions for the Environment (E-HYPE), and Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land (HTESSEL), which differ in their underlying hydrological formulation, but are all forced by meteorological forecasts from ECMWF's fifth generation seasonal forecast system (SEAS5). We postprocessed the streamflow forecasts using quantile mapping (QM) and analyzed several forecast quality metrics. Forecast performance was assessed based on the available reforecast period, as well as on individual summer seasons. QM increased forecast skill for nearly all metrics evaluated. Averaged over the reforecast period, forecasts were skillful for up to 4 months in spring and early summer. Later in summer the skillful period deteriorated to 1–2 months. When investigating specific years with either low- or high-flow conditions, forecast skill increased with the extremity of the event. Although raw forecasts for both E-HYPE and EFAS were more skillful than HTESSEL, bias correction based on QM can significantly reduce the difference. In operational mode, the three forecast systems show comparable skill. In general, dry conditions can be forecasted with high success rates up to 3 months ahead, which is very promising for successful use of Rhine streamflow forecasts in downstream reservoir management.


SIGNIFICANCE STATEMENT: Lake IJssel is the Netherlands' largest freshwater reservoir, with its main water source coming from a branch of the river Rhine. We investigate whether seasonal forecasts of river discharge can help in managing the lake level to create extra buffer capacity for dry periods. We compare three seasonal forecast systems and assess their quality. We find that statistical corrections are needed for all systems to be used. In spring discharge can be predicted up to 4 months ahead due to snow processes. In summer this time is shorter, but it increases with event extremity: severe low-flow events can be predicted longer ahead. This offers potential for water managers to base their lake management on other similar reservoirs.

KEYWORDS: Europe; Rivers; Hydrology; Statistical techniques; Seasonal forecasting

1. Introduction

Lake IJssel in the Netherlands is a large freshwater reservoir. About 50% of the country's surface area is drained into the lake, and during summer droughts about 50% of the total freshwater use can be supplied from the lake (Waterman et al. 1998). The lake is fed by local precipitation and streamflow from a few rivers, of which about 85% is provided by the river IJssel, a distributary from the river Rhine. The lake is drained by evaporation, water transport into the surrounding region for regional use, and streamflow to the Wadden Sea. In 2018,

the routine water level policy switched from a fixed seasonal cycle—with lake levels of -0.40 m MSL in winter and -0.20 m MSL in summer—to a flexible lake level management (Rijkswaterstaat 2018). This was, among other reasons, to anticipate possibly enhanced future water shortages (e.g., van den Hurk et al. 2014). By raising lake levels prior to, or early during a drought, an extra buffer is created. To optimally manage the lake level, early indications of upcoming droughts are extremely valuable. Because the region recently experienced three dry years in a row (2018, 2019, 2020), attention for this subject has greatly increased (Buitink et al. 2020; Rakovec et al. 2022).

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Ruud Hurkmans, r.hurkmans@hkv.nl



This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

DOI: 10.1175/JHM-D-22-0107.1

© 2023 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](http://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

The water balance of Lake IJssel is governed by local precipitation, river inflow, evaporation out of the lake, and water extraction from the surrounding land. See [section 2](#) for details. The main influx is by far the inflow from river IJssel, which is highly correlated with discharge in the river Rhine, and the other components are driven by local meteorology. The forecast skill of meteorological seasonal forecasts is known to be relatively low in Europe due to the weak influence from large-scale controls of variability such as ENSO, which is a dominant source of seasonal predictability in many other areas (e.g., [Doblas-Reyes et al. 2013](#); [Yossef et al. 2017](#); [Lucatero et al. 2018](#)). Moreover, the North Atlantic Oscillation (NAO) provides limited skill in Europe, and mainly in winter ([Bierkens and van Beek 2009](#); [Scaife et al. 2014](#); [Sánchez-García et al. 2019](#)). On the other hand, forecasts of river streamflow generally have more skill, because it is not only determined by meteorology, but to a large extent by the system memory to initial conditions: groundwater, soil moisture, and snow ([Crochemore et al. 2020](#); [Pechlivanidis et al. 2020](#)). A correct estimation of the hydrological state at the beginning of the forecast, therefore, improves forecast skill ([Bierkens and van Beek 2009](#); [Arnal et al. 2018](#)). Because streamflow from the river IJssel is the main water source for Lake IJssel and seasonal forecasts of the governing meteorological variables are not expected to add forecast skill to the lake water balance, the focus of this study is on streamflow from the river IJssel. A skillful seasonal streamflow forecast would substantially help water authorities to anticipate lake level adjustments. In 2018, for example, a decision to raise lake levels was taken when IJssel streamflow was already insufficient to do so (Rijkswaterstaat 2020, personal communication), highlighting the need for seasonal predictions and early decision-making.

A number of approaches exist to forecast streamflow at the seasonal scale. When, for example, streamflow is largely determined by snow and relationships are well defined, regression approaches, based on data alone can be successful (e.g., [Abudu et al. 2010](#)). Other approaches employ (dynamic) hydrological models but are forced by observations, such as the ensemble streamflow prediction method (ESP; [Wood and Lettenmaier 2006](#); [Bennett et al. 2017](#)), which assumes no skill in meteorological forecasts and uses ensembles of historical meteorological forcing. Skill, therefore, only originates from initial conditions ([Harrigan et al. 2018](#)). However, meteorological forecast skill has improved over past decades, and currently a number of forecasting systems are available that take this source of forecast skill into account (e.g., [Arnal et al. 2018](#); [Greuell et al. 2018](#); [Wanders et al. 2019](#); [Pechlivanidis et al. 2020](#)). Most of these are based on meteorological seasonal forecasts from the ECMWF SEAS5 prediction system ([Johnson et al. 2019](#)) or its predecessor SEAS4, with different implementations of a hydrological model and the consequent process representation.

These forecasting systems are typically applied to the entire European continent, often without specific attention to the Rhine basin. [Greuell et al. \(2018\)](#) did not find the Rhine basin

to be among the regions with significant forecast skill. Studies specific to the Rhine basin include [Rutten et al. \(2008\)](#), who used correlation analyses based on large-scale oceanic and atmospheric patterns to predict summer discharge, and found these provided little information. [Demirel et al. \(2015\)](#) analyzed forecast skill of different hydrological models up to 90 days for the Moselle, a tributary of the Rhine, with promising results. [Sutanto and Van Lanen \(2022\)](#) analyzed streamflow drought forecasts for European rivers and found relatively large forecast potential for the Rhine because of the large baseflow (and snowmelt) contribution to streamflow, and confirmed the importance of land memory as a source of seasonal forecast skill. However, seasonal forecasts are not yet operationally used in the Netherlands. Because forecast skill is expected to originate mainly from land memory and the hydrological conditions, the objective of this study is to compare different available forecasting systems for the Rhine River, based on different hydrological schematizations, and to assess the potential for operational use.

Specifically, the research questions we will investigate are 1) whether streamflow forecasts provide meaningful information for reservoir management, 2) whether statistical postprocessing improves the forecast usability, and 3) which of the forecast systems provides the most useful information. To answer these questions, we compare a number of seasonal streamflow forecasting systems. A statistical postprocessing method (quantile mapping; QM) is applied to assess whether forecast skill can be improved. We compare the forecasts using a number of forecast skill metrics, each assessing a different aspect of forecast quality, by analyzing a long hindcast period, but also investigate specific years with both high- and low-flow conditions. We categorize the streamflow forecasts into terciles to assess the corresponding statistics. Finally, to assess the differences between the employed systems and their significance we use the Diebold–Mariano significance test ([Diebold and Mariano 1995](#)).

The paper is organized as follows. In [section 2](#), we present the study area along with the available meteorological forecasts and hydrological forecasting systems. [Section 3](#) presents the methodology for forecast postprocessing and evaluation. [Section 4](#) presents the results, followed by a discussion in [section 5](#). Finally, [section 6](#) states the conclusions.

2. Study area, data, and forecasting systems

a. Study area

Lake IJssel was cut off from the open sea in 1932 and has been a freshwater reservoir since then. [Figure 1](#) shows the geography of the region including streamflow and intake locations for the surrounding regions. In dry periods, a large part of the Netherlands, approximately represented by light shading in [Fig. 1](#), depends on Lake IJssel for freshwater supply. Sources of water include local precipitation and a number of inflows, of which the IJssel is by far the largest. Besides evaporation from the lake, surrounding areas extract water from it during summer and the excess is flushed toward the Wadden

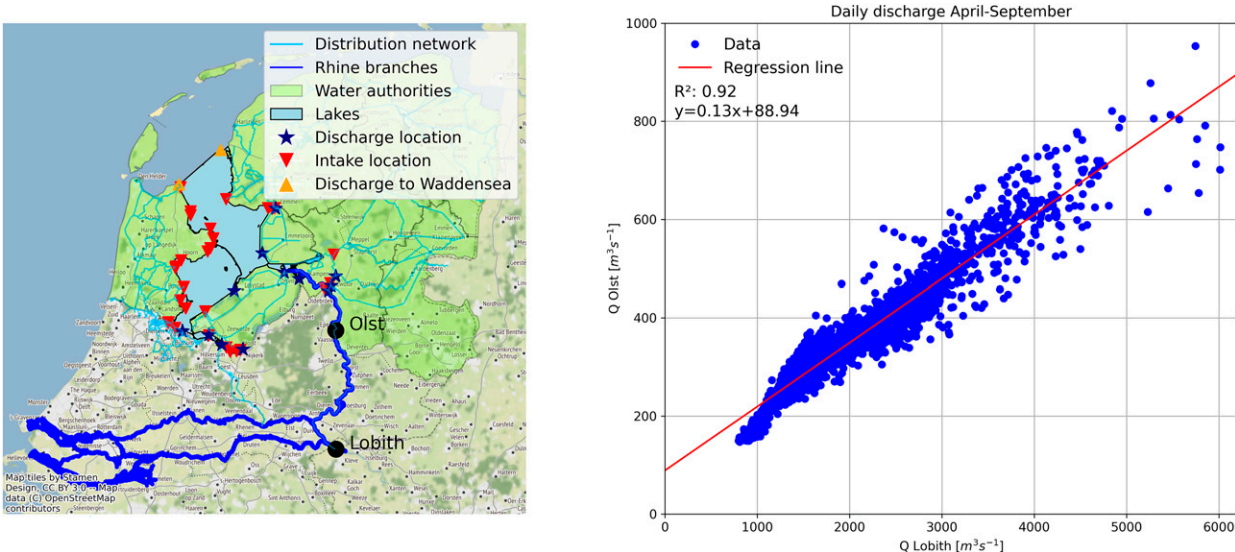


FIG. 1. (left) Geography of the region. Lakes IJssel and Marken are shaded in light blue. Light green areas indicate regions that (partially) depend on freshwater from Lake IJssel in dry periods with the main branches of the distribution network shown in light blue. Blue stars indicate inflow locations, red triangles extraction locations, and orange triangles discharge locations to the sea. Rhine branches are shown in dark blue. (right) The relationship between daily IJssel streamflow at Olst (the most downstream observation location in the IJssel) and daily Rhine streamflow at Lobith.

Sea. To illustrate the relative importance of the discharge from the IJssel for the reservoir water balance, Table 1 gives an overview of the lake water balance. IJssel discharge is the main input to the lake, and it also governs the main flux going out of the lake, discharge to the Wadden Sea, as the lake level is kept mostly constant. One of the objectives to implement the flexible lake level management policy in 2018 is to allow raising water levels, and create extra buffers, when droughts are anticipated. The gauging station in the IJssel closest to the lake is Olst. However, we opt to use the station of Lobith, which is just upstream of the bifurcations in the Dutch part of the Rhine (Fig. 1) to maximize the upstream catchment area. The relation between discharge at Olst and Lobith is well defined (right panel of Fig. 1).

During peak flows, about 11% of the Rhine streamflow is routed through the IJssel. Under drier conditions, this fraction can be increased by weirs in the Nederrijn, another Rhine branch. Figure 1 shows the relation between streamflow at Olst (in the river IJssel) and Lobith (in the Rhine upstream of the IJssel branch), averaged over April–September for the period 1993–2018. On average in that period, about 15% of

Rhine streamflow is routed through the IJssel, depending on specific conditions.

b. Data and forecasting systems

We analyze streamflow (re)forecasts from three seasonal forecasting systems driven by the ECMWF’s fifth generation seasonal forecast system (SEAS5) forecasts (ECMWF 2017): streamflow derived from SEAS5, hereafter referred to as HTESEL (Johnson et al. 2019); streamflow from the European Flood Awareness System, hereafter referred to as EFAS (Arnal et al. 2018); and streamflow from the Swedish Meteorological and Hydrological Institute (SMHI) European seasonal forecasting service, hereafter referred to as E-HYPE (Pechlivanidis et al. 2020). Table 2 shows relevant properties of all forecast systems. From all systems, we extract hindcasts and forecasts for the location Lobith, where the river Rhine enters the Netherlands (Fig. 1). We consider hindcasts with 25 ensemble members for the period 1993–2015 and forecasts for the exceptionally dry summer of 2018, with 51 members. For all forecast systems, the daily time series are aggregated to weekly averages. Aggregation is necessary to capture the

TABLE 1. Overview of the (relative) magnitude of all fluxes flowing into and out of Lake IJssel. Fluxes are taken from a simulation of the Dutch National Water Model (NWM), spanning a period of 111 years (Mens et al. 2021), and averaged over the period typically associated with dry periods in the region (March–October). Numbers are expressed as equivalent change in lake level (using an area of 1956 km²) and as a percentage of the total inbound (or outbound) flux.

| Flux into the lake | Absolute (mm day ^{−1}) | Relative (%) | Flux out of the lake | Absolute (mm day ^{−1}) | Relative (%) |
|--------------------|----------------------------------|--------------|----------------------|----------------------------------|--------------|
| Precipitation | 2.05 | 10.7 | Evaporation | 2.63 | 13.8 |
| IJssel | 14.3 | 74.8 | Extraction | 0.92 | 4.8 |
| Other | 2.78 | 14.5 | Waddensea | 15.5 | 81.4 |
| Total in | 19.1 | 100 | Total out | 19.1 | 100 |

TABLE 2. Overview of the three forecasting systems.

| System | HTESSEL | EFAS | E-HYPE |
|------------|--------------------------------|--|--|
| Model | HTESSEL | LISFLOOD | HYPE |
| Resolution | $36 \times 36 \text{ km}^2$ | $5 \times 5 \text{ km}^2$ | 215 km^2 |
| Variable | Subsurface + surface flow (mm) | Streamflow ($\text{m}^3 \text{ s}^{-1}$) | Streamflow ($\text{m}^3 \text{ s}^{-1}$) |
| Routing | No | Kinematic wave | Delay and dampening |

relevant signal and reduce noise in the forecasts (Zhu et al. 2014; Robertson and Vitart 2018). A weekly time step is a relevant interval for low-flow events on one hand, but on the other hand leaves sufficient data points to allow a robust statistical analysis.

1) HTESSEL

For the ECMWF SEAS5 forecasts, we aggregate total runoff from the land surface scheme HTESSEL (Balsamo et al. 2009) at native resolution [about 36 km according to ECMWF (2017)] over the Rhine catchment. In this case, no surface routing is applied and also reservoirs are not considered. Because we aggregate the daily time series to weekly averages and the travel time between Switzerland and Lobith is about 5 days (Khanal et al. 2019), the lack of routing is assumed to have a negligible impact on the results. Different from the other systems, the land surface model HTESSEL has a focus on accurate representation of the water and energy balances, and is not tuned on simulating a correct river hydrograph. Especially during dry periods, the representation of river streamflow strongly relies on baseflow and this may be expected to be less accurate than the streamflow representation in the other models. To illustrate the quality of a seasonal forecast for a meteorological variable, we also extract total precipitation from the ECMWF-SEAS5 forecasts and aggregate it to the Rhine catchment in the same way as runoff.

2) EFAS

In EFAS, the underlying hydrological model is LISFLOOD (Burek et al. 2013), a GIS based, distributed model running on a $5 \times 5 \text{ km}^2$ resolution over entire Europe. It was calibrated on the period 1993–2002 at 693 European catchments (Arnal et al. 2018). LISFLOOD can represent snow, glaciers, frozen soils, and soil moisture variability, which is parameterized using the Xinanjiang formulation, similar to, for example, the Variable Infiltration Capacity (VIC) model (Liang et al. 1994). Groundwater is included by means of a quick (shallow groundwater and macropore flow) reservoir and a slow (baseflow) reservoir, similar to the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Lindström et al. 1997). The resulting runoff is routed using a kinematic wave function through the river network, including reservoirs and lakes.

3) E-HYPE

The HYPE (Hydrological Predictions for the Environment) model is in its European application referred to as E-HYPE (Hundecha et al. 2016), is a semidistributed model, where

hydrological response units (HRUs) are based on land cover and soil. HRUs have an average size of about 215 km^2 . The soil is schematized using three layers, where the bottom layer accounts for groundwater. Deep aquifers are not taken into account, however, both snow and glacier processes are. Streamflow is routed through the river network, where reservoirs and lakes also dampen the streamflow according to rating curves. The ECMWF SEAS5 meteorological forecasts are bias-corrected using the distribution-based scaling method (DBS; Yang et al. 2010) and the HydroGFD dataset (Berg et al. 2018) as reference.

4) OBSERVATIONS

All forecast systems are benchmarked against discharge observations from the Dutch national water authorities, disseminated through the open data portal <https://waterinfo.rws.nl>. Daily discharge observations at Lobith were obtained, spanning the period from 1 January 1901 to 1 January 2019.

3. Methodology

a. Quantile mapping

Quantile mapping is a relatively simple approach to match the cumulative density functions (CDF) of forecasts and observations (Panofsky and Brier 1958; Wetterhall et al. 2015; Crochemore et al. 2016; Ratri et al. 2019), since it has the advantage to correct the entire statistical distribution (Thrasher et al. 2012). In this application we compare all forecast values for a given forecast starting month and with lead times binned in 30-day intervals (where every lead time interval contains 4-weekly time steps). A specific month and lead time has, therefore, 92 (4 weeks per month times 23 years during the 1993–2015 period) data points. We then compose the observed and forecasted CDFs using 2% intervals and for each bin, we calculate a multiplicative factor.

We do not fit the multiplication factors to all years in the hindcast period, but employ a leave-one-out procedure; during the verification all metrics below show the average of all years, where the parameters for each year are fitted on all other years. This is justified because, for every calendar month, the autocorrelation of monthly averaged streamflow, with a lag of one year, is well below the significance level of 0.18 (based on an observed time series of 119 years, at a significance level of 5%). For illustration, Fig. 2 shows the quantile mapping for the forecasts starting on 1 April, and Fig. 3 shows the multiplication factors as a function of lead time for the relevant forecast months.

From Figs. 2 and 3 it is clear that HTESSEL, as expected, typically requires the largest correction. Also higher correction

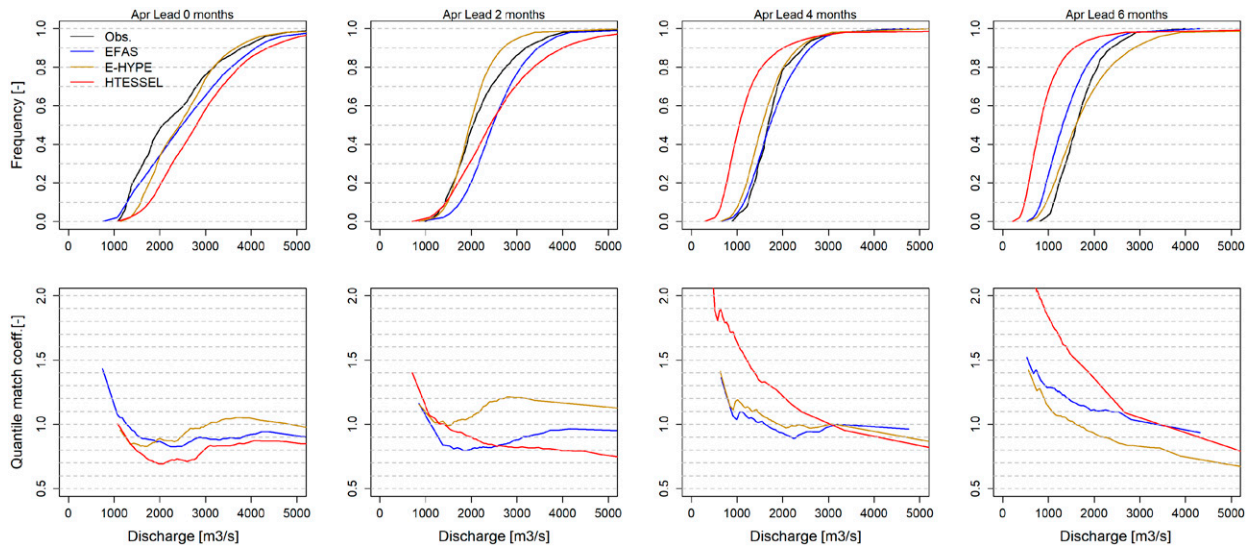


FIG. 2. (top) CDFs from EFAS (blue), E-HYPE (golden), HTESEL (red), and observations (black), derived from forecasts issued on 1 Apr for lead months 0 (April), 2 (June), 4 (August) and 6 (October); (bottom) mapping factors between observed and forecasted CDFs.

coefficients are needed at lower streamflows, where all products typically underestimate low streamflows to some extent. For HTESEL, the required correction is largest for intermediate lead times during the summer months. For lead times of

6 and 7 months, the forecasted discharge distribution is closer to the observed distribution than for lead times between 3 and 5 months. In the next section we discuss the effect of applying QM on the forecast skill.

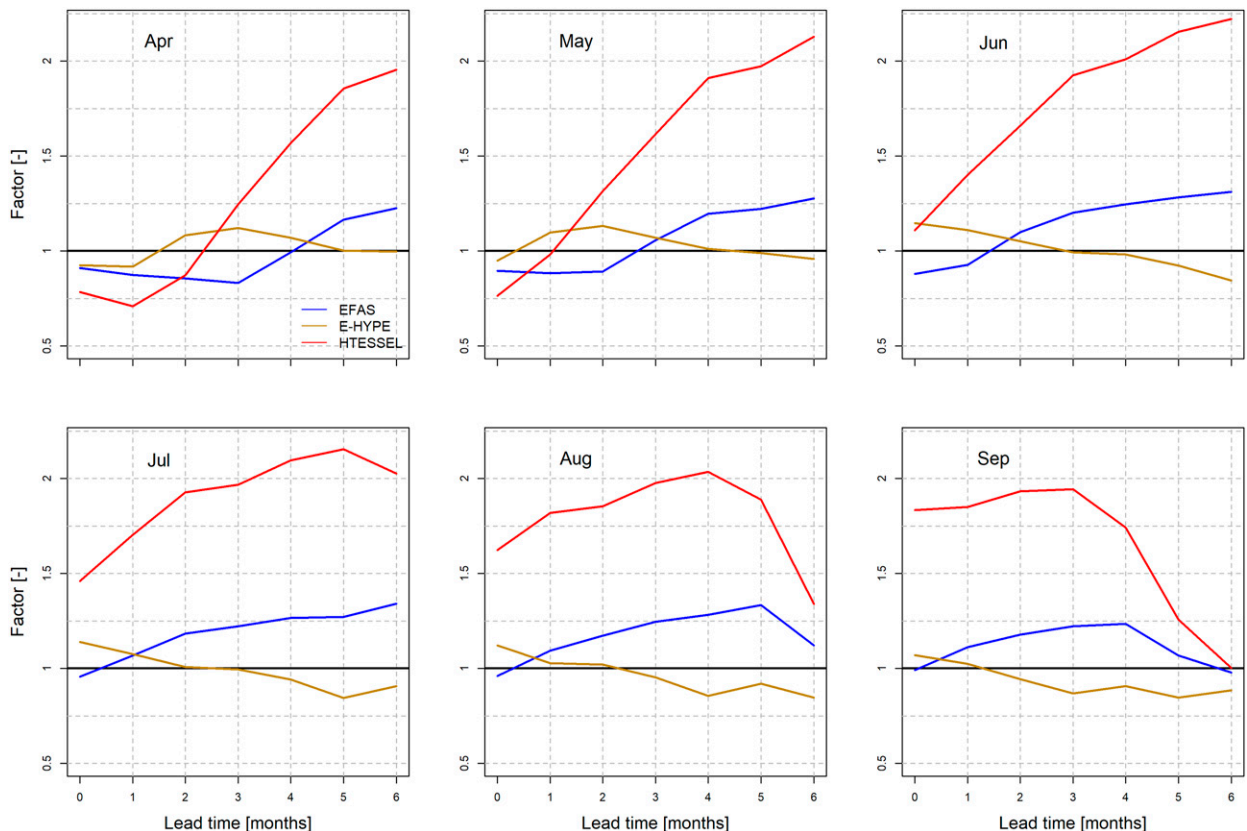


FIG. 3. Multiplication factors as a function of lead time for six summer months (April–September) for EFAS, E-HYPE, and HTESEL. All forecasts shown in a panel refer to the same month. For instance, the forecast with a lead time of two for the month of May was issued in March.

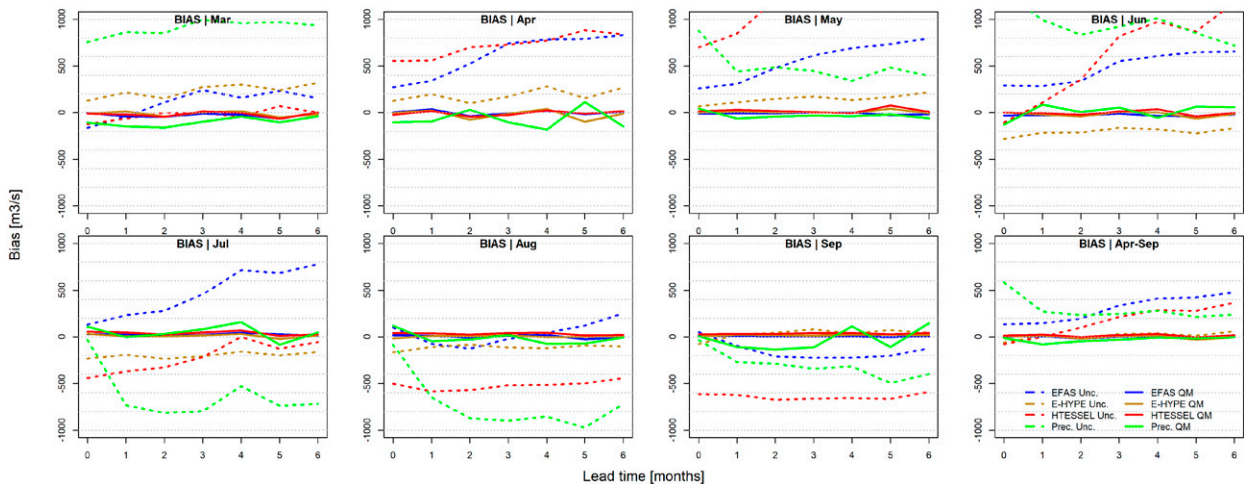


FIG. 4. ME for all hindcasts, with and without bias correction, for each lead time (months).

b. Evaluation of forecast skill

Forecast skill is evaluated in two ways. First, we analyze forecast systems averaged over the entire reforecast period (1993–2015; section 1). In addition, the performance during specific years with varying streamflow conditions is investigated in section 2.

1) REFORECAST PERIOD

We assess the forecast skill and how it is impacted by QM for the reforecast period (1993–2015). Three measures are used to illustrate this. First, to verify the effect of forecast postprocessing using QM, the mean error (ME) of the ensemble forecasts medians is shown. Second, as a measure of general forecast performance, we use the continuous ranked probability score (CRPS; Matheson and Winkler 1976), which shows the deviation between the observed and forecasted CDF. Third, to focus on low-flow forecasts, we used the Brier score (BS; Brier 1950) for underexceedance of lower tercile (33%) flows for each calendar month. For comparison, we also compute the BS for exceedance of the higher tercile (66%). All score metrics use observed streamflow as reference. We calculated the skill scores of CRPS and BS with the skill defined as

$$S_{\text{skill}} = 1 - \frac{S}{S_{\text{clim}}}, \quad (1)$$

where S_{skill} is the skill score, S is the calculated score (CRPS or BS), and S_{clim} is the score when using observed streamflow climatology, calculated over the hindcast period (1993–2015), as a benchmark forecast. If S_{skill} is 1 the forecast is perfect, while negative skill implies lower performance than using climatology as a forecast system. In the following, the skill scores based on CRPS and BS are denoted as CRPSS and BSS, respectively. Similar to the QM analysis, lead times are taken as 30-day intervals. For reference, we calculate the CRPSS of precipitation from the ECMWF SEAS5 forecasts, averaged over the Rhine basin, using spatially averaged

EOBS-precipitation v21.e (Cornes et al. 2018) as benchmark. By including precipitation, we illustrate the difference between meteorological and hydrological forecasts and the effect of land memory.

2) INDIVIDUAL YEARS

To assess the forecast skill for individual events, we use tercile plots, which is a widely used approach to assess probabilistic forecasts (e.g., Doblas-Reyes et al. 2013; Johnson et al. 2019). Tercile plots show the probability that (in this case) monthly discharge will fall in one of three categories: below normal, normal, or above normal. The 33% and 67% percentiles (the terciles) form the boundaries between categories and were derived from the reforecast period. Associated to the tercile plots is the ranked probability skill score (RPSS), which is essentially the Brier skill score generalized to more than two categories (Weigel et al. 2007). Similar to CRPSS and BSS, RPSS is benchmarked using an observed streamflow climatology as a reference.

In addition, another metric is extracted from the terciles: the absolute difference between the probability of higher-than-normal versus lower-than-normal streamflow. For every calendar month, we bin the absolute value of this difference for all forecasts with 20% intervals. Subsequently, we score for every bin how often the forecast proved to be right, i.e., the realization was in the same tercile as the forecast, excluding cases where the observation category was “normal.” The resulting metric is related to a reliability plot, which shows the forecast probability versus the observed relative frequency.

4. Results

a. Impact of postprocessing on forecast skill

Here, we analyze the forecasts before and after postprocessing aiming to assess the impact of QM on streamflow forecast skill. Figure 4 shows the mean error for the forecasts medians, and Figs. 5–7 show the skill scores (March–September) as function of lead time for CRPSS and BSS,

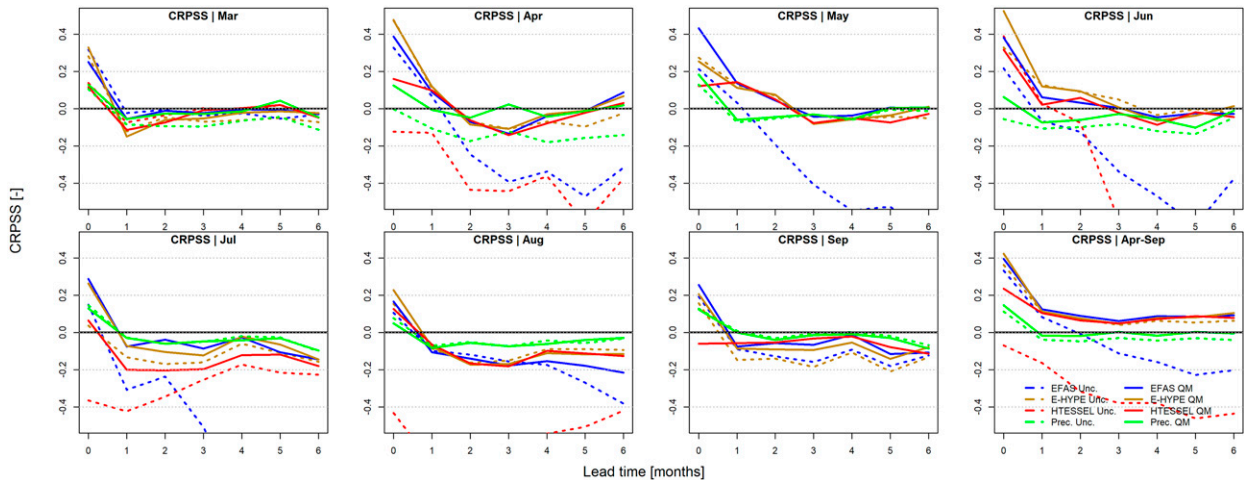


FIG. 5. CRPSS for all hindcasts, with and without bias correction, for each lead time (months). Also shown is the forecast of precipitation aggregated over the Rhine basin (see text). For a given calendar month we evaluate forecasts at seven issue dates, so, for example, in the top-left panel we evaluate forecasts issued between September and March.

respectively. BSS is shown for both underexceedances of the lower tercile and exceedances of the higher tercile.

As can be seen in Fig. 4, there is still significant bias present, especially in raw HTESSEL and, to a lesser degree, in raw EFAS. Note that meteorological input for E-HYPE has already been bias-corrected using the DBS method. Also in the raw precipitation forecasts from ECMWF there is considerable bias in individual months. Interestingly, this bias is positive in spring and strongly negative in summer, particularly July and August. Aggregated over the entire period (April–September), the bias is, therefore, relatively small. The same pattern of positive bias in spring and negative in summer is also present in the runoff from H-TESSEL and to a lesser degree in E-HYPE. For this region, climate scenarios show precipitation increases in winter and spring, and decreases in summer and autumn (e.g., Jacob et al. 2014). The

timing of the transition and the severity of the drought are uncertain. Possibly, this uncertainty is reflected in the bias we find here. In all cases, QM effectively removes the bias in raw HTESSEL and EFAS and further reduces the bias in E-HYPE, with all systems performing similarly in all forecast months and lead times.

Similar conclusions are drawn for the probabilistic metrics CRPSS and BSS (Figs. 5 and 6). Largest forecast skill is present in spring and early summer, when streamflow appears predictable up to 4 months ahead. Presumably this is due to the melting of the Alpine snowpack. Later in the summer, when streamflow is rain dominated, the skill is much lower than in spring and gets negative after about 2 months. Regarding prediction system intercomparison, the differences between forecast systems are small: only in June E-HYPE shows slightly higher CRPSS scores than

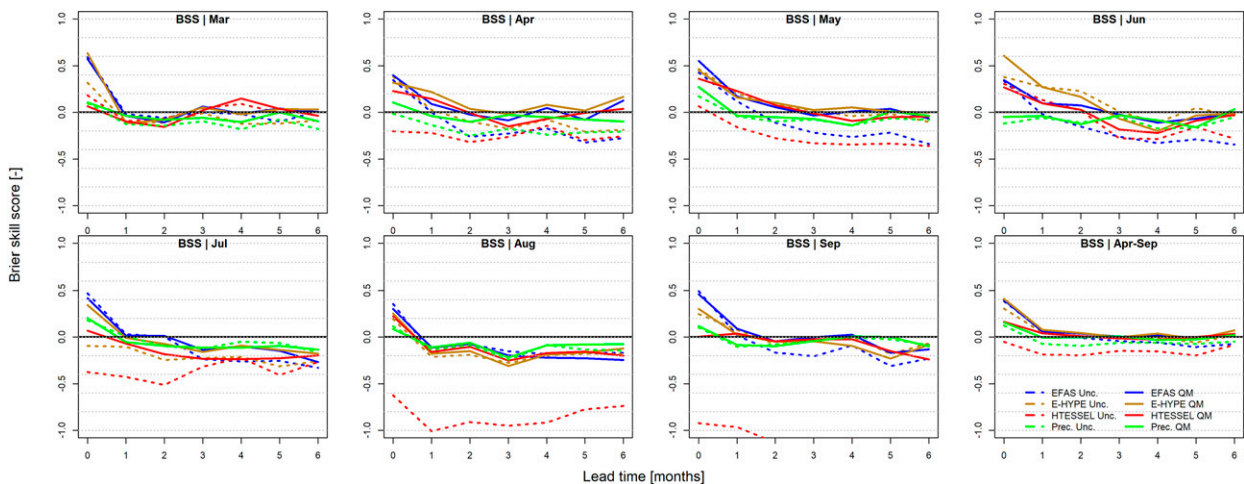


FIG. 6. As in Fig. 5, but for the BSS. The threshold for calculating the Brier score is based on an underexceedance of the lower tercile (33rd percentile).

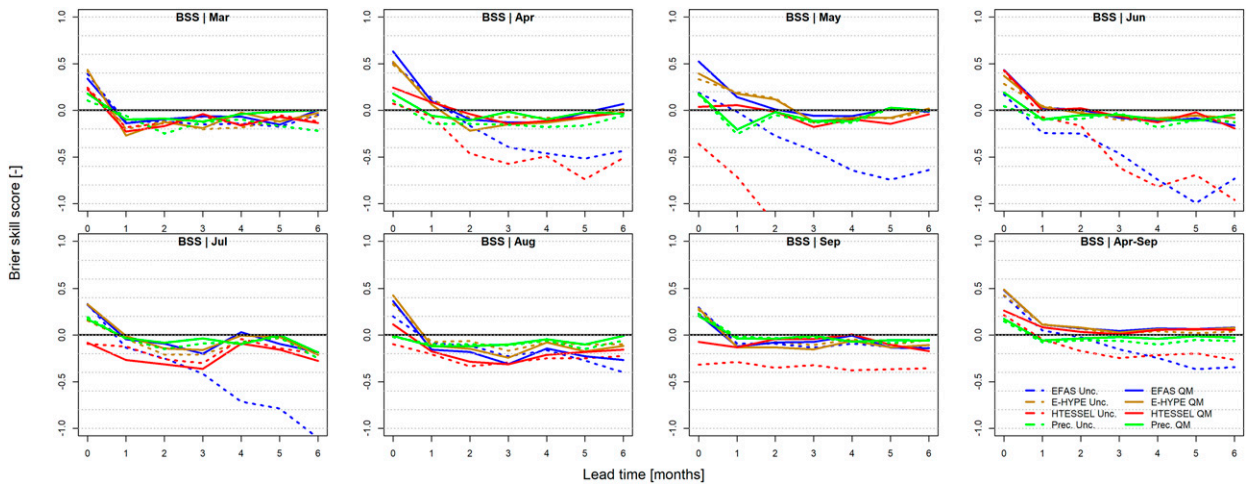


FIG. 7. As in Fig. 6, but here the threshold for calculating the Brier score is based on an exceedance of the upper tercile (66th percentile).

EFAS and HTESSEL. For precipitation forecasts, CRPSS and BSS scores are always negative beyond lead times of a month. After the first month, precipitation forecast skills remain close to zero and become less negative than the discharge forecast skills. Precipitation forecasts return to climatological values soon after the first month; hence the scores resemble those of the climatology and the skill scores do not become strongly negative. Due to the stronger (and longer) influence of initial conditions, streamflow forecasts generally deviate more from climatological values both in a positive and in a negative sense. Although differences between CRPSS and BSS (Figs. 5–7) are small, low-flow conditions are slightly more predictable than high-flow conditions, particularly in July. When considering

exceedances of a threshold, as does the Brier score (Figs. 6 and 7), the forecast skill is generally higher than when considering the entire distribution as the CRPS does (Fig. 5).

To analyze whether a specific forecast is better than another in certain parts of the year, we carried out a number of Diebold–Mariano tests (Diebold and Mariano 1995), specifically developed to test the hypothesis of equal forecast accuracy, and assessed whether prediction errors of one forecast ensemble median are significantly lower (at $p < 0.05$) than another forecast median. Figure 8 shows a summary of the results per calendar month and lead time. We used one-sided tests, to be able to determine whether one forecast is significantly better or worse.

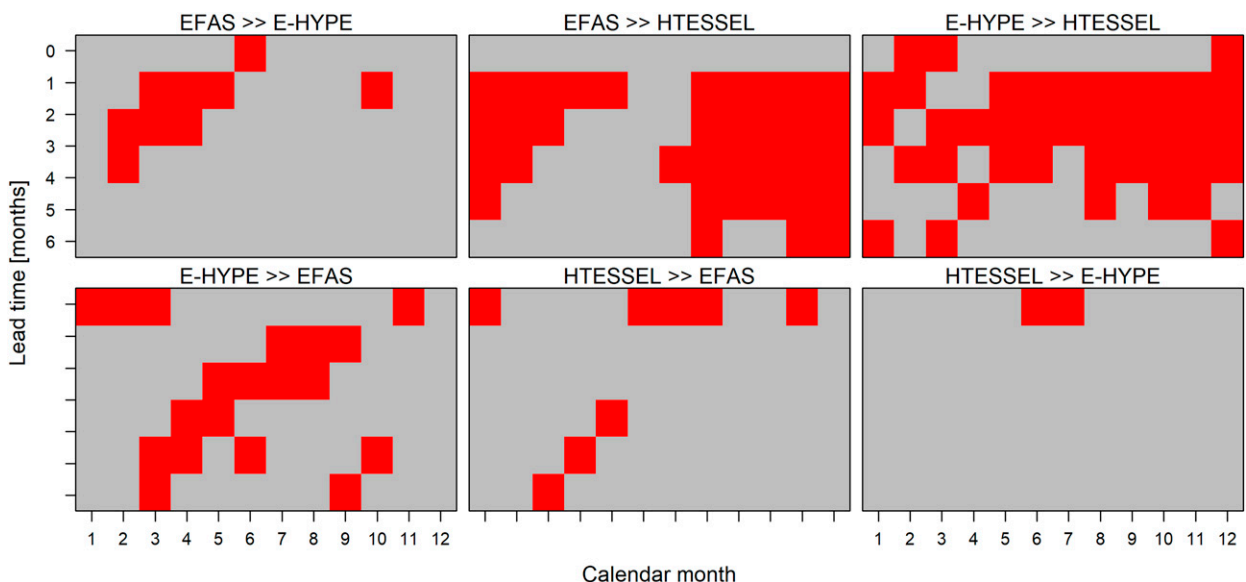


FIG. 8. Summary of the results of Diebold–Mariano tests between bias-corrected forecast medians per calendar month and lead time. Significance of differences is plotted as a function of lead time (horizontal) and calendar month (vertical). The two rows show opposite tests: i.e., the upper left shows where EFAS is significantly better than E-HYPE; the lower left shows where E-HYPE is significantly better than EFAS. Significant differences are expressed by red colors.

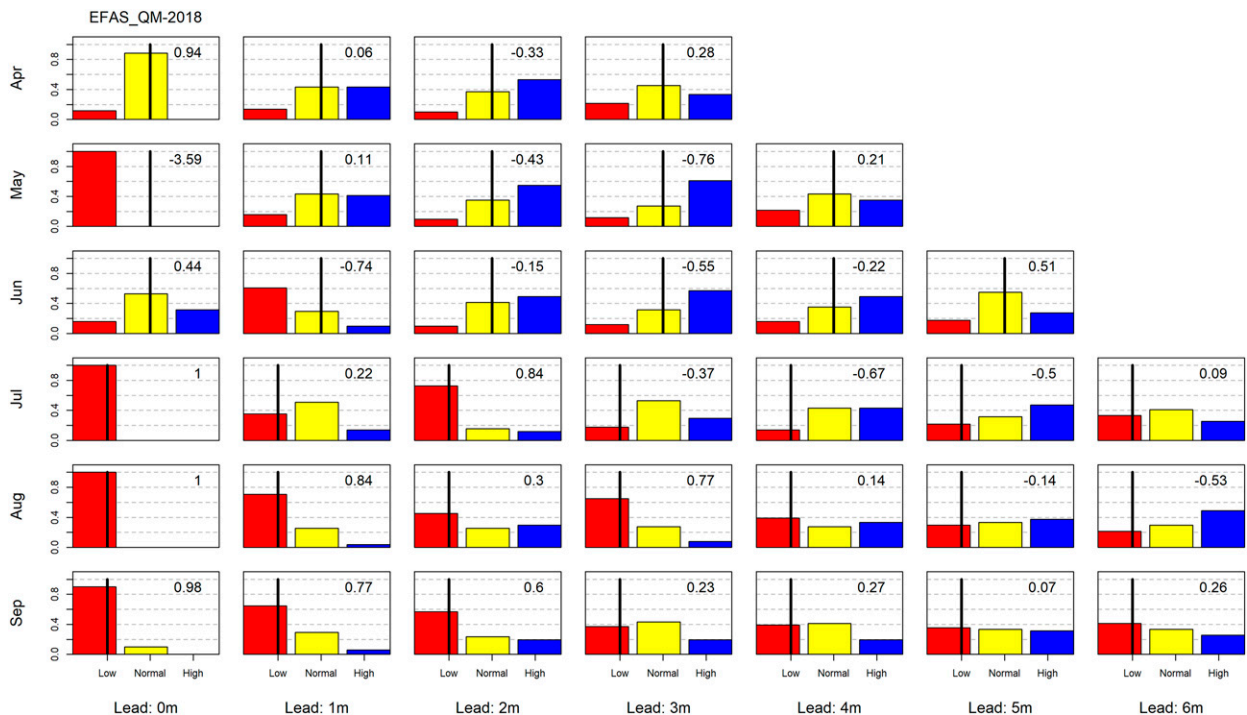


FIG. 9. Example of a tercile plot using the QM-corrected EFAS forecasts at Lobith, for 2018. Red bars indicate below normal streamflow, and blue bars are above normal. The black line is the observed realization, and the number indicates the RPSS, consistent with the thresholds and using observed climatology as a reference [Eq. (1)]. The bars show the probabilities for the three categories. Panels from top to bottom show the calendar months of 2018, with lead time (aggregated to 30-day bins) in horizontal rows.

Figure 8 shows that bias-corrected EFAS and E-HYPE outperform bias-corrected HTESSSEL throughout most of the seasonal cycle. Before correction with QM, this difference is larger than post-bias correction: raw forecasts for E-HYPE and EFAS are nearly always better than for HTESSSEL, whereas raw E-HYPE forecasts are generally better than raw EFAS forecasts. It should be noted here that the “raw” E-HYPE forecasts were already partly corrected (see section 2b). The differences between bias-corrected E-HYPE and EFAS are slightly more complex, with better performance for EFAS in spring (for short lead times) and E-HYPE in summer, which indicates the potential for multimodeling in the region.

b. Assessment in specific years

Next, we assessed the streamflow forecast skill for selected “interesting” years from a decision-making perspective. As mentioned before, 2018 is highly relevant because a period with low Rhine (and IJssel) discharges coincided with a period of high demand from the reservoir.

Figure 9 presents the forecasts at Lobith for a given month and lead time during 2018. We consider the bias-corrected results, and carried out a leave-two-year-out validation and fit parameters based on the other years. Because at long lead times data from the previous year are also used, we leave both the target and the previous year out of the fitting procedure. So, for example, when considering the summer of 2003, QM factors are derived during 2000, 2001, and 2004–15.

Forecasts are promising for the summer of 2018; for up to 4 months ahead lower-than-normal conditions were forecasted with probabilities higher than 0.5 (red bars in Fig. 9), consistent with observations. We made similar plots to Fig. 9 for basin averaged precipitation and temperature compared to E-OBS observations, to investigate meteorological predictability during extreme conditions, and found virtually no skill after the first month. As these findings confirm results from literature (e.g., Lucatero et al. 2018), we do not show all of these results.

To explore the forecast performance characteristics in other periods, we ranked April–October averaged streamflow seasonal forecasts from the reforecast period (1993–2015) and 2018 and selected the four years with highest and lowest average streamflow. Figure 10 shows time series of observed streamflow with ensemble medians of EFAS, E-HYPE, and HTESSSEL, both raw and bias-corrected for the forecasts initialized on 1 April, of the year with the highest (2001) and lowest (2003) streamflow from the ranking.

Figure 10 illustrates the impact of QM correction. Especially during low streamflows, the HTESSSEL forecasts severely underestimate observed streamflow highlighting the need for a high correction factor (Fig. 2). During extremely dry summer seasons, such as 2003 and 2018 (the latter is not shown here), the raw HTESSSEL forecasts are well in line with the observations. This explains the promising forecast performance as shown in Fig. 9; however the raw HTESSSEL

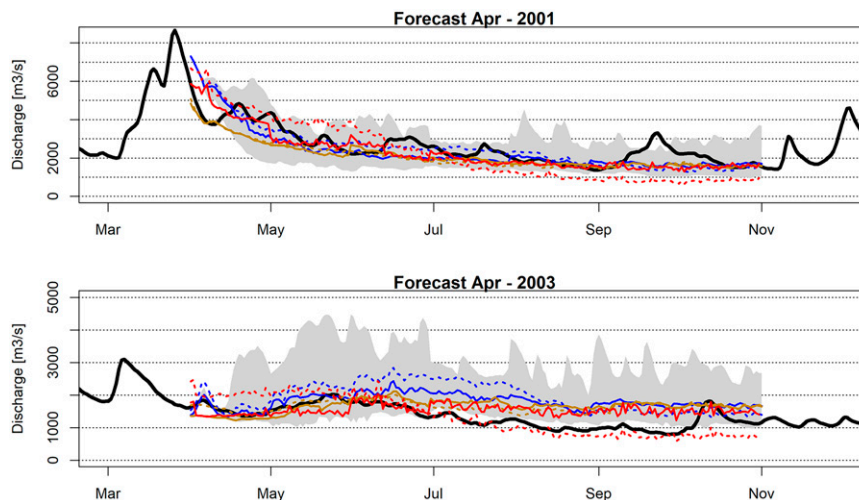


FIG. 10. Example of forecasts from 1 Apr for the year with the (top) highest (2001) and (bottom) lowest (2003) average summer streamflow. Observations are shown in black; ensemble medians of EFAS in blue, E-HYPE in gold, and H-TESSEL in red. Solid lines refer to QM-corrected forecasts, while dotted lines refer to raw values. The gray shading indicates the spread in the ensemble for raw EFAS forecasts.

forecasts also show low streamflow in normal and above-normal periods, which thus could lead to a high false-alarm rate. After bias correction, the differences between the forecast systems are relatively small. Figure 10 shows that after about 3–4 months, all forecasts converge to the climatological values, as expected. Figure 11 contains a graphical summary of the monthly RPSS scores that are associated with the individual panels in the tercile plots (similar to Fig. 9).

In general, the RPSS values for three forecast systems are quite similar (Fig. 11). Also it shows that the long-lead forecasts for August and September 2018 have good performance. However, in years with extremely high or low streamflow, forecast skill seems to be above average (in relevance to Fig. 5). To better illustrate this, Fig. 12 shows the averaged RPSS over lead times up to 4 months and the summer half-year plotted against rank numbers of years ordered from years with lowest (2003, left) to highest (2001, right) streamflow, averaged over the period April–October.

In some cases, the tercile categorization amplifies relatively small differences in absolute streamflow, occasionally resulting in a considerable decrease of the monthly RPSS score. For example, in August 2011, the negative RPSS scores are related to a temporary increase in observed streamflow at the end of July or beginning of August, which causes to shift the observed condition to “normal,” whereas still lower than normal streamflow was forecasted. The sensitivity of the RPSS score to the threshold used highlights the need for visual inspection using figures such as shown in Figs. 9 and 10, in addition to analyzing metrics alone. The negative RPSS values in Fig. 11 are all associated with such events, in which observed streamflow is categorized in the middle tercile. From this visual inspection, there appears to be relevant information in the difference between the probability of higher-than-normal versus lower-than-normal discharge, although this is not

always reflected in the RPSS score. We present this information in Fig. 13 by plotting the fraction of correct hindcasts (hindcasts with the corresponding observation in the same tercile) as a function of the absolute difference between the high and low hindcast probabilities, according to the approach described in section 2. If either the low or the high tercile contains all ensemble members, therefore, the difference on the horizontal axis is 100%. The metric shows that forecasts become more reliable, in terms of the observation being in the same tercile, with increasing difference between the probabilities of high and low discharge. It is related to a reliability plot, which shows the observed frequencies of an event as a function of forecast probability. A reliable forecast, therefore, is as close as possible to the diagonal line in the plot (Wilks 2011). A reliability diagram, however, appeared to be more difficult to interpret by local water managers. For reference, we show reliability plots for underexceedance of the lower tercile and exceedance of the upper tercile as well.

Reliability plots provide information about both the reliability and the discrimination of the forecast. Due to relatively small datasets, they are noisy and it is difficult to draw conclusions from them. In general, forecasts during low-flow conditions (left column in Fig. 13) are more reliable than those during high-flow conditions (middle column). For high flows, generally the forecast discrimination is also worse than for low flows: especially for H-TESSEL the high end of the observed frequency distribution is never forecasted. This is consistent with the strong underestimation of discharge by H-TESSEL in the summer months we saw earlier. Forecasts from all three systems are more reliable in spring, as the lines for April, May, and June are generally closer to the diagonal line than the lines for the summer months, and E-HYPE generates more accurate forecasts in the summer months (see also Fig. 8). The information as presented in Fig. 13 may be

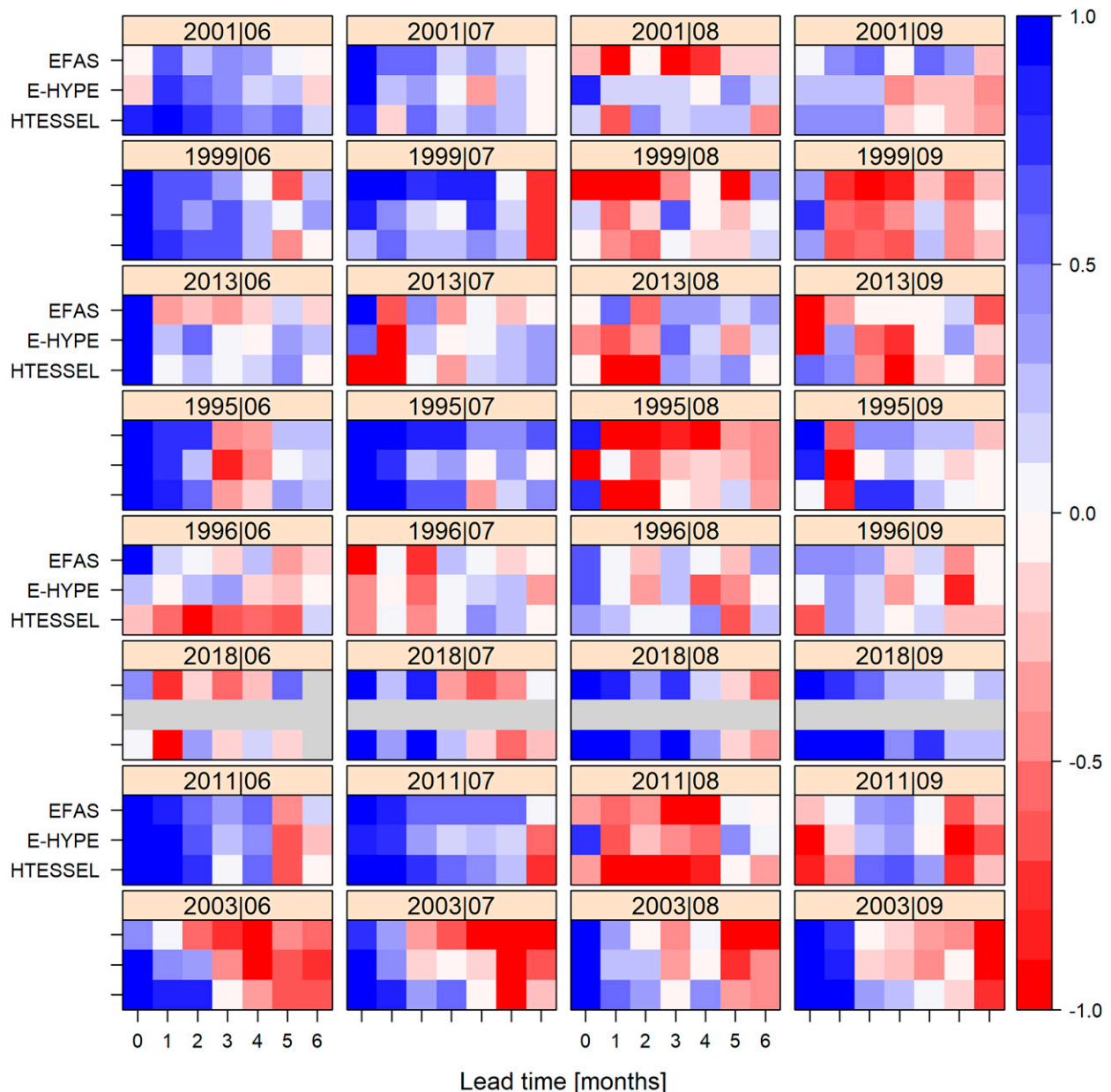


FIG. 11. RPSS scores of QM-corrected monthly streamflow forecasts of EFAS, E-HYPE, and HTESEL. The panels show, from top to bottom, the four years with highest (four upper panels) and lowest (four lower panels) average summer streamflow. From left to right, the calendar months June–September are shown. Positive values indicate skillfulness relative to climatology. Gray colors indicate nonavailability of data for a specific month and lead time.

useful for operational managers, where instead of an average skill score, the probability for certain forecasts being correct is presented. With the forecast archive growing, the noise in the results will decrease as the curves become more stable.

5. Discussion

Our results largely confirm earlier results by (Arnal et al. 2018), who found increased forecast skill in spring in (partly) snow-fed rivers and lowest forecast skill in summer. In winter,

the North Atlantic Oscillation (NAO) positively affects forecast skill in European rivers (Bierkens and van den Hurk 2007; Scaife et al. 2019). Arnal et al. (2018) indicated a few regions in Europe with increased forecast skill in summer, but the Rhine appeared not to be one of them. To ensure statistical robustness, these analyses resulted in average forecast skill over a large number of years. Although we confirmed that average summer forecast skill is low and varies between years, we also found that skill increases with the extremity of the event: summers with extremely low discharges were skillfully

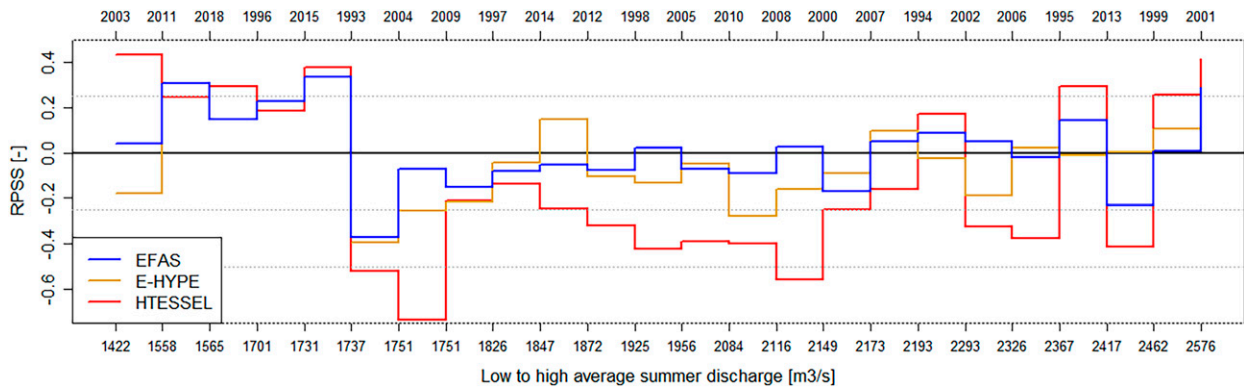


FIG. 12. RPSS scores, averaged over the summer half year (April–October) and lead times up to 4 months for years ranked from low (left) to high (right) averaged streamflow.

forecast longer ahead, up to 4 months. Ionita and Nagavciuc (2020) found similar results for the summer of 2018 based on statistical forecasting systems. They found especially sea surface temperature in parts of the northern Atlantic Ocean to be a good predictor of Rhine River discharge for long lead times. Meißner et al. (2017) found such statistical forecasting methods to outperform more physically based methods such as SEAS5. As statistical and physically based methods each have their strengths and weaknesses, a multimodel forecasting system such as has been recently developed by, for example, Muhammad et al. (2018), Samaniego et al. (2019), and Wanders et al. (2019), would ideally incorporate both methods.

Our finding of increasing predictability with event extremity suggests that individual forecasts contain useful information, which could be discarded as noise by statistical analyses. As was also noted by Viel et al. (2016) and Meißner et al. (2017), small forecast skill does not mean forecasts are not useful for decision-makers. Our results indicate that when a large fraction of the forecast ensemble is in the lower or upper tercile, the probability of the forecast being correct is high. This, therefore, is very important information for decision-makers in reservoir management in anticipation of extreme low-flow conditions.

Given the identical meteorological forcing, the differences between forecast systems found in our study are presumably caused by the schematization and/or parameterization of the hydrological models. By considering and comparing the relevant processes separately (e.g., groundwater flow, glacier melt) the reasons for these differences could be further explored. The representation of glacial melt, for example, is more sophisticated in E-HYPE compared to LISFLOOD (Arheimer et al. 2020), which might contribute to slightly higher skill in summer for E-HYPE, whereas the opposite is case for snowmelt. LISFLOOD includes, for example, snowmelt due to precipitation, a relevant process at lower altitudes (Burek et al. 2013), possibly causing slightly higher skills for EFAS in early spring. The absence of hydraulic processes in H-TESSSEL is likely to cause overall lower forecast skills. E-HYPE, which generally performs best in terms of forecast skill, is semidistributed and has the lowest effective spatial resolution. In modeling studies in the Rhine basin,

relatively simple, conceptual models were found to outperform more complex distributed models in reproducing streamflow at Lobith (Hurkmans et al. 2008; te Linde et al. 2008) because they are relatively easy to tune. This capability appears to translate to forecast skill: when forecast skill would be assessed at more locations representing smaller subcatchments, distributed models are more suitable and the higher resolution of LISFLOOD and the EFAS forecasts would probably enhance forecast skill. In our analyses, however, using weekly averages and the entire catchment, this is not the case. Fully disentangling these differences due to model formulation and resolution requires model output per component for all forecast systems, which was not feasible in the current study. Moreover, the meteorological input for E-HYPE was already corrected prior to our analyses, which further complicates a correct interpretation of the differences found.

We focused on one specific location in the Rhine basin. However, discharge at Lobith nearly integrates the entire Rhine basin, which covers, with 160 000 km², a substantial part of western Europe. In more spatially extensive studies, (e.g., Arnal et al. 2018), results for the Rhine appeared comparable to many other areas in Europe, suggesting that our results are applicable to other areas in Europe. In general, streamflow is more predictable in river systems with long memory due to snow processes, groundwater contribution and dampening from lakes and reservoirs and groundwater contribution, all of which apply to the Rhine, and less in arid climates with fast hydrological response to precipitation (Pechlivanidis et al. 2020). Our finding of the higher predictability of low-flow extremes would, in our view, also translate to other catchments with similar characteristics. This remains to be confirmed in a future study.

a. Limitations in experimental setup

To assess the employed forecasts, we have selected a number of skill metrics, some of which depend on specific threshold values. For instance, the BSS was calculated with quantile thresholds of 33% and 66% to be consistent with the terciles, which are commonly used in the dissemination of seasonal (streamflow) forecasts (e.g., ECMWF 2017; Arnal et al. 2018).

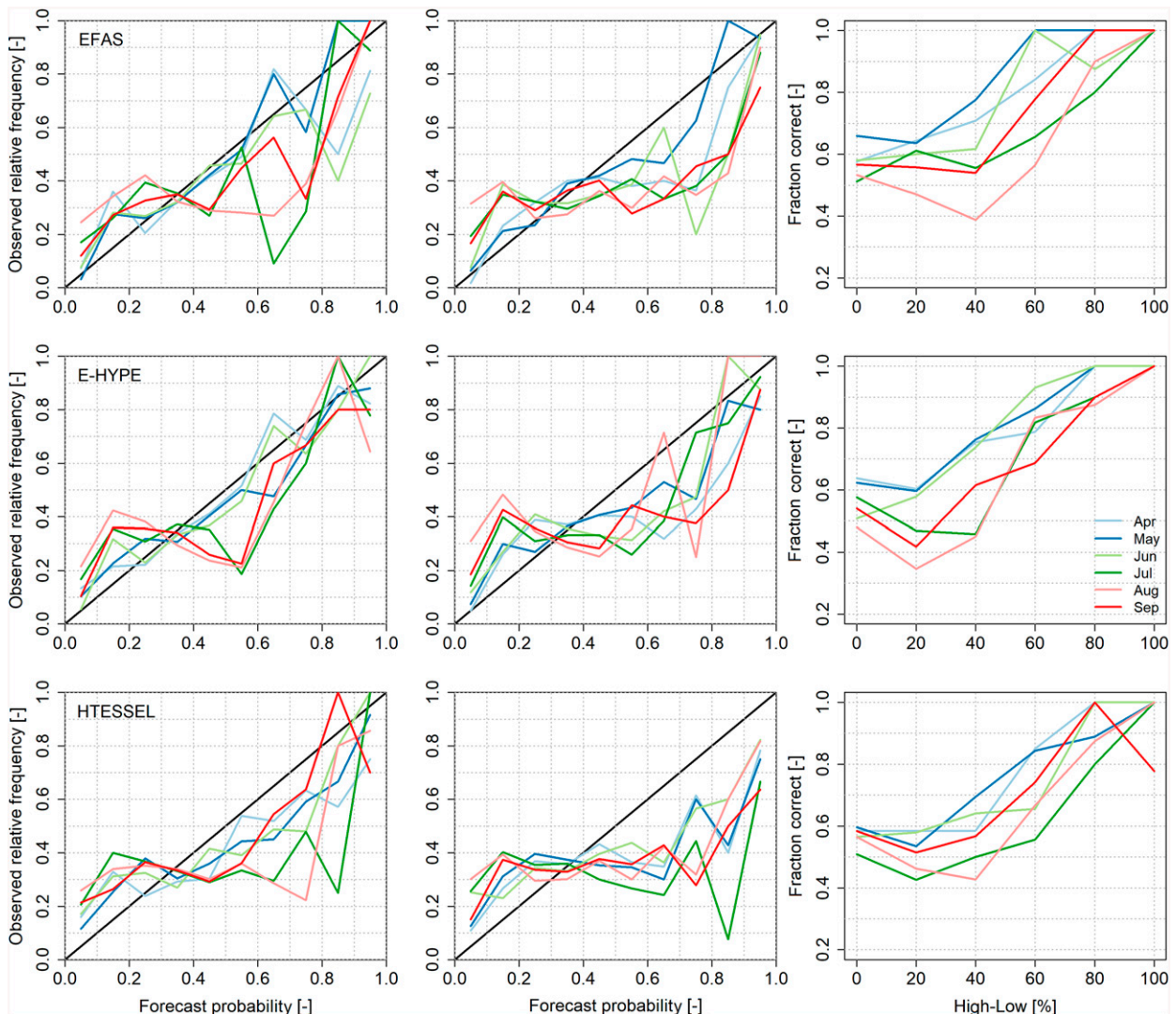


FIG. 13. Forecast reliability plots for streamflow based on EFAS, E-HYPE, and HTESSEL (left) not exceeding the 33% lower tercile value and (center) exceeding the 67% upper tercile for different calendar months, considering all lead times. (right) The fraction of correct forecasts as a function of the difference between the upper and lower terciles.

We also calculated the BSS based on thresholds of 10% and 20%. Results were similar, although they were more uncertain as the number of underexceedances of the threshold decreased. Related to this, is the extent of the hindcast dataset. With the period 1993–2015 available, and metrics computed for each calendar month and lead time, the number of data points is relatively small, as was also noted by Arnal et al. (2018), particularly when one assesses the extremes. When the reforecast dataset is extended, or combined with forecasts, the skill scores could become more stable. To optimize this number here, we have analyzed lead times in monthly intervals, with streamflow aggregated over weeks. Weekly aggregates are appropriate for typical low-flow events of the river Rhine (e.g., Hurkmans et al. 2010).

In this study, we focused on streamflow at Lobith as this location represents a (relatively) large catchment area. In the

current application we use a fixed relation with IJssel streamflow based on historical data (Fig. 1). In dry conditions, the fraction of Lobith streamflow routed over the river IJssel may be managed by weirs in the other branches (van Malde 1988). To accommodate for this, machine learning algorithms may be able to include this management aspect in the relationship (e.g., Suntaranont et al. 2020).

b. Moving forward and service implementation

The other components of the reservoir water balance apart from streamflow, i.e., precipitation, evaporation, and regional water extractions, depend on local meteorological conditions and are therefore less predictable. Ongoing research pertains to both improving the relationship between Rhine and IJssel streamflow and assessing how streamflow forecast skill propagates through the lake water balance in the lake level forecast skill.

The differences between the forecast systems, especially E-HYPE and EFAS, are difficult to disentangle without examining the stores and fluxes within the underlying hydrological models. Recently, a high-resolution, distributed model specific for the Rhine basin has been developed (Imhoff et al. 2020). This has a number of features that might increase forecast skill, such as a more detailed representation of groundwater and soil moisture and upstream surface water reservoirs. Producing seasonal forecasts based on this model and assessing whether forecast skill indeed improves are included in ongoing research.

Along with this future research a prototype of a (pre)operational system is being developed. Recently, much effort has been put into bridging the gap between science and operational use (e.g., van Den Hurk et al. 2016; Soares et al. 2018; Lavers et al. 2020; White et al. 2022). Skillful hydrological forecasts are the first step in successful operational use. In addition, the ability of end-users to interpret the ensemble probabilities and make optimal decisions is needed (Giuliani et al. 2020). To achieve this, operational systems and their interfaces are best created in cooperation between developers and end-users (Golding et al. 2019). Development of the prototype will, therefore, be in close cooperation between the developers and Rijkswaterstaat.

6. Conclusions

We compared various metrics of forecast skill for a number of streamflow forecast systems for the Rhine catchment, with a focus on low flows and monthly streamflow values. The investigated forecast systems were EFAS, E-HYPE, and HTESSSEL. All generate operational hydrological seasonal forecasts for the entirety of Europe, hence their usability for regional applications can be explored. All three systems are forced by ECMWF SEAS5 seasonal forecasts and mainly differ in the underlying hydrological model, while in the case of E-HYPE the meteorological forecasts are first bias corrected.

We show that streamflow forecast skill scores are high up to 4 months ahead in spring, when streamflow is dominated by snowmelt, while the length of the skillful period decreases to about 1–2 months in summer when streamflow is mainly driven by rainfall. We also noted a relationship between the absolute difference between tercile probabilities and skill. Moreover, both the forecast skill and the lead time with positive skill increase with the extremity of the hydrological event. From an operational perspective, when forecasts have a large difference between the tercile probabilities at long lead times, the probability of an upcoming anomalous event is relatively high, which is a relevant notion for water managers.

After postprocessing the streamflow forecasts using a bias correction, based on quantile mapping, the difference between the three systems is small. In addition, the categorization of streamflow into terciles reduced the difference between the forecast systems. Consequently, we concluded that all bias-corrected forecast systems were able to provide useful information for reservoir operations. That was not the case for raw forecasts, since bias correction proved to be essential. Depending on the season and lead time, there were differences between the forecast systems. To

combine the strengths of all models and accurately represent forecasted uncertainty, a multimodel ensemble is proposed for operations, in which depending on the season, one forecasting system could get more weight than the others. Finally, we note that the forecasts will be implemented in an operational reservoir management system and the usefulness (in terms of economic benefit and general decision-making) in managing reservoir water levels will be assessed.

Acknowledgments. This research was financially supported by NWO project SWM-EVAP (ALWTW.2016.049). The EFAS seasonal forecasts are produced by the EFAS computational center in support of the Copernicus Emergency Management Service (EMS) and Early Warning Systems (EWS) 198702. I.G.P. was partially funded by the EU Horizon 2020 project I-CISK (Innovating climate services through integrating scientific and local knowledge) under Grant Agreement 101037293. We thank the two reviewers for their constructive comments, as well as the handling editor.

Data availability statement. The ECMWF SEAS5 data were downloaded from the ECMWF Web API. Observed data are freely available from the referenced sources. The R-scripts that were used for the analyses and any other underlying datasets are available from the authors upon request.

REFERENCES

- Abudu, S., J. P. King, and T. C. Pagano, 2010: Application of partial least-squares regression in seasonal streamflow forecasting. *J. Hydrol. Eng.*, **15**, 612–623, [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000216](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000216).
- Arheimer, B., R. Pimentel, K. Isberg, L. Crochemore, J. C. M. Andersson, A. Hasan, and L. Pineda, 2020: Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation. *Hydrol. Earth Syst. Sci.*, **24**, 535–559, <https://doi.org/10.5194/hess-24-535-2020>.
- Arnal, L., H. L. Cloke, E. Stephens, F. Wetterhall, C. Prudhomme, J. Neumann, B. Krzeminski, and F. Pappenberger, 2018: Skillful seasonal forecasts of streamflow over Europe? *Hydrol. Earth Syst. Sci.*, **22**, 2057–2072, <https://doi.org/10.5194/hess-22-2057-2018>.
- Balsamo, G., A. Beljaars, K. Scipal, P. Viterbo, B. van den Hurk, M. Hirschi, and A. K. Betts, 2009: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *J. Hydrometeorol.*, **10**, 623–643, <https://doi.org/10.1175/2008JHM1068.1>.
- Bennett, J. C., Q. J. Wang, D. E. Robertson, A. Schepen, M. Li, and K. Michael, 2017: Assessment of an ensemble seasonal streamflow forecasting system for Australia. *Hydrol. Earth Syst. Sci.*, **21**, 6007–6030, <https://doi.org/10.5194/hess-21-6007-2017>.
- Berg, P., C. Donnelly, and D. Gustafsson, 2018: Near-real-time adjusted reanalysis forcing data for hydrology. *Hydrol. Earth Syst. Sci.*, **22**, 989–1000, <https://doi.org/10.5194/hess-22-989-2018>.
- Bierkens, M. F. P., and B. J. J. M. van den Hurk, 2007: Groundwater convergence as a possible mechanism for multi-year persistence in rainfall. *Geophys. Res. Lett.*, **34**, L02402, <https://doi.org/10.1029/2006GL028396>.

- , and L. P. H. van Beek, 2009: Seasonal predictability of European discharge: NAO and hydrological response time. *J. Hydrometeorol.*, **10**, 953–968, <https://doi.org/10.1175/2009JHM1034.1>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Buitink, J., and Coauthors, 2020: Anatomy of the 2018 agricultural drought in the Netherlands using in situ soil moisture and satellite vegetation indices. *Hydrol. Earth Syst. Sci.*, **24**, 6021–6031, <https://doi.org/10.5194/hess-24-6021-2020>.
- Burek, P. A., J. van der Knijff, and A. de Roo, 2013: LISFLOOD distributed water balance and flood simulation model: Revised user manual. JRC Tech. Rep. 78917, 139 pp., <https://doi.org/10.2788/24719>.
- Cornes, R. C., G. van der Schrier, E. J. van den Besselaar, and P. D. Jones, 2018: An ensemble version of the E-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.*, **123**, 9391–9409, <https://doi.org/10.1029/2017JD028200>.
- Crochemore, L., M.-H. Ramos, and F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, **20**, 3601–3618, <https://doi.org/10.5194/hess-20-3601-2016>.
- , —, and I. Pechlivanidis, 2020: Can continental models convey useful seasonal hydrologic information at the catchment scale? *Water Resour. Res.*, **56**, e2019WR025700, <https://doi.org/10.1029/2019WR025700>.
- Demirel, M. C., M. J. Booij, and A. Y. Hoekstra, 2015: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. *Hydrol. Earth Syst. Sci.*, **19**, 275–291, <https://doi.org/10.5194/hess-19-275-2015>.
- Diebold, F. X., and R. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.
- Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. Rodrigues, 2013: Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, <https://doi.org/10.1002/wcc.217>.
- ECMWF, 2017: SEAS5 user guide version 1.1. ECMWF Tech. Rep., 44 pp., https://www.ecmwf.int/sites/default/files/media/library/2017-10/System5_guide.pdf.
- Giuliani, M., L. Crochemore, I. Pechlivanidis, and A. Castelletti, 2020: From skill to value: Isolating the influence of end user behavior on seasonal forecast assessment. *Hydrol. Earth Syst. Sci.*, **24**, 5891–5902, <https://doi.org/10.5194/hess-24-5891-2020>.
- Golding, N., C. Hewitt, P. Zhang, M. Liu, J. Zhang, and P. Bett, 2019: Co-development of a seasonal rainfall forecast service: Supporting flood risk management for the Yangtze River basin. *Climate Risk Manage.*, **23**, 43–49, <https://doi.org/10.1016/j.crm.2019.01.002>.
- Greuell, W., W. H. Franssen, H. Biemans, and R. W. Hutjes, 2018: Seasonal streamflow forecasts for Europe – Part I: Hindcast verification with pseudo- and real observations. *Hydrol. Earth Syst. Sci.*, **22**, 3453–3472, <https://doi.org/10.5194/hess-22-3453-2018>.
- Harrigan, S., C. Prudhomme, S. Parry, K. Smith, and M. Tanguy, 2018: Benchmarking ensemble streamflow prediction skill in the UK. *Hydrol. Earth Syst. Sci.*, **22**, 2023–2039, <https://doi.org/10.5194/hess-22-2023-2018>.
- Hundecha, Y., B. Arheimer, C. Donnelly, and I. Pechlivanidis, 2016: A regional parameter estimation scheme for a pan-European multi-basin model. *J. Hydrol. Reg. Stud.*, **6**, 90–111, <https://doi.org/10.1016/j.ejrh.2016.04.002>.
- Hurkmans, R., W. Terink, R. Uijlenhoet, P. Torfs, D. Jacob, and P. A. Troch, 2010: Changes in streamflow dynamics in the Rhine basin under three high-resolution regional climate scenarios. *J. Climate*, **23**, 679–699, <https://doi.org/10.1175/2009JCLI3066.1>.
- Hurkmans, R. T. W. L., H. de Moel, J. C. J. H. Aerts, and P. A. Troch, 2008: Water balance versus land surface model in the simulation of Rhine River discharges. *Water Resour. Res.*, **44**, W01418, <https://doi.org/10.1029/2007WR006168>.
- Imhoff, R. O., W. J. van Verseveld, B. van Osnabrugge, and A. H. Weerts, 2020: Scaling point-scale (pedo)transfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrologic modeling: An example for the Rhine River. *Water Resour. Res.*, **56**, e2019WR026807, <https://doi.org/10.1029/2019WR026807>.
- Ionita, M., and V. Nagavciuc, 2020: Forecasting low flow conditions months in advance through teleconnection patterns, with a special focus on summer 2018. *Sci. Rep.*, **10**, 13258, <https://doi.org/10.1038/s41598-020-70060-8>.
- Jacob, D., and Coauthors, 2014: EURO-CORDEX: New high-resolution climate change projections for European impact research. *Reg. Environ. Change*, **14**, 563–578, <https://doi.org/10.1007/s10113-013-0499-2>.
- Johnson, S. J., and Coauthors, 2019: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.*, **12**, 1087–1117, <https://doi.org/10.5194/gmd-12-1087-2019>.
- Khanal, S., A. F. Lutz, W. W. Immerzeel, H. Vries, N. Wanders, and B. Hurk, 2019: The impact of meteorological and hydrological memory on compound peak flows in the Rhine river basin. *Atmosphere*, **10**, 171, <https://doi.org/10.3390/atmos10040171>.
- Lavers, D. A., and Coauthors, 2020: A vision for hydrological prediction. *Atmosphere*, **11**, 237, <https://doi.org/10.3390/atmos11030237>.
- Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.*, **99**, 14 415–14 458, <https://doi.org/10.1029/94JD00483>.
- Lindström, G., B. Johansson, M. Gardelin, and S. Bergström, 1997: Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.*, **201**, 272–288, [https://doi.org/10.1016/S0022-1694\(97\)00041-3](https://doi.org/10.1016/S0022-1694(97)00041-3).
- Lucatero, D., H. Madsen, J. C. Refsgaard, J. Kidmose, and K. H. Jensen, 2018: On the skill of raw and post-processed ensemble seasonal meteorological forecasts in Denmark. *Hydrol. Earth Syst. Sci.*, **22**, 6591–6609, <https://doi.org/10.5194/hess-22-6591-2018>.
- Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, <https://doi.org/10.1287/mnsc.22.10.1087>.
- Meißner, D., B. Klein, and M. Ionita, 2017: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in central Europe. *Hydrol. Earth Syst. Sci.*, **21**, 6401–6423, <https://doi.org/10.5194/hess-21-6401-2017>.
- Mens, M., B. Minnema, K. Overmars, and B. van den Hurk, 2021: Dilemmas in developing models for long-term drought risk management: The case of the National Water Model of the Netherlands. *Environ. Modell. Software*, **143**, 105100, <https://doi.org/10.1016/j.envsoft.2021.105100>.
- Muhammad, A., T. A. Stadnyk, F. Unduche, and P. Coulibaly, 2018: Multi-model approaches for improving seasonal ensemble streamflow prediction scheme with various statistical post-processing techniques in the Canadian prairie region. *Water*, **10**, 1604, <https://doi.org/10.3390/w10111604>.
- Panofsky, H., and G. Brier, 1958: *Some Applications of Statistics to Meteorology*. Pennsylvania State University, 224 pp.

- Pechlivanidis, I., L. Crochemore, J. Rosberg, and T. Bosshard, 2020: What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resour. Res.*, **56**, e2019WR026987, <https://doi.org/10.1029/2019WR026987>.
- Rakovec, O., L. Samaniego, V. Hari, Y. Markonis, V. Moravec, S. Thober, M. Hanel, and R. Kumar, 2022: The 2018–2020 multi-year drought sets a new benchmark in Europe. *Earth's Future*, **10**, e2021EF002394, <https://doi.org/10.1029/2021EF002394>.
- Ratri, D. N., K. Whan, and M. Schmeits, 2019: A comparative verification of raw and bias-corrected ECMWF seasonal ensemble precipitation reforecasts in Java (Indonesia). *J. Appl. Meteor. Climatol.*, **58**, 1709–1723, <https://doi.org/10.1175/JAMC-D-18-0210.1>.
- Rijkswaterstaat, 2018: Peilbesluit IJsselmeergebied (in Dutch). https://puc.overheid.nl/rijkswaterstaat/doc/PUC_162522_31/.
- Robertson, A., and F. Vitart, 2018: *Sub-seasonal to Seasonal Prediction: The Gap between Weather and Climate Forecasting*. Elsevier, 585 pp.
- Rutten, M., N. van de Giesen, M. Baptist, J. Icke, and W. Uijttewaai, 2008: Seasonal forecast of cooling water problems in the river Rhine. *Hydrol. Processes*, **22**, 1037–1045, <https://doi.org/10.1002/hyp.6988>.
- Samaniego, L., and Coauthors, 2019: Hydrological forecasts and projections for improved decision-making in the water sector in Europe. *Bull. Amer. Meteor. Soc.*, **100**, 2451–2472, <https://doi.org/10.1175/BAMS-D-17-0274.1>.
- Sánchez-García, E., J. Voces Aboy, B. Navascués, and E. Rodríguez Camino, 2019: Regionally improved seasonal forecast of precipitation through best estimation of winter NAO. *Adv. Sci. Res.*, **16**, 165–174, <https://doi.org/10.5194/asr-16-165-2019>.
- Scaife, A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, <https://doi.org/10.1002/2014GL059637>.
- , and Coauthors, 2019: Does increased atmospheric resolution improve seasonal climate predictions? *Atmos. Sci. Lett.*, **20**, e922, <https://doi.org/10.1002/asl.922>.
- Soares, M. B., M. Alexander, and S. Dessai, 2018: Sectoral use of climate information in Europe: A synoptic overview. *Climate Serv.*, **9**, 5–20, <https://doi.org/10.1016/j.cliser.2017.06.001>.
- Suntaranont, B., S. Aramkul, M. Kaewmoracharoen, and P. Champrasert, 2020: Water irrigation decision support system for practical weir adjustment using artificial intelligence and machine learning techniques. *Sustainability*, **12**, 1763, <https://doi.org/10.3390/su12051763>.
- Sutanto, S. J., and H. A. Van Lanen, 2022: Catchment memory explains hydrological drought forecast performance. *Sci. Rep.*, **12**, 2689, <https://doi.org/10.1038/s41598-022-06553-5>.
- te Linde, A. H., J. C. J. H. Aerts, R. T. W. L. Hurkmans, and M. Eberle, 2008: Comparing model performance of two rainfall-runoff models in the Rhine basin using different atmospheric forcing data sets. *Hydrol. Earth Syst. Sci.*, **12**, 943–957, <https://doi.org/10.5194/hess-12-943-2008>.
- Thrasher, B., E. P. Maurer, P. B. Duffy, and C. McKellar, 2012: Bias correcting climate model simulated daily temperature extremes with quantile mapping. *Hydrol. Earth Syst. Sci.*, **16**, 3309–3314, <https://doi.org/10.5194/hess-16-3309-2012>.
- van den Hurk, B., P. Siegmund, and A. Klein Tank, Eds., 2014: KNMI14: Climate change scenarios for the 21st century – A Netherlands perspective. Tech. Rep. WR2014-01, 120 pp., <https://cdn.knmi.nl/knmi/pdf/bibliotheek/knmipubWR/WR2014-01.pdf>.
- van Den Hurk, B. J. J. M., and Coauthors, 2016: Improving predictions and management of hydrological extremes through climate services: www.imprex.eu. *Climate Serv.*, **1**, 6–11, <https://doi.org/10.1016/j.cliser.2016.01.001>.
- van Malde, J., 1988: Hydrology of disasters. *Hydrological Aspects of Pollution of the Rhine*, James and James (Science Publishers), 164–171.
- Viel, C., A.-L. Beaulant, J.-M. Soubeyroux, and J.-P. Céron, 2016: How seasonal forecast could help a decision maker: An example of climate service for water resource management. *Adv. Sci. Res.*, **13**, 51–55, <https://doi.org/10.5194/asr-13-51-2016>.
- Wanders, N., S. Thober, R. Kumar, M. Pan, J. Sheffield, L. Samaniego, and E. F. Wood, 2019: Development and evaluation of a pan-European multimodel seasonal hydrological forecasting system. *J. Hydrometeorol.*, **20**, 99–115, <https://doi.org/10.1175/JHM-D-18-0040.1>.
- Waterman, R. E., R. Misdorp, and A. Mol, 1998: Interactions between water and land in The Netherlands. *J. Coast. Conserv.*, **4**, 115–126, <https://doi.org/10.1007/BF02806503>.
- Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, <https://doi.org/10.1175/MWR3280.1>.
- Wetterhall, F., H. C. Winsemius, E. Dutra, M. Werner, and E. Pappenberger, 2015: Seasonal predictions of agrometeorological drought indicators for the Limpopo basin. *Hydrol. Earth Syst. Sci.*, **19**, 2577–2586, <https://doi.org/10.5194/hess-19-2577-2015>.
- White, C., and Coauthors, 2022: Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Amer. Meteor. Soc.*, **103**, E1448–E1472, <https://doi.org/10.1175/BAMS-D-20-0224.1>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- Wood, A. W., and D. P. Lettenmaier, 2006: A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bull. Amer. Meteor. Soc.*, **87**, 1699–1712, <https://doi.org/10.1175/BAMS-87-12-1699>.
- Yang, W., J. Andréasson, L. Phil Graham, J. Olsson, J. Rosberg, and F. Wetterhall, 2010: Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrol. Res.*, **41**, 211–229, <https://doi.org/10.2166/nh.2010.004>.
- Yossef, N. C., R. van Beek, A. Weerts, H. Winsemius, and M. F. P. Bierkens, 2017: Skill of a global forecasting system in seasonal ensemble streamflow prediction. *Hydrol. Earth Syst. Sci.*, **21**, 4103–4114, <https://doi.org/10.5194/hess-21-4103-2017>.
- Zhu, H., M. C. Wheeler, A. H. Sobel, and D. Hudson, 2014: Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Mon. Wea. Rev.*, **142**, 1556–1569, <https://doi.org/10.1175/MWR-D-13-00222.1>.