https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions

The suitability of a hybrid framework including data driven approaches for hydrological forecasting

Sandra M. Hauswirth[1], Marc F. P. Bierkens[1,2], Vincent Beijk[3], and Niko Wanders[1]

[1]Utrecht University, Department of Physical Geography, Princetonlaan 8a, Utrecht, The Netherlands
[2]Deltares, Daltonlaan 600, 3584 BK Utrecht, The Netherlands
[3]Rijkswaterstaat, Water, Verkeer en Leefomgeving, Griffioenlaan 2, Utrecht, The Netherlands

Correspondence: Sandra M. Hauswirth (s.m.hauswirth@uu.nl)

Abstract. Hydrological forecasts are important for operational water management and near future planning, even more so in light of increased occurrences of extreme events such as floods and droughts. Having a flexible forecasting framework that can deliver this information in fast and computational efficient manner is critical. In this study, the suitability of a hybrid forecasting framework, combining data-driven approaches and seasonal (re)forecasting information to predict hydrological variables was explored. Target variables include discharge and surface water levels for various stations at national scale with the Netherlands as focus. Five different ML models, ranging from simple to more complex and trained on historical observations of discharge, precipitation, evaporation and sea water levels, were run with seasonal (re)forecast data (EFAS and SEAS5) of these driver variables in a hindcast setting. The results were evaluated using the evaluation metrics Anomaly Correlation Coefficient (ACC), Continuous Ranked Probability (Skill) Score (CRPS and CRPSS), and Brier Skill Score (BSS) in comparison to a climatological reference hindcast. Aggregating results of all stations and ML models revealed that the hindcasting framework outperformed the climatological reference forecasts by roughly 60% for discharge predictions (80% for surface water level predictions). Skilful prediction for the first lead month, independently of initialization month, can be made for discharge. The skill extends up to 2-3 months for spring months due to snow melt dynamics captured in the training phase of the model. Surface water levels hindcasts showed similar skill and skilful lead times. While the different ML models showed differences in performance during a testing and training phase using historical observations, running the ML framework in a hindcast setting showed only minor differences between the models, which is attributed to the uncertainty in seasonal forecasts. However, despite being trained on historical observations, the hybrid framework used in this study shows similar skilful predictions as previous large scale forecasting systems. With our study we show that a hybrid framework is able to bring location specific skilful seasonal forecast information with global seasonal forecast inputs. At the same time our hybrid approach is flexible and fast, and as such a hybrid framework could be adapted to make it even more interesting to water managers and their needs, for instance a part of a fast model-predictive control framework.

1 Introduction

Forecasting in combination with local system knowledge plays an important role in increasing the readiness for imminent extreme events such as floods and droughts. Especially over the last few years, where the effects and impacts of climate change

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions
Open Access
EGU

have become more and more distinct with an increasing recurrence of extreme events require adaptive planning based on skilful forecasts. For instance, knowledge of upcoming water surplus or shortage is important to limit damages to infrastructure and impacts on society during floods as well as to increase and sustain the water availability prior but also during droughts.

Over the past years, platforms of open source forecasting services and data sets that deliver meteorological and hydrological predictions have increased. These data sets can differ in leadtime (e.g. short- to medium range, sub-seasonal, and seasonal time scales) and include uncertainty by consisting of various numbers of ensemble members. Examples of open source seasonal forecasting system are the operational European Flood Awareness System (EFAS, Thielen et al. (2009); Arnal et al. (2018)) and the Global Flood Awareness System (GloFAS, Alfieri et al. (2013)), as well as ECMWF latest seasonal forecasting system SEAS5 Johnson et al. (2019). Forecasting systems like these are run by large scale, physically based models (e.g. Lisflood Van Der Knijff et al. (2010); De Roo et al. (2000) in case of EFAS), which require a lot of information regarding parametrization, can be slow and computational intensive as well as require large data storage facilities. Another example of forecasting systems facing similar challenges are multimodel ensemble systems, which combine several general circulation models and hydrological models (Wanders et al., 2019; Samaniego et al., 2019). One of the earliest approaches for streamflow forecasting includes ensemble streamflow prediction (ESP, originally named extended streamflow prediction Day (1985)), where a physically based model is run with observed historical meteorological time series from multiple years but initialized with the current hydrological conditions.

Even though continuously improved in terms of ease of use and interoperability, the main challenges of handling such large, computational and data intensive systems remain. Furthermore, Samaniego et al. (2019) highlighted that in case forecasting systems are used for decision making, prediction horizons, spatial scales, model choices, storage and computational requirements and reported variables can limit the applicability of forecasting systems to local water management. We hypothesize that incorporating data-driven approaches to support seasonal forecasting systems can be beneficial not only in terms of reducing computational requirements but also their flexibility and data use. Especially, if the forecasting system can be kept simple, for example regarding input variables or algorithms incorporated, the threshold of applying it on various spatial and temporal scales would be even further lowered and more readily applicable. In this study we explore these opportunities by incorporating data driven approaches in a seasonal forecasting framework, combining both local and global information.

Data driven approaches, including machine learning (ML) models, have been explored and tested out increasingly in hydrological assessments over the last few years (Shen, 2018; Shen et al., 2021), either as standalone models or also in hybrid-settings (coupled with physically based models) (Kratzert et al., 2018; Koch et al., 2019). ML can be used for any spatial and temporal scale study, as long as there is sufficient data available for training and validation. Besides using local observations and remote sensing information an upcoming trend is also to incorporate knowledge based learning (Koch et al., 2021), where ML models are also trained with information provided by physically based model or in hybrid model setups. ML has shown to be promising in simulating hydrological variables such as discharge and groundwater levels but also in contributing to operational water management.

However, most of the previous research focused on successfully simulating past observations or current hydrological states but incorporating ML in a seasonal forecasting framework has only scarcely been explored in the hydrological field. Work by

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions

Hunt et al. (2022) being one of the most recent examples, where LSTMs were explored in a hybrid forecasting setup to predict discharge for short term scale. A substantial issue as to using ML for seasonal forecasting is often the limited amount of samples for training. This is often resolved by including long climate model simulations for training as an example, however depending on the scale and resolution these might not always be ideal for more local studies. Nevertheless, if sufficient local data are available, it is worthwhile investigating how one can exploit the assets of ML for seasonal forecasting (limited complexity of forecasting setup, computational demand and handling of large data amounts) in order to create efficient and flexible operational settings and increase the support for water management. This can be especially useful for floods but also drought occurrences, where local information has to be available and updated within a short time frame (floods) and changes in water management planning have to be reassessed both ahead and in time of an event to optimise water availability (droughts).

To be able to have such a data-driven forecasting system that can support water managers as an example, a first step is to build a framework that can be explored in a so called hindcasting experiment, where the forecasting framework is tested based on historical observations. Once this is successful, the forecasting framework could be switched to seasonal forecasting. The aim of this study is to explore the first step: test the suitability of a data-driven forecasting framework in a hindcast experiment. The framework will build on an existing ML model framework based on a previous study by Hauswirth et al. (2021) in combination with (re)forecast information as input variables. Hauswirth et al. (2021) tested out different ML algorithms, ranging from simple to more complex methods, for simulating hydrological variables at national scale based on a simple input dataset including water management aspects. In this study we want to test the suitability of these models based on their previous performance regarding discharge and surface water levels.

The ML models (trained on historical observations), will be run with seasonal (re)forecasting data replacing the previous input dataset consisting of discharge, precipitation, evaporation and sea water level observations. Running the models with seasonal (re)forecasting information creates an ensemble of the target variables for each station of interest. These ensembles will be analysed and compared with historical observations to assess the skill of the ML framework for seasonal forecasting for general hydrological predictions but also extreme events such as droughts by computing different skill scores common to evaluate seasonal forecasting frameworks. Furthermore, the benefits but also challenges of such a simple setup will be explored and listed to assess whether the framework is suitable for current practices and whether it opens up possibilities for future assessments.

In the following sections the study's approach will be laid out, followed by an evaluation of the performance of the hindcast experiment by assessing skill scores both for general and droughts. Thereafter, the findings will be summarized, discussed and put into a bigger context of the field, followed by the main conclusions.

2 Material and Methods

This section is divided into subsections covering the general concept of the hindcast framework used in this study, the seasonal (re)forecast data and its preparation as well as the skill scores used to assess the forecasting skills of the data-driven framework.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

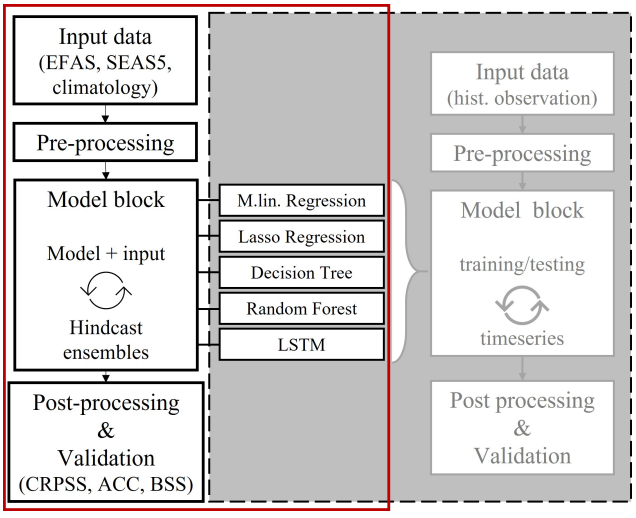

Figure 1. Schematization of hindcast framework highlighted in red frame, grey box indicating model framework previously developed by Hauswirth et al. (2021) and used as base for the model block.

2.1 Hindcast framework

The hindcasting framework used in this study can be described as a simple three block system including: pre-processing of the input data, a main model block including a selection of ML models, as well as post-processing of the target variable and skill score calculation (Fig. 1). The main model block, which is based on the ML model study of Hauswirth et al. (2021), was run on seasonal (re)forecasting information, which has first undergone the pre-processing block. The hindcast results were evaluated based on different skill scores included in the post-processing block to assess how skilful the data-driven framework is in hindcasting historical observations. The spatial and temporal setting of the hindcast experiment is focusing on the Netherlands and the time period 1993-2018. Target variables include discharge and surface water level of selected stations throughout the observation network of the National Water Authority.

2.2 Data and pre-processing

The input dataset based on seasonal (re)forcasting information covers the period 1993 - 2018, including a lead time of 7 months (215 days) and consisting of 25 ensemble members (50 from 2017 on). The input dataset replaces the previously used historic data of the ML framework, which consists of a simple set of variables including discharge, precipitation, evaporation, and sea level observations at specific locations of the case study area. The seasonal (re)forecasting data was obtained from the open source platform Copernicus Climate Data Store. Meteorological information was taken from the SEAS5 and includes precipitation, u and v wind components, temperature, surface net solar radiation, and mean sea level pressure (Johnson et al., 2019). The latter three variables were used to calculate the Makkink reference potential evaporation (de Bruin and Lablans, 1998). Seasonal reforecasting information on discharge was obtained from the European Flood Awareness System (EFAS)

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions

(Thielen et al., 2009; Wanders et al., 2014; Arnal et al., 2018). For the sea level data, the water level for the historic period was simulated by using the pytide python module. The difference between the observed and pytide predicted sea level fluctuations (including only tidal components) was computed to get the anomalies. A simple multi-linear regression model was then used to compute the final sea level ensemble set based on the u and v wind speed and the previously computed sea level anomalies.

The input dataset was processed in a similar manner as done by Hauswirth et al. (2021), for example by including the lagged time series of every input variable. In this case the seasonal (re)forecasting data in a first step was bias corrected using cumulative density function (CDF) matching approach before extending the input variable by incorporating the lagged times series approach (Wanders et al., 2014).

For every ML model (representing a station of interest) the input data set after data pre-processing finally consists of a set of ensembles, including ensemble members of all input variables. The models were run with every set of ensemble members (e.g. input dataset based on first ensemble members of discharge, precipitation, evaporation and sea level information).

2.2.1 Data-driven model setup

We applied here a recently developed ML model framework by Hauswirth et al. (2021). This framework has a focus on a simple setup using only readily available input data to simulate target variables such as discharge, surface water level, surface water temperature, and groundwater levels for several stations at a national scale in the Netherlands. Furthermore, the framework is able to incorporate the influence of water management aspects. For every station of interest a ML model was trained on historic observations of the target variable and the input dataset, consisting of the five variables: precipitation, evaporation, Rhine discharge, Meuse discharge, sea-level observations. Different ML methods were tested, ranging from simple to more complex methods including: Multi-linear Regression (MReg), Lasso Regression (Lasso), Decision Tree (DT) and Random Forest (RF), as well as LSTM. For more information we refer to Hauswirth et al. (2021).

For this study the pretrained models were rerun based on a prepared seasonal (re)forecast input dataset. The input dataset is made up of the same variables as in the previous study but taken from seasonal (re)forcasting datasets such as EFAS and SEAS5. The models were not retrained, so the input data was used for an extensive validation of the simulation of seasonal forecasting skill. Using the pretrained models has the benefit of saving computational time, which would have otherwise been needed for testing and training the models. Secondly, this study aims to test the suitability of this ML framework for seasonal forecasting in an operational setting. In such an operational setting one would like to keep a consistent modelling framework that has been validated on an extensive hindcast archive. On the other hand, not retraining the models on ensemble datasets limits the potential improvement the model could experience by seeing forecasting data in the training phase. As we want to test the suitability of the developed ML framework for hindcasting, we are putting a focus on the pretrained model in combination with the seasonal (re)forecast input dataset. This allows us to test the performance of the models based on information that the models have definitely not seen before. Running the model based on a seasonal (re)forecast input dataset, consisting of several ensemble members, creates an ensemble of time series for the target variables discharge and surface water levels. These ensemble simulations were then analysed by computing frequently used skill scores.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions

2.3 Skill scores

To evaluate the performance of the data-driven hindcast framework, the target variables were compared to observations as bi-weekly and monthly averages, time scales which would also be of interest to water managers for mid- to longterm planning. In this study, the hindcasts (including 7 months of lead time) will be addressed by their lead time, where lead one equals the first months of the forecast in which it was initiated, lead two equals the second month after initialization, etc. The performance was evaluated by different skill scores, shedding a light on various aspect of hindcast skills, for example overall performance, accuracy and reliability. These skill scores are common in the forecasting community and include: Continuous Ranked Probability (Skill) Score (CRPS and CRPSS), Brier (Skill) Score (BS and BSS), as well as Anomaly Correlation Coefficient (ACC).

2.3.1 CRPS and CRPSS

The CRPS, which is one of the most common used evaluation benchmarks used in ensemble forecasting studies (Pappenberger et al., 2015), was used to assess the overall performance of the hindcasting framework. It compares the differences in the hindcast and observed Cumulative Distribution Functions (CDF) and ranges from 0 to infinity. The lower the computed score, the better the performance of the hindcasting framework (Arnal et al., 2018; Pappenberger et al., 2015). Equation 1 taken from Hersbach (2000) (where P(x) is the cumulative density function of the hindcast and P_a observation probability) is computed) was used and the CRPS computed over all ensemble members for each lead day of every hindcast before aggregating it to other temporal scales. As a skilful benchmark (baseline) we also compare the hindcast framework with a forecasts based on the historical distribution of observations. In other words, for each forecast day we look at the historical observations for that day and select values for all the years of this historical observations to generate an observation based climatological hindcast ensemble. Both the hindcast CRPS and the baseline CRPS were used to compute the CRPSS (Eq. 2).

$$CRPS = CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx \tag{1}$$

$$CRPSS = 1 - \frac{CRPS_{hindcast}}{CRPS_{ref}} \tag{2}$$

2.3.2 BS and BSS

To determine the accuracy and the performance of the hindcasts for simulating high and low flow periods the BS and BSS can be used. To assess these specific categories, thresholds can be defined e.g. the lowest 20^{th} percentile data to account for droughts. This allows one to analyse events which are either higher or lower than the usual observations for a given month (Candogan Yossef et al., 2017). The BS is calculated by Eq. 3, where N equals the number of hindcasting instances, f and o are the hindcast and observed probability of exceeding a threshold, respectively (Candogan Yossef et al., 2017). Score values range between 0 and 1, whereas 0 is indicating the best performance.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{3}$$

$$BSS = 1 - \frac{BS_{hindcast}}{BS_{ref}} \tag{4}$$

Furthermore the BSS (Eq. 4) can be used to compare the accuracy and performance of the hindcasting framework compared to a reference system. We used a set of random created hindcasts, corrected for seasonality based on the climatology, as reference forecasting system.

2.3.3 ACC

To measure the quality of the hindcasting framework, the Anomaly Correlation Coefficient (ACC) between the hindcasts and the observations is computed. This was done using the ensemble mean of each forecast for ever lead day of the hindcasts, before gathering all years for every month to calculate the ACC per lead day and finally aggregating it to different temporal scales. The ACC helps to verify the hindcast and observed anomalies, compared to the normal correlation where seasonality can influence the calculation results. Therefore, the ACC can also be seen as skill score in comparison with the climate. The ACC score ranges from -1 to 1, with 1 representing a perfect correlation between observations and forecast. For representation purposes the significance level was computed based on the number of observational years (in general) and only stations with less than 10 missing observation months were considered (same criteria of station selection was used for the other scores).

3 Results

The results will be presented such that first an overview of the general performance of the hindcast framework for one target variable will be given. Subsequently the focus will be directed towards one model (Random Forest, RF) and an example station to provide a more in-depth insight into the different evaluation scores and differences in temporal resolution. The scores were calculated for all the initialization months of the hindcast and for different temporal resolutions (daily, weekly, monthly). For demonstration we highlight the performance of the hindcast framework for selected months providing weekly to monthly scores, which are temporal scales of interest for long term water management decisions. Further background information on evaluation score results based on different initialization months, temporal scales, target variables and ML models can be found in the Appendix.

3.1 General performance

To obtain understanding of the overall performance of the hindcast framework, the CRPSS aggregated over all hindcasts for all stations and all ML models was computed. This was done by computing all the individual daily CRPSS results of all the hindcasts of every station and methods, before aggregating the individual CRPSS scores to different temporal scales

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

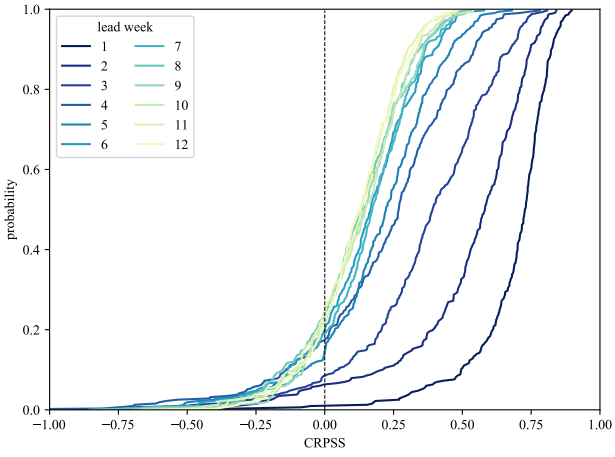

Figure 2. CDFs of weekly CRPSS shown for different lead weeks with CRPSS being aggregated over all models and station for discharge hindcasts. CRPSS is decreasing with increasing lead weeks but even up to 12 weeks roughly 60% of all stations and models show a better performance than the reference hindcasts.

(weekly in Fig. 2). The CDFs for the first 12 lead weeks can be seen in Fig. 2, with CRPSS ranging from -1 to 1, with values above 0 indicating the hindcast framework outperforming the climatological reference. As expected, the CRPSS decreases with increasing lead time (with CDF lines moving up and the zero line being crossed earlier) and naturally converges after 7 weeks. However, up to lead week 12 roughly 60% of all stations and models show a better performance than the climatological reference. Even better results can be seen for surface water levels (Fig. A1), where up to 80% show a better performance. Even though the separate evaluation scores can vary slightly between target variable and station (locations) (shown later on), the overall hindcasting framework shows a positive tendency compared to hindcasts solely based on climatology.

During the analysis of the evaluation scores for all the different ML model hindcasts only a minor differences between the models are noticeable, which is seen throughout most stations along the main river network, especially for discharge hindcasts (Fig. 3). Minor differences between methods are observed for surface waterlevel depending on the station location (Fig. A2), where for the example stations shown the more simpler methods show a slightly better skill. The minor differences are likely due to the limited impact of the model selection compared to the inherent uncertainty (represented by the ensemble spread) in the dynamical meteorological and hydrological forecast data. In addition, the high temporal aggregation (monthly) and post-processing of the results before calculating the different evaluation scores, smoothing out the original differences in hindcasts results reduce the differences in performance between models.

Shifting the focus from the whole hindcast framework to a more detailed exploration of the evaluation metrics, the following paragraphs focus on results of one ML model and later on one example station. As hindcast results indicate that the differences between the models are minor, we will focus on the RF model, which previously already showed a promising performance Hauswirth et al. (2021). Figure 4 shows the weekly Anomaly Correlation Coefficient (ACC) for the discharge hindcasts at various stations (each represented by a pie chart) throughout the Netherlands for initialization months a) January, b) April, c)

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

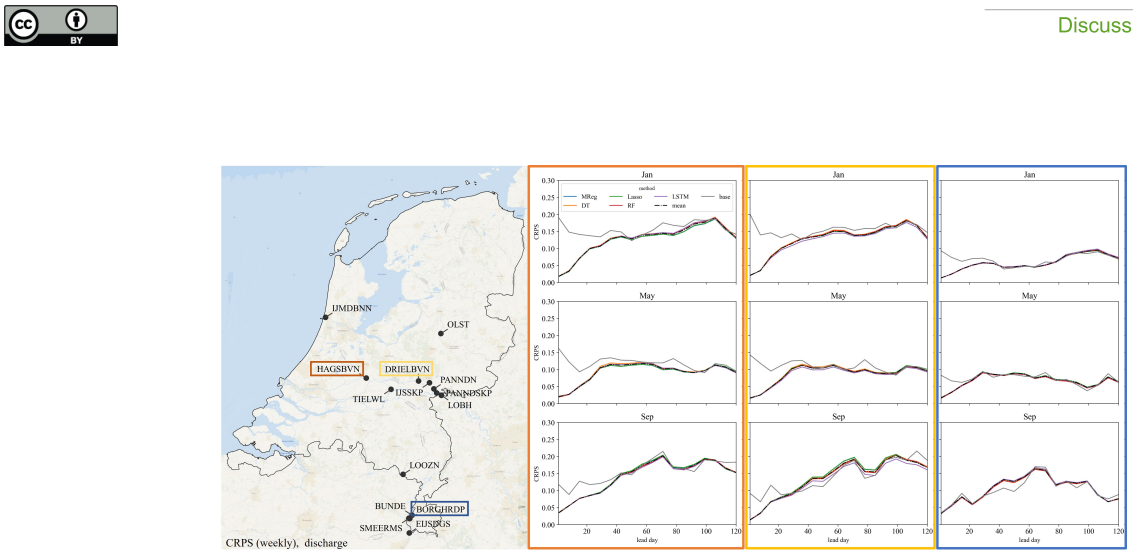

Figure 3. Overview of weekly CRPS values for month January, May and September. Different ML model scores, average of ML models score (dashed line) as well as climatological reference (grey) are shown for three discharge stations. For most months shown and first few lead weeks, the CRPS of the hindcast framework shows a lower score than the climatological reference. However, only minor differences between the ML models were observed. Maps were created using QGIS (QGIS Development Team, 2022), HCMGIS plugin and basemap data from ESRI Ocean (Sources: Esri, GEBCO, NOAA, National Geographic, DeLorme, HERE, Geonames.org, and other contributors) and GADM database.

July and d) October. The ACC values per week are indicated by the pie slices arranged clockwise and their colour, dark blue indicating a high correlation coefficient while light yellow slices show weeks with a lower coefficient (note only significant values are shown). Looking at the results for the different months in Fig. 4 indicate that for all months shown, the ACC decreases with increasing lead weeks. However, for all months the first few weeks (min. 3-4 weeks) show a high and significant score. This can be observed for all stations along the main river networks, both Rhine and Meuse, while stations which are located at smaller streams or channels, which are strongly influenced by water management, can be more challenging (e.g. station close to the sea which is located at a shipping channel).

The observations from countrywide ACC analysis are supported by a more detailed analysis for station Hagenstein Boven located on the Rhine river network, roughly in center of the Netherlands and influenced by water management. We clearly see that there are differences in ACC per lead and initialization month related to the initialization month of the hindcast and the length of the forecast (Fig. 5). In addition, it shows that the differences on the ACC in temporal aggregation from daily, weekly to monthly temporal scale have a minor impact and that the skill assessment is robust. Significant ACC values can be observed throughout the first lead month for all initialization months. For early spring and summer months (March-July), significant ACC values for discharge predictions can be seen until two months in advance, in all temporal aggregation levels. The increase in significant lead time for the early spring months (March and April) is due to the snow melt dynamics in upstream catchment that were captured in the model training period (done prior to this study) and the physical model inputs from the EFAS system at the Lobith and Eijsden stations. The observation of more significant ACC values during the spring months due to the snow melt dynamic can be found throughout the stations along the Rhine, and less pronounced for the stations along the Meuse. Discharge predictions from late summer on show lower ACC values, likely due to the lower predictability in atmospheric

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions

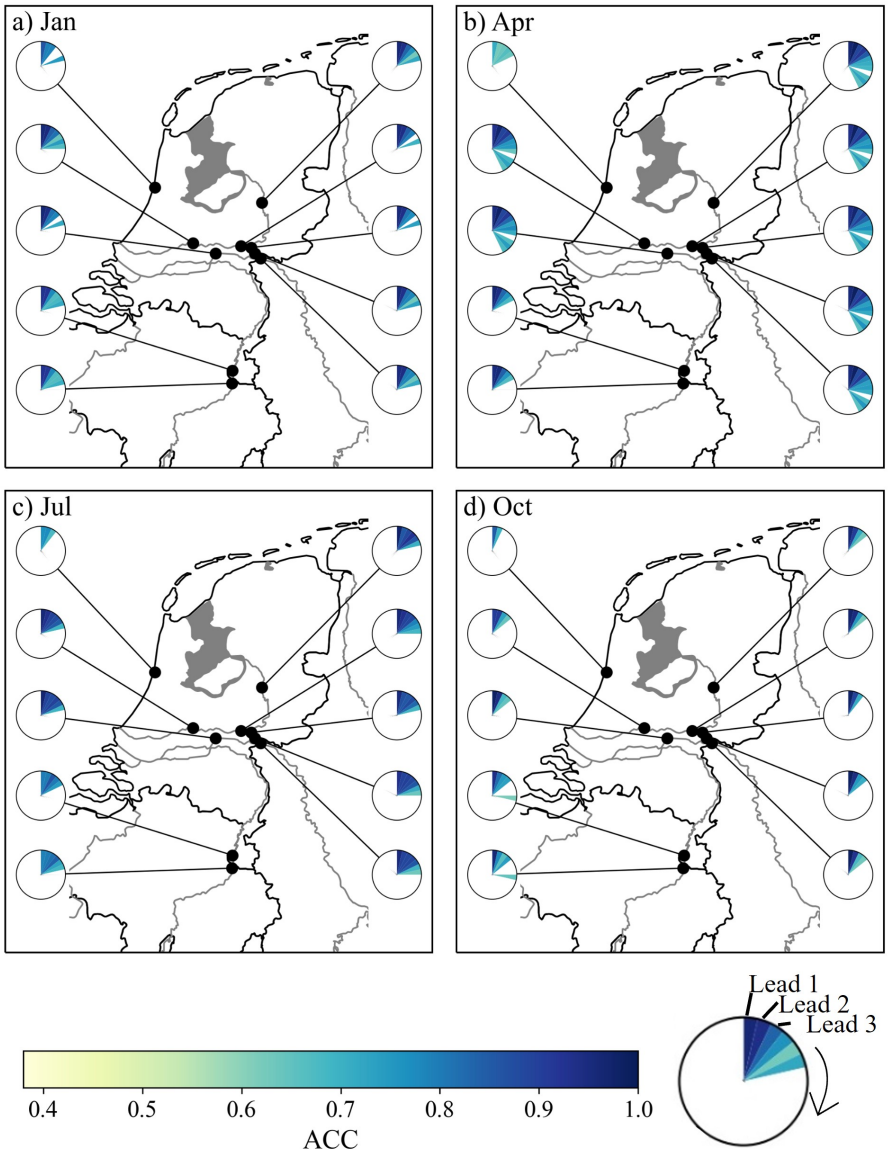

Figure 4. Anomaly Correlation Coefficient (ACC, weekly) for a) January, b) April, c) July and d) October for discharge hindcasts computed by the RF model showing the results for different stations (limited selection for visual purpose) in the national monitoring network. Maps were created using the python package Cartopy (Elson et al., 2022), which uses basemap data from Made with Natural Earth and © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

weather patterns and reduced water storage in highly predictable stores like snow and groundwater. Unrealistic long lead times with significant values are likely due to lower observation records that can occur throughout the years, despite the selection of stations with limited missing records. Overall, ACC values for the discharge hindcasts show that hindcast anomalies are

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

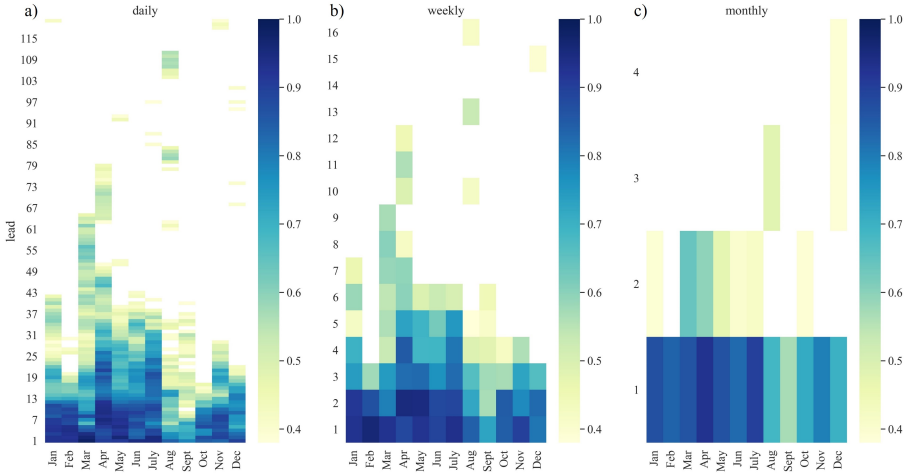

Figure 5. ACC results for station Hagenstein Boven over different temporal scales (a) daily, b) weekly , c) monthly) using the RF model. Only significant values are shown.

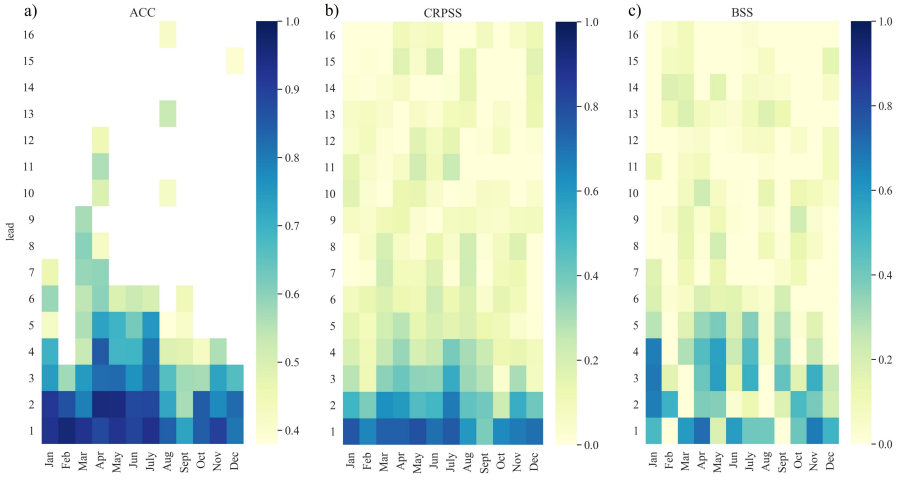

Figure 6. Heatmaps of weekly ACC (same as in Fig. 5), CRPSS and BSS for example station Hagenstein Boven. a) ACC only significant values are shown, b) CRPSS and c) BSS indicating a good performance (in dark blue and values above 0) compared to climatological reference.

240 captured well for lead times up to one or two months for all initialization months compared to the observed anomalies for a complex station like Hagenstein Boven, which is affected by water management and upstream water reallocation. Furthermore, the hindcast framework was able to capture the general snow melt dynamic in early spring, resulting in significant values up to a lead time of two months at the onset of summer.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions

To check the robustness of the results we also analyzed the forecast performance with the CRPSS and BSS. The CRPSS was computed to assess the general performance of the hindcast framework. Figure 6 represents the weekly CRPSS values for the same example station Hagenstein Boven in the center panel. The heatmap giving an overview of the skill score throughout the year with values ranging from 0 to 1, with values above 0 representing lead weeks where the hindcast framework outperforms the climatological reference. Similar to the ACC, the first few lead weeks show consistently good performance, indicating that the hindcast spread is close to the one of observations, while with increasing lead time the CRPSS decreases. This pattern can be observed by stations along the main river networks.

3.2 Hydrological extremes - low flows

To assess the hindcast frameworks capability of simulating low flow events, the BSS was computed using a threshold for the lowest 20^{th} percentile of discharge observations and hindcasts. Figure 6 represents the weekly BSS on the right panel again for the example station Hagenstein Boven. Blue tiles on the heatmap indicate lead times where the hindcast framework outperforms the climatological reference strongly. The BSS confirms the earlier findings and shows the same trend of increased performance in the first lead weeks, additional skill of several weeks is found for early spring and early summer months. However, tiles with lower performance throughout long lead periods, late summer and winter months can be spotted for this station. Some of these weeks appear to be more difficult to predict compared to early months in the year. This is likely due to unequal distribution of low flow occurrences throughout the year: where during summer months low flows can be more common and therefore chances to not fully capturing every event are higher, the low flows during winter are less common and captured relatively well with the snow melt dynamic as seen in previous scores. The described observations for the example station are in line with findings at other stations along the Rhine river network and less pronounced along the Meuse.

3.3 Difference between target variables and models

The result section so far has been focusing on the discharge hindcasts to be able to focus on the different evaluation scores in more detail. However, the hindcast framework was also used to hindcast surface water levels. Surface water level prediction skill shows a similar pattern and trend, regarding ACC and CRPSS results for stations along the main river network (Fig. A3 and Fig. A4 for an example station). Yet, surface water levels seem to be more challenging to hindcast with the evaluation scores being slightly lower especially for stations in smaller channels and further away from the main river network. Similar findings are found for BSS, where the performance of the hindcast for low flow periods was tested. As can be expected, stations which are not along the main river network and located downstream the main input variables (Rhine at Lobith, Meuse at Eijsden) show lower skills in capturing low flow periods compared to stations closer to input variables and the Rhine river. While ACC and BSS show a slightly lower performance, the CRPSS ranges are in the same range as for discharge hindcasts.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions

4 Discussion

In this study we tested the suitability of ML models for seasonal predictions of several hydrological target variables at local scales throughout the Netherlands. This framework incorporated ML models over varying complexity, ranging from Multilinear Regression, Lasso Regression, Decision Tree to Random Forests and LSTM. While the methods have shown differences in their performance during training and testing phase on historical observations (especially their ability to reproduce extreme events Hauswirth et al. (2021)), interestingly applying the same subset of models on seasonal (re)forecasting information did not lead to large differences in model performance. We hypothesize that this is caused by the large uncertainty in the meteorological and hydrological input data, that outweighs the relative difference in performance by the different ML algorithms. In addition, the minor differences seen between the ML algorithms in the original hindcasts were further smoothed out while calculating the evaluation scores on different temporal scales. While the ML models were previously trained on direct observations, the seasonal (re)forecasting information from SEAS5 and EFAS introduces additional uncertainty from their forecasting system. We deliberately decided not the retrain the ML models on the forecasting information, as this more closely mimics the normal operational setting where an already trained model is used to produce forecasts. However, this provides an additional challenge, as we add another source of potential uncertainty as the ML models might not be well tuned to the forecast information. Retraining the models would also open up the opportunity for overfitting the ML models on the forecast data, which is something that should be avoided. Therefore, we preferred to use the more realistic operational scenario and use ML models trained on historic observations only, over a setup that uses ML models specifically trained on forecast data. Assessing the approach of additionally retraining the models for different cases, e.g. focus on extreme events or climate change trends are opportunities for future projects.

We extended our runs including water management, in line with the approach previously explored by Hauswirth et al. (2021). However, incorporating variables that represent water management settings in the ML models lead to negligible improvement and the improved performance as seen in the previous study could not be detected. We think that the uncertainty included in the seasonal (re)forecast input data is having a larger influence than the one of added water management information and therefore the strength of incorporating the additional information as seen in the previous study could not be observed.

The evaluation metrics show that for the majority of the discharge and surface water level stations, skilful predictions for the first lead month one can be made. For early spring and summer month the skill increases up to 2-3 months, due to the snow melt dynamic being captured by the models in their training phase and the presence of this signal in the seasonal reforecasts of the discharge at Lobith used as input to the ML models. The skilful prediction for the first few lead months are comparable with other studies which have evaluated physically based systems (Wanders et al., 2019; Arnal et al., 2018; Girons Lopez et al., 2021; Pechlivanidis et al., 2020). However, contrary to the large scale physically based forecasting systems, hybrid frameworks such as the one presented in this study show to skilfully forecast target variables at specific locations which would not be feasible and at a fraction of the computation demand. This can be interesting for water management needs at smaller scale or scenario analysis. Besides the fast running times of the models an additional benefit for the current framework is the input data set, which can be easily replaced by other input sources regarding precipitation, evaporation, discharge and surface

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022


Hydrology and Earth System Sciences
Discussions

water level as it was done in this study with EFAS and SEAS5 data. This framework however only focuses on time series at existing stations and therefore does not address the challenge of predicting at ungauged basins. However, recent advances in deep learning methods show that forecasting at ungauged sites may be a possibility if auxiliary geographically distributed variables (elevation, soil, river network topology) are incorporated (Kratzert et al., 2019).

Similar to Hunt et al. (2022) we show that a hybride forecasting system can provide added benefits compared to physical forecasting system. In addition to Hunt et al. (2022) this work confirms that the benefits of hybride forecasting can also be obtained for long-term forecasting. In this study we also show that these hybrid forecasting systems have the ability to provide more local information compared to large-scale physically based systems. As with other models, using ML models in hydrology comes with benefits and drawbacks. While the data availability can be a limiting factor for effectively train a model, the flexibility and low computational demand compared to large scale physically based models is an advantage. We think that with the right data available, ML models like the ones used here can easily be (re)trained for more specific studies and cases as well. Additional training on low flow periods for example could enhance drought predictions while incorporating climate change aspects and land use could help to assess future trends regarding water availability under increased human influence.

5 Conclusions

In this study we explored the suitability of a hybrid hindcasting framework, combining data-driven approaches and seasonal (re)forecasting information to predict hydrological variables locally for multiple stations at national scale for the Netherlands. Different ML models, previously trained on historical observations, were run with a simple input data set based on forecast data from EFAS and SEAS5 and evaluated using the evaluation metrics Anomaly Correlation Coefficient (ACC), Continuous Ranked Probability (Skill) Score (CRPS and CRPSS), and Brier Skill Score (BSS). The hindcast framework's skill was compared to the skill of a climatological reference hindcast. Aggregating the hindcasts of all stations and ML models revealed that the hindcasting framework was outperforming the climatological reference forecast by roughly 60% and 80% for discharge and surface water level hindcasts. ACC results further show that independently of the discharge prediction's initialization month, a skilful prediction for the first lead month can be made. For spring months the skill extends up to 2-3 months due to stronger link to snow melt dynamic and temperature related impacts on the hydrological cycle that were captured in the training phase of the model. CRPSS and BSS show a similar pattern of skilful predictions for the first few lead weeks compared to the climatological reference forecasts. Skilful discharge predictions are particularly observed along the main river networks, Rhine and Meuse, which can be linked to the close proximity of the discharge input variables. This distribution of performance is also observed for surface water level hindcasts. We also observed that the difference between different ML models in the hindcast results are only minor, contrary to the differences observed when reproducing historical timeseries. This reduction in differences in performance between ML models is attributed to the relatively large uncertainties in seasonal (re)forecast data, reducing the relative impact of the model uncertainty in the total hindcast uncertainty. Even though the current hindcast framework is trained on historical observations, the hybrid framework used in this study shows similar skilful predictions as previous large scale forecasting systems. With the focus on creating a hindcast framework that is simple in its setup, fast and also locally applicable,

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

challenges that can come with large scale operational forecasting systems for local users can be lowered. In addition, the ML hindcast framework also significantly reduces the computation demand and allows decision makers to explore more options and better quantify forecast uncertainty using a variety of ML models and inputs. Adapting the framework to special interests, e.g. droughts or climate change trends, by retraining the original ML models for specifically this purpose could further increase its performance. We conclude that the ML framework as developed in this study provide a valuable way forward, to making seasonal (re)forecast information more accessible to local and regional decision makers in the field of operational water management. In this study we purposely used publicly available seasonal forecast information which is globally available. This allows us to deploy this framework around the world and potentially provide relevant forecasting information for water managers and decision makers outside of the study area.

Data availability. Seasonal forecasting and (re)forecasting data was acquired over the Copernicus Climate Data Store (SEAS5 and EFAS).

Author contributions. Conceptualization of this research has been done by NW, MB and VB. Data acquisition and Analysis was performed by SH. Discussion, Writing was done by SH, NW, MB and VB, while SH took the lead in writing.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. SH acknowledges funding from the Cooperate Innovation Program and the Department of Water, Transport and Environment at the Dutch National Water Authority, Rijkswaterstaat. NW acknowledges funding from NWO 016.Veni.181.049.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

355 **Appendix A: Results**

A1 General performance

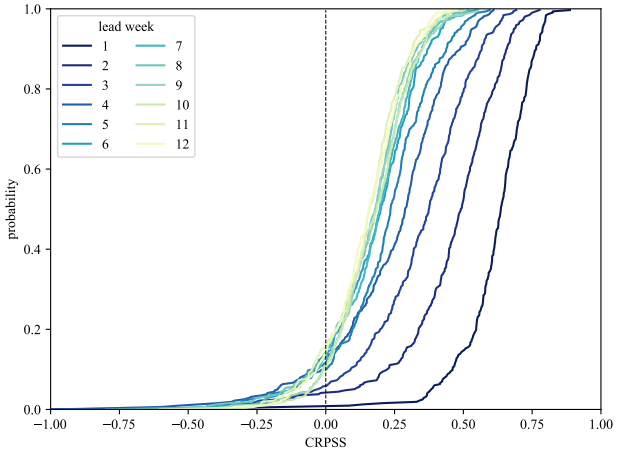

Figure A1. CDFs of weekly CRPSS shown for different lead weeks with CRPSS being aggregated over all models and stations for surface water level hindcasts.

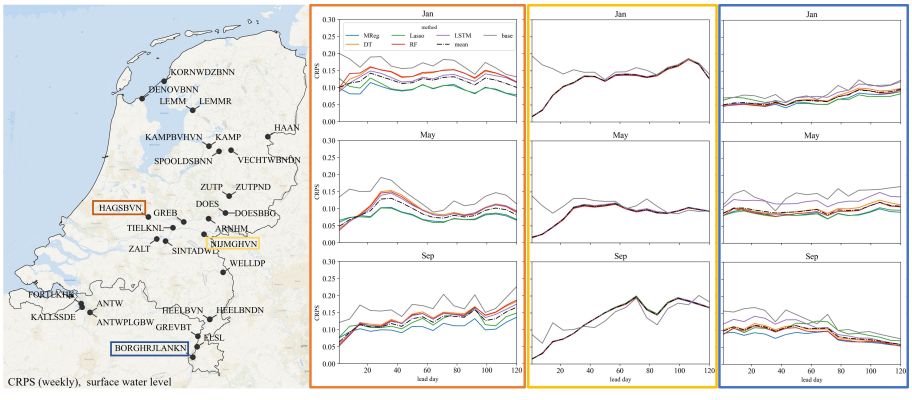

Figure A2. Overview of weekly CRPS scores for month January, May and September. Different ML model scores, average of ML models score (dashed line) as well as climatological reference (grey) are shown for three surface water level stations. Maps were created using QGIS (QGIS Development Team, 2022), HCMGIS plugin and basemap data from ESRI Ocean (Sources: Esri, GEBCO, NOAA, National Geographic, DeLorme, HERE, Geonames.org, and other contributors) and GADM database.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions

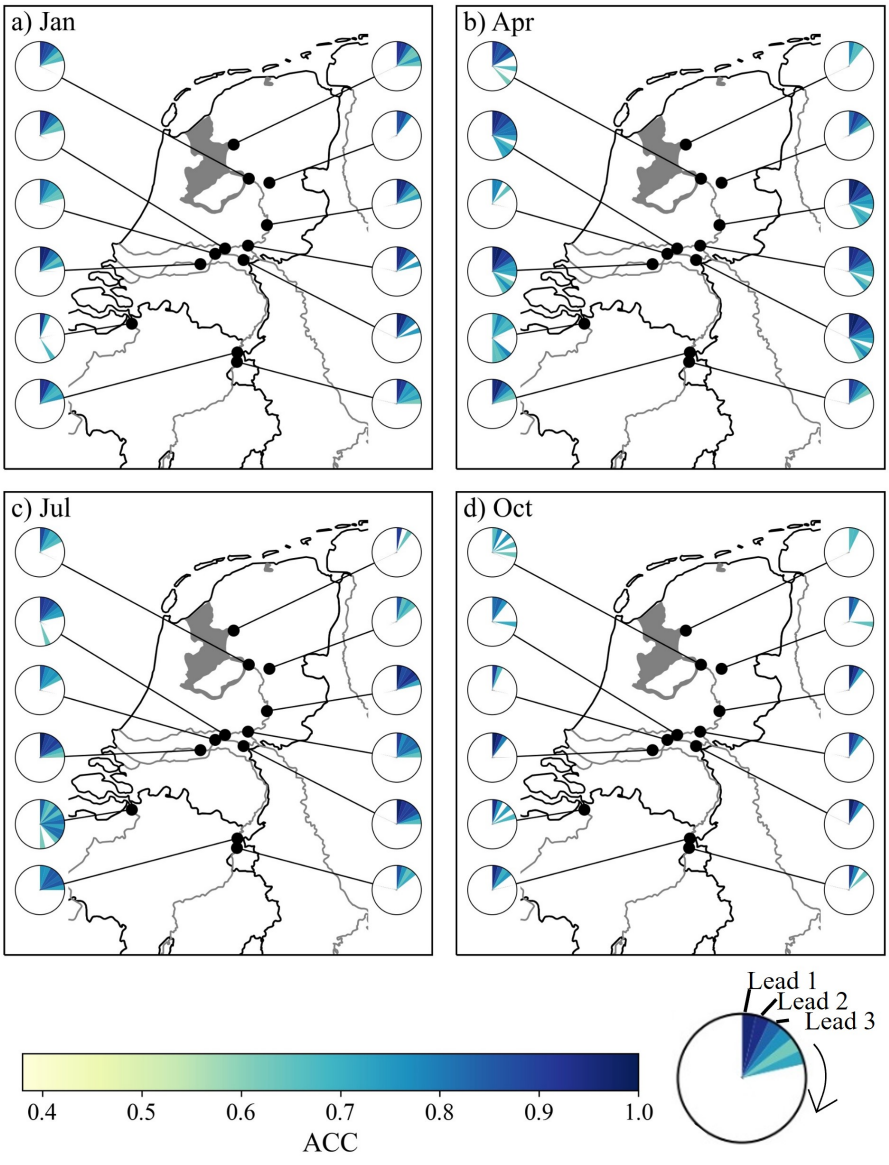

Figure A3. Anomaly Correlation Coefficient (ACC, weekly) for a) January, b) April, c) July and d) October for surface water level hindcasts computed by the RF model showing the results for different stations (limited selection for visual purpose) in the national monitoring network. Maps were created using the python package Cartopy (Elson et al., 2022), which uses basemap data from Made with Natural Earth and © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

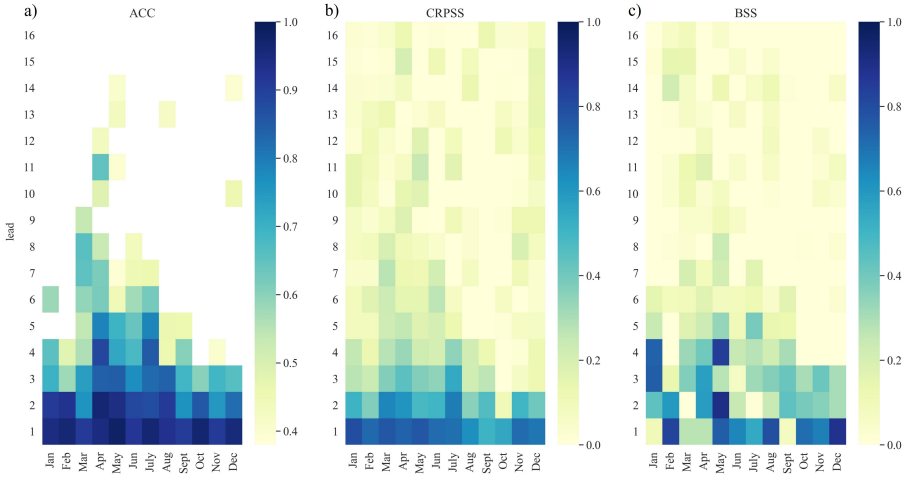

Figure A4. Overview of weekly evaluation scores for surface water level hindcasts. a) ACC overview throughout the year for example station Nijmegen, b) CRPSS and c) BSS heatmaps.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Hydrology and Earth System Sciences
Discussions

References

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS – global ensemble streamflow forecasting and flood early warning, Hydrology and Earth System Sciences, 17, 1161–1175, https://doi.org/10.5194/hess-17-1161-2013, 2013.

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, Hydrology and Earth System Sciences, 22, 2057–2072, https://doi.org/10.5194/hess-22-2057-2018, 2018.

Candogan Yossef, N., van Beek, R., Weerts, A., Winsemius, H., and Bierkens, M. F. P.: Skill of a global forecasting system in seasonal ensemble streamflow prediction, Hydrology and Earth System Sciences, 21, 4103–4114, https://doi.org/10.5194/hess-21-4103-2017, 2017.

Day, G. N.: Extended Streamflow Forecasting Using NWSRFS, Journal of Water Resources Planning and Management, 111, 157–170, https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157), 1985.

de Bruin, H. A. R. and Lablans, W. N.: Reference crop evapotranspiration determined with a modified Makkink equation, Hydrological Processes, 12, 1053–1062, https://doi.org/10.1002/(SICI)1099-1085(19980615)12:7<1053::AID-HYP639>3.0.CO;2-E, 1998.

De Roo, A. P. J., Wesseling, C. G., and Van Deursen, W. P. A.: Physically based river basin modelling within a GIS: the LISFLOOD model, Hydrological Processes, 14, 1981–1992, https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1099-1085%2820000815/30%2914%3A11/12%3C1981%3A%3AAID-HYP49%3E3.0.CO%3B2-F, 2000.

Elson, P., de Andrade, E. S., Lucas, G., May, R., Hattersley, R., Campbell, E., Dawson, A., Raynaud, S., scmc72, Little, B., Snow, A. D., Donkers, K., Blay, B., Killick, P., Wilson, N., Peglar, P., lbdreyer, Andrew, Szymaniak, J., Berchet, A., Bosley, C., Davis, L., Filipe, Krasting, J., Bradbury, M., Kirkham, D., stephenworsley, Clément, Caria, G., and Hedley, M.: SciTools/cartopy: v0.20.2, https://doi.org/10.5281/zenodo.5842769, 2022.

Girons Lopez, M., Crochemore, L., and Pechlivanidis, I. G.: Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden, Hydrology and Earth System Sciences, 25, 1189–1209, https://doi.org/10.5194/hess-25-1189-2021, 2021.

Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N.: The potential of data driven approaches for quantifying hydrological extremes, Advances in Water Resources, 155, 104 017, https://doi.org/10.1016/j.advwatres.2021.104017, 2021.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hunt, K. M. R., Matthews, G. R., Pappenberger, F., and Prudhomme, C.: Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States, Hydrology and Earth System Sciences Discussions, pp. 1–30, https://doi.org/10.5194/hess-2022-53, publisher: Copernicus GmbH, 2022.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H., and Monge-Sanz, B. M.: SEAS5: the new ECMWF seasonal forecast system, Geoscientific Model Development, 12, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019, 2019.

Koch, J., Berger, H., Henriksen, H. J., and Sonnenborg, T. O.: Modelling of the shallow water table at high spatial resolution using random forests, Hydrology and Earth System Sciences, 23, 4603–4619, https://doi.org/10.5194/hess-23-4603-2019, 2019.

https://doi.org/10.5194/hess-2022-89
Preprint. Discussion started: 31 March 2022

Koch, J., Gotfredsen, J., Schneider, R., Troldborg, L., Stisen, S., and Henriksen, H. J.: High Resolution Water Table Modeling of the Shallow Groundwater Using a Knowledge-Guided Gradient Boosting Decision Tree Model, Frontiers in Water, 3, 701 726, https://doi.org/10.3389/frwa.2021.701726, 2021.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resources Research, 55, 11 344–11 354, https://doi.org/10.1029/2019WR026065, 2019.

Pappenberger, F., Ramos, M., Cloke, H., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, Journal of Hydrology, 522, 697–713, https://doi.org/10.1016/j.jhydrol.2015.01.024, 2015.

Pechlivanidis, I. G., Crochemore, L., Rosberg, J., and Bosshard, T.: What Are the Key Drivers Controlling the Quality of Seasonal Streamflow Forecasts?, Water Resources Research, 56, https://doi.org/10.1029/2019WR026987, 2020.

QGIS Development Team: QGIS Geographic Information System, QGIS Association, https://www.qgis.org, 2022.

Samaniego, L., Thober, S., Wanders, N., Pan, M., Rakovec, O., Sheffield, J., Wood, E. F., Prudhomme, C., Rees, G., Houghton-Carr, H., Fry, M., Smith, K., Watts, G., Hisdal, H., Estrela, T., Buontempo, C., Marx, A., and Kumar, R.: Hydrological Forecasts and Projections for Improved Decision-Making in the Water Sector in Europe, Bulletin of the American Meteorological Society, 100, 2451–2472, https://doi.org/10.1175/BAMS-D-17-0274.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2019.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resources Research, 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018.

Shen, C., Chen, X., and Laloy, E.: Editorial: Broadening the Use of Machine Learning in Hydrology, Frontiers in Water, 3, https://www.frontiersin.org/article/10.3389/frwa.2021.681023, 2021.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, Hydrology and Earth System Sciences, 13, 125–140, https://doi.org/10.5194/hess-13-125-2009, 2009.

Van Der Knijff, J. M., Younis, J., and De Roo, A. P. J.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189–212, https://doi.org/10.1080/13658810802549154, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658810802549154, 2010.

Wanders, N., Karssenberg, D., de Roo, A., de Jong, S. M., and Bierkens, M. F. P.: The suitability of remotely sensed soil moisture for improving operational flood forecasting, Hydrology and Earth System Sciences, 18, 2343–2357, https://doi.org/10.5194/hess-18-2343-2014, 2014.

Wanders, N., Thober, S., Kumar, R., Pan, M., Sheffield, J., Samaniego, L., and Wood, E. F.: Development and Evaluation of a Pan-European Multimodel Seasonal Hydrological Forecasting System, Journal of Hydrometeorology, 20, 99–115, https://doi.org/10.1175/JHM-D-18-0040.1, 2019.