

# Exploring the Use of European Weather Regimes for Improving User-Relevant Hydrological Forecasts at the Subseasonal Scale in Switzerland

ANNIE Y.-Y. CHANG<sup>a,b</sup>, KONRAD BOGNER<sup>a</sup>, CHRISTIAN M. GRAMS<sup>c</sup>, SAMUEL MONHART<sup>d</sup>,  
DANIELA I. V. DOMEISEN<sup>b,e</sup> AND MASSIMILIANO ZAPPA<sup>a</sup>

<sup>a</sup> Swiss Federal Institute WSL, Birmensdorf, Switzerland

<sup>b</sup> ETH Zurich, Zurich, Switzerland

<sup>c</sup> Institute of Meteorology and Climate Research (IMK-TRO), Department Troposphere Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

<sup>d</sup> MeteoSwiss, Federal Office of Meteorology and Climatology, Locarno Monti, Switzerland

<sup>e</sup> University of Lausanne, Lausanne, Switzerland

(Manuscript received 29 December 2021, in final form 6 June 2023, accepted 9 June 2023)

**ABSTRACT:** Across the globe, there has been an increasing interest in improving the predictability of subseasonal hydro-meteorological forecasts, as they play a valuable role in medium- to long-term planning in many sectors, such as agriculture, navigation, hydropower, and emergency management. However, these forecasts still have very limited skill at the monthly time scale; hence, this study explores the possibilities for improving forecasts through different pre- and postprocessing techniques at the interface with a Precipitation–Runoff–Evapotranspiration Hydrological Response Unit Model (PREVAH). Specifically, this research aims to assess the benefit of European weather regime (WR) data within a hybrid forecasting setup, a combination of a traditional hydrological model and a machine learning (ML) algorithm, to improve the performance of subseasonal hydro-meteorological forecasts in Switzerland. The WR data contain information about the large-scale atmospheric circulation in the North Atlantic–European region, and thus allow the hydrological model to exploit potential flow-dependent predictability. Four hydrological variables are investigated: total runoff, baseflow, soil moisture, and snowmelt. The improvements in the forecasts achieved with the pre- and postprocessing techniques vary with catchments, lead times, and variables. Adding WR data has clear benefits, but these benefits are not consistent across the study area or among the variables. The usefulness of WR data is generally observed for longer lead times, e.g., beyond the third week. Furthermore, a multimodel approach is applied to determine the “best practice” for each catchment and improve forecast skill over the entire study area. This study highlights the potential and limitations of using WR information to improve subseasonal hydro-meteorological forecasts in a hybrid forecasting system in an operational mode.


**KEYWORDS:** Climate classification/regimes; Hydrology; Operational forecasting; Machine learning; Ensembles


## 1. Introduction

There is an increasing interest in improving the predictability of subseasonal (i.e., weekly to monthly) hydro-meteorological forecasts, as they play a valuable role in medium- to long-term planning in many sectors, such as agriculture, navigation, hydro-power production, and hazard warning (Anghileri et al. 2019; Arnal et al. 2018; Bogner et al. 2018; Hwang et al. 2019; Monhart et al. 2018; White et al. 2017, 2022). Numerous meteorological services around the world currently perform operational meteorological forecasts at the subseasonal scale (Buizza et al. 2005; Vitart et al. 2017), which opens up the possibility of performing predictions of hydrological variables for a subseasonal forecast horizon. However, along with other important drivers

such as initial hydrological conditions (Girons Lopez et al. 2021; Pechlivanidis et al. 2020), the skill of subseasonal hydrological forecasts relies greatly on the quality of the meteorological input data (Jörg-Hess et al. 2015). The skill of long-range meteorological forecasting itself remains low at this time scale in central Europe under the influence of several teleconnections (Domeisen et al. 2015).

As the current subseasonal meteorological models still exhibit systematic biases, many statistical techniques have been developed to postprocess the meteorological forecasts (Vitart et al. 2017). Given that meteorological data are input to hydrological models, this processing technique is one approach to improve the hydrological forecasts. This approach is hereby referred to as “preprocessing” from the perspective of the hydrological model. Monhart et al. (2018) investigated the effect of two different bias correction methods: a mean debiasing method and a quantile mapping method on temperature and precipitation, where the quantile mapping method performed better for both variables. In the same study, the lead-time-dependent predictability of extended-range (i.e., S2S) temperature and precipitation forecasts was assessed in detail for 1637 ground stations in Europe. Through a weekly aggregated verification study, they showed that the forecast skill diminishes after 1 week for precipitation and after 2–3 weeks for temperature. In a subsequent study (Monhart et al. 2019), the same methodology was applied to the

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

 Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-21-0245.s1>.

Corresponding author: Annie Y.-Y. Chang, [annie.chang@env.ethz.ch](mailto:annie.chang@env.ethz.ch)

DOI: 10.1175/JHM-D-21-0245.1

© 2023 American Meteorological Society. This published article is licensed under the terms of a Creative Commons Attribution 4.0 International (CC BY 4.0) License



meteorological forecasts, and these preprocessed forecasts were used to drive a hydrological model for three selected catchments in Switzerland. It showed that even low skill in the meteorological S2S predictions could translate into useful skill in hydrological predictions, extending the skillful forecast horizon. However, forecast skill highly depends on the characteristics of the catchments, where snow-dominated catchments have a stronger positive response to improved meteorological forecast. Furthermore, [Bogner et al. \(2018\)](#) focused on extended-range hydrological prediction skill for 307 catchments in Switzerland and assessed hydrological variables as well as areal precipitations. It was found that, despite low precipitation forecast skill, the hydrological predictions can still obtain decent skill, particularly for the baseflow for catchments in the Swiss midlands. Since previous studies have explored the influence of meteorological variables' predictability on hydrological variables, we focus only on the predictability of the hydrological output in this study.

Forecasts at longer lead times are associated with high uncertainty, and one way to make the forecasts useful for the relevant users is to categorize forecasts in terciles. Terciles can be simply defined as three categories with thresholds set by the 33rd and 66th percentile of the climatological means. The normal conditions fall between the 33rd and 66th percentile, while anything lower and higher is considered below and above the normal conditions, respectively. For example, a tercile temperature forecast at three weeks lead time can be "colder than normal," "normal," or "warmer than normal" with respect to the mean climatological temperature of the same week in the past. Such a tercile approach in meteorological forecasts has already been implemented by [MeteoSwiss \(2021c\)](#), and several other studies (e.g., [Hamill et al. 2004](#); [Tippett et al. 2007](#)), but the use of terciles on hydrological forecasts beyond streamflow as suggested by this present study is one of the first (e.g., [Arnal et al. 2018](#); [Liu et al. 2019](#); [Delorit et al. 2017](#); [Sahu et al. 2017](#)). With the ensemble forecasts in this study, we can compute the probabilities of each tercile, making it possible to communicate the confidence of our forecasts.

Machine learning (ML) applications in environmental sciences started in the 1990s ([Heish 2009](#); [Papacharalampous and Tyrakis 2022](#)), and they have been gaining more attention and popularity in recent years due to the increasing computing power. [Heish \(2009\)](#) gave a thorough overview of many applications of ML techniques in environmental sciences, ranging from satellite data processing, weather prediction, and hydrological forecasting to ecological modeling and beyond. On the other hand, for subseasonal and seasonal forecasts, the resources at national forecast centers have mostly been allocated to dynamical models over the past two decades. However, studies have shown improved skill when applying ML techniques. "Hybrid forecast," an approach combining dynamical models and statistical techniques, was recommended for subseasonal to seasonal forecasts ([Cohen et al. 2019](#)). In a review of different hybrid forecasting systems, [Slater et al. \(2023\)](#) summarizes the strengths of such systems, including bias minimization, the ability to combine different predictability sources at varying forecasting horizons, and their speed as well as operational efficiency. Among these strengths, [Slater et al. \(2023\)](#) also shines light on a ML model's ability to integrate large datasets compared to physics-based models in the way that large datasets

can be useful for a ML model, whereas a physics-based model tends to saturate its ability to adapt with limited data. An ensemble system consisting of two nonlinear regression models using customized algorithms significantly improved the skill of the operational U.S. Climate Forecast System, showing improvements of 40%–50% for temperature and 129%–169% for precipitation ([Hwang et al. 2019](#)). Forecast skill on seasonal time scales almost doubled when combining a regression model and a dynamical model ([Dobrynin et al. 2018](#)). Streamflow forecasts were improved by postprocessing with a quantile regression model and an artificial neural network model ([Bogner et al. 2016](#)). Five different ML techniques were applied to predict energy consumption and production in Switzerland between January 2015 and October 2018, and the results also demonstrated the possibility of skill improvements ([Bogner et al. 2019](#)).

Previously, [Bogner et al. \(2022\)](#) successfully trained ML models with a Gaussian process (GP) algorithm to postprocess the outputs of the Precipitation–Runoff–Evapotranspiration Hydrological Response Unit (HRU) Model (PREVAH) for Switzerland. Two variables (runoff and soil moisture) were aggregated into terciles and investigated on daily and weekly time scales for a 32-day forecasting horizon. The same aforementioned preprocessing method by [Monhart et al. \(2018\)](#) was also applied and compared with the postprocessing performance of the raw forecast. [Bogner et al. \(2022\)](#) explored the potential of postprocessing tercile hydrological forecasts using a ML algorithm and thereby laid a strong foundation for this study.

Building upon the findings of [Bogner et al. \(2022\)](#), linking large-scale weather regimes (WRs) with local hydrological events is the key exploratory and novelty part of this study. The rationale behind including WR data is that WRs contain information about the large-scale atmospheric circulation in the entire North Atlantic–European region in an aggregated way, and thus allows to exploit potential flow-dependent predictability. The U.K. Met Office uses eight coarse WRs to describe the large-scale conditions in the North Atlantic–European region and for postprocessing monthly and seasonal forecasts ([Neal et al. 2016](#)). Meteorological drought forecasts over Europe were improved using WR predictors, especially for intense droughts ([Lavaysse et al. 2018](#)). In North America, four WRs were identified with distinct relationships to rainfall and surface temperatures at a monthly time scale ([Vigaud et al. 2018](#)). Despite the promising results shown in meteorological forecasts, there has been less effort in applying the WR approach directly to hydrological forecasts, as outlined in this study. A recent study combined atmospheric circulation patterns and ML to predict extreme floods in the United States ([Schlef et al. 2019](#)). For a different application, [Grams et al. \(2017\)](#) derived a life cycle definition of seven year-round European WRs to quantify the effect of subseasonal meteorological variability on wind-power output in Europe. Both studies highlighted the potential of using WR data for forecasting purposes.

Different from most of the existing hydrological forecast research that predominately focuses on runoff prediction (e.g., [Madadgar et al. 2014](#); [Yuan et al. 2018](#)), and building upon the previous study of [Bogner et al. \(2022\)](#) on runoff and soil moisture, this study adds two additional hydrological variables, namely, baseflow and snowmelt. These two variables play important roles in forecasting floods and drought, yet they have been

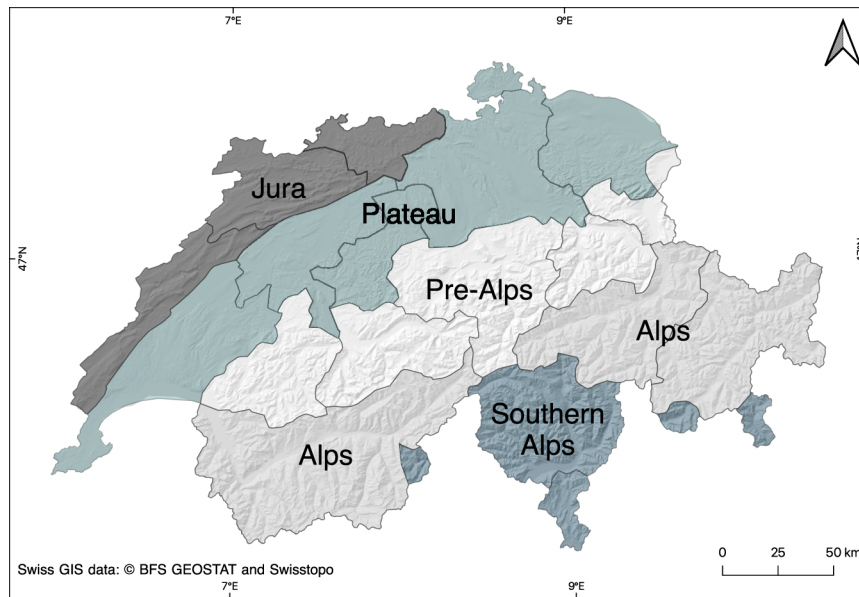


FIG. 1. Map of the study area, Switzerland, with the names of different regions.

frequently disregarded in previous research. The main objective of this study is to assess the feasibility of incorporating European WR data into a hybrid forecasting setup combining a traditional hydrological model and an ML algorithm to improve the performance of subseasonal hydrological forecasts in Switzerland. We aim to explore the potential of a forecasting system subject to different pre- and postprocessing methods to improve subseasonal hydrological forecast skill. Importantly, this study focuses on evaluating the overall skill of the different model setups over the year as a function of lead time without investigating the seasonal variability of the skill. We take a multimodel ensemble approach that enables us to generate tailored best practice maps for each variable, providing critical insights into the most effective processing techniques for different regions. Figure 1 shows the study area and the region names we refer to in this work. Our findings have implications for improving water resource management and decision-making in Switzerland and potentially wider applications in other regions with similar hydrological characteristics such as the European Alpine Space (Stephan et al. 2021).

## 2. Data

Similar to the study setup of Bogner et al. (2022), this study utilizes four sets of hydrological data: raw forecast, preprocessed

forecast, reference simulation, and climatology simulation. A novelty of this study is the introduction of European WR data. The concept is to postprocess raw forecast and preprocessed forecasts with the additional information provided by the WR data via an ML model with the aim to better match the reference simulation as in previous applications. Table 1 provides an overview of the characteristics of these datasets.

### a. Simulated hydrological data

The hydrological data in Bogner et al. (2022) include outputs from the PREVAH model for runoff and baseflow, but newly added to this study are PREVAH outputs for snowmelt and soil moisture. PREVAH is a distributed conceptual hydrological model consisting of several modules accounting for processes including evapotranspiration, interception, snowmelt and icemelt, soil moisture storage, groundwater storage, and runoff generation (Viviroli et al. 2009). It has a distinct glacier module for firn melt, icemelt, and snowmelt, which is an important component of hydrological variability in Swiss Alpine areas (Klok et al. 2001). HRUs, short for “hydrological response units,” are unit areas in a basin with similar (expected) hydrological behavior (Viviroli et al. 2009). The model version used in this study has the same setup as the one used in the previous study of Brunner et al. (2019a), which is a gridded realization of the model for all of Switzerland at  $500\text{ m} \times 500\text{ m}$  grid resolution consisting of

TABLE 1. Characteristics summary of the five datasets used in this study.

Data	Type	Spatial resolution	Temporal resolution	Lead time	Ensemble member
Raw forecast	Hydrological	500 m	Daily	32 days	51
Preprocessed forecast	Hydrological	500 m	Daily	32 days	51
Reference simulation	Hydrological	500 m	Daily	—	1
Climatology simulation	Hydrological	500 m	Daily	—	1
WR forecast	Atmospheric	Continental	6-h	46 days	51

300 catchments. No HRU aggregation is performed for the gridded version. The required meteorological inputs include fields of wind speed, 2-m temperature, 2-m humidity, sunshine duration, precipitation, and solar radiation. All gridded meteorological fields, forecasted and observed, are first stored at 2-km resolution and then downscaled to 500-m grid resolution with bilinear interpolation during simulation (Brunner et al. 2019a). Temperatures are adjusted based on a lapse rate as first described in Fundel et al. (2013). Eighty percent of the basins have an area between 47 and 228 km<sup>2</sup>, with a median of 117 km<sup>2</sup>. The biggest catchment is over 400 km<sup>2</sup>, while the smallest basins are at the border of political Switzerland with an area of approximately 10 km<sup>2</sup>.

All PREVAH forecasts are available at daily resolution with a lead time of 32 days. The first four days of the forecasts are discarded when aggregating data from daily time steps to weekly, yielding four weekly values (e.g., the week-1 value is the average of day 5–day 11, and the week-2 value is the average of day 12–day 18, etc.). The tercile values are then computed based on the weekly values with respect to climatology (described below). Values below the 33rd percentile of climatology are categorized as “below normal.” Values between the 33rd and 66th percentile are categorized as “normal.” Values above the 66th percentile are categorized as “above normal.” Figure S1 in the online supplemental material is an example of the weekly tercile forecast for total runoff generated on 18 April 2019, with lead times up to 4 weeks. A more detailed description of tercile analysis can be found in Bogner et al. (2022).

The four sets of hydrological data are the following:

- *Reference simulation:* To generate this set of outputs, the PREVAH model is forced with spatially interpolated observed meteorological data from real-time measuring stations provided by the Federal Office of Meteorology and Climatology of Switzerland (MeteoSwiss). Information from about 300 stations is used, of which 70 stations have all variables available. This approach assumes that the meteorological input is perfect, and the hydrological model error remains the same over time for different initializations so that the forecasts can be compared with the reference simulation to evaluate forecast performance. The advantage of comparing the forecasts with the reference simulations is that the effect of errors from the hydrological model, measurement, and parameter estimation can be neglected. The data period is from March 2012 to May 2022.
- *Climatology simulation:* This is the long-term historical average (from 1981 to 2010) of PREVAH outputs forced with spatially interpolated observed meteorological data from real-time measuring stations. It is used to set the thresholds of the tercile classes, which are the 33rd and 66th percentile marks. The data period is from January 1981 to December 2010.
- *Raw forecast:* This is the PREVAH forecast forced with raw meteorological forecasts as input with no preprocessing applied. The raw meteorological inputs provided by MeteoSwiss are available in an ensemble of 51 members, based on the 51 members of the European Centre for Medium-Range Weather Forecasts (ECMWF) extended-range forecast (Vitart et al. 2019),

which result in an ensemble of 51 members for the raw forecast. For each forecast run, the initial condition is obtained from the reference simulation forced with observation data as outlined above. The data period is from March 2012 to May 2022.

- *Preprocessed forecast:* This set of preprocessed hydrological forecasts uses the preprocessed meteorological forecasts as model input. The preprocessed inputs undergo a quantile mapping bias correction process, and they are provided by MeteoSwiss. Quantile mapping is a statistical bias correction method that matches the distribution of the entire forecast data to the distribution of the observational data. To perform preprocessing of the meteorological input, the meteorological reforecasts have been used to estimate the correction factor that is applied to the operational forecast. Thus, the reforecasts themselves are not preprocessed. The setup of this preprocessing technique is described in Monhart et al. (2018, 2019). Out of all the required meteorological inputs, only mean daily temperature and precipitation are preprocessed, while other inputs such as relative humidity, global radiation, wind speed, and sunshine duration do not undergo bias correction. Initial conditions are obtained in the same way as for the raw forecast. The deployed preprocessing technique came into operation in 2018, and the preprocessed meteorological reforecasts are not available prior to 2018 due to lack of resources to generate such data for all of Switzerland. Thus, this dataset in the end limits the available data for this study. The preprocessed forecast is included in this study to compare the effect of postprocessing using WR data with the effect of preprocessing, aiming for a more comprehensive and unbiased analysis of the added value of the WR data. This set of forecasts also consists of 51 ensemble members. The data period is from March 2018 to May 2022.

The common period of these four sets of data is March 2018–May 2022. Raw forecast data prior to March 2018 are evaluated and presented in Bogner et al. (2022). Both raw and preprocessed PREVAH forecasts are currently operational, thus representative of the real-time potential of the hydrometeorological process in the catchments. The forecast frequency of PREVAH follows the forecast frequency of the meteorological forecasts at ECMWF. During the data period, the Integrated Forecast System (IFS) of ECMWF underwent several cycle updates from cycle 38r2 to cycle 47r3, which may cause inconsistencies in the forecast skill. Information on the IFS updates can be found at <https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model>. Given that this study aims to evaluate the relative improvement of the deployed postprocessing techniques, the effects due to the IFS updates are assumed to be canceled out and not further investigated in this study. The raw tercile forecasts are published on drought.ch, an open platform that provides information on current and potential drought in Switzerland (Zappa et al. 2014).

#### b. Weather regime data

To investigate the link between large-scale WRs and local hydrological events and determine the benefits of using WRs to postprocess hydrological data, we use the year-round

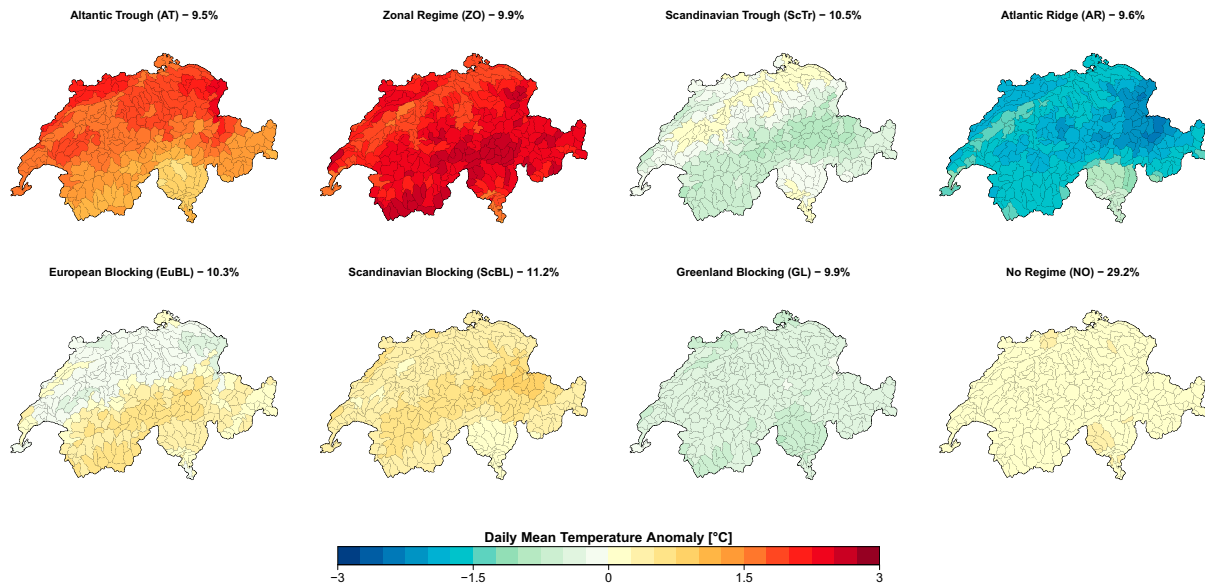


FIG. 2. Mean temperature anomalies associated with each of the seven weather regimes plus the no-regime regime in Swiss hydrological catchments from 1981 to May 2022. Anomalies computed with respect to climatology between the years 1981 and 2010, which is the same climatology period used for the tercile calculation of PREVAH forecasts. Daily mean temperature data (TabsD) used in this analysis are part of the MeteoSwiss Grid-Data Products (MeteoSwiss 2021a).

definition of seven Atlantic–European WRs, as established by Grams et al. (2017), to identify WRs in both ECMWF operational extended-range forecasts and operational analyses, following the forecast setup outlined in Grams et al. (2020) and Büeler et al. (2021). For each ensemble member and each 6-h lead time (or the operational analysis), we compute seven “weather regime indices” (IWRs), which are the normalized projection of the 10-day low-pass-filtered normalized Z500 anomalies into the respective regime pattern. Based on the IWRs, a WR life cycle (LC) is derived, which allows us to attribute each time step to a specific regime and identify days that do not exhibit regime behavior (no regime days). The technical steps in computing IWR and LC attribution are as in Büeler et al. (2021). In contrast to Büeler et al. (2021), here we use an updated version of the regime definition: we use ERA5 reanalysis data covering 1979–2019 and compute 10-day low-pass-filtered Z500 anomalies with respect to an ERA5 1979–2019 calendar day climatology instead of using ERA-Interim 1979–2018 and 5-day low-pass-filtered data used by Büeler et al. (2021). Same as in Büeler et al. (2021), we apply a bias correction for the Z500 anomalies to account for systematic lead-time-dependent biases. The WR forecast data are available for initializations each Monday and Thursday from January 2018 to May 2022 with 46-day lead times.

For illustrative purposes, we show the time series of IWR for the years 2018–22 of our common data period in Fig. S2. Active LCs are indicated in bold and—by definition—exhibit an IWR greater than 1.0 for at least 5 days. The color code in the bottom row in each panel of Fig. S2 indicates regimes that exhibit an active LC and the maximum IWR of all IWRs. This criterion is used for a unique LC attribution of each time step. In addition, we show mean temperature and precipitation anomalies in the hydrological catchments of Switzerland during WRs (based on the LC

attribution) between 1981 and 2022 in Figs. 2 and 3. The daily mean temperature (TabsD) and daily precipitation (RhiresD) data used in Figs. 2 and 3 are the interpolated values that are measured at the operational station network SwissMetNet of MeteoSwiss, and they are part of the MeteoSwiss Grid-Data Products (MeteoSwiss 2021a,b). Note that the LC attributions are only used in Figs. 2 and 3 for illustrating the surface weather impact of regimes, and not in the postprocessing phase. On average, the cyclonic regimes Atlantic Trough (AT) and Zonal Regime (ZO) exhibit warmer-than-usual conditions, and the anticyclonic regimes Atlantic Ridge (AR) and, to a lesser extent, Greenland Blocking (GL) cooler-than-usual conditions across Switzerland. Scandinavian Blocking (ScBL) exhibits a weak warm anomaly. Scandinavian Trough (ScTr) and European Blocking (EuBL) show contrasting temperature anomalies depending on the region considered (Jura Mountains, Swiss Plateau, Alps, or Southern Alps). Note that there are important seasonal differences (Fig. S3), in particular for the anticyclonic regimes (AR, EuBL, ScBL, GL) and ScTr. For instance, ScBL exhibits on average a negative temperature anomaly across Switzerland in winter and spring, but a warm anomaly in summer and autumn. WRs also modulate daily precipitation (Fig. 3) with some seasonal variability (Fig. S4). On average, AT, ScTr, and GL exhibit more than usual precipitation in most regions, whereas during ZO, AR, EuBL, and ScBL, negative precipitation anomalies occur (Fig. 3; except southern Switzerland for ScBL). Days that are not attributed to a regime (No regime days) exhibit only negligible anomalies. The different patterns suggest a link between the European WRs and local Swiss weather, which indicates that the inclusion of WR data in the postprocessing step of the hydrological model can potentially further help to improve the forecast skill.

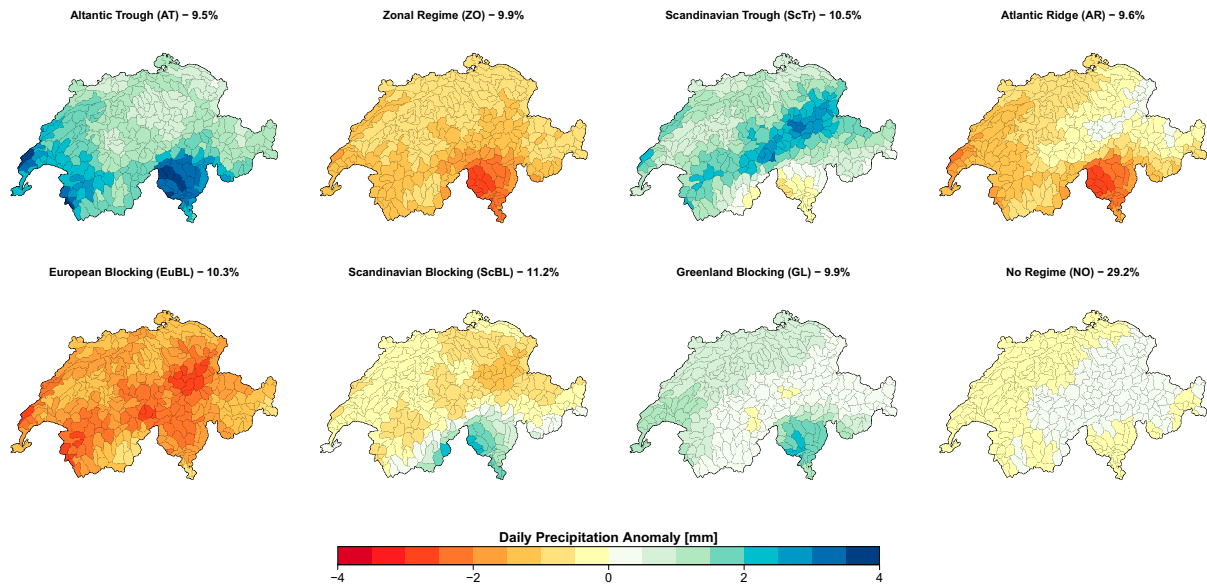


FIG. 3. As in Fig. 2, but for precipitation anomalies. Daily mean precipitation data (RhiresD) used in this analysis are part of the MeteoSwiss Grid-Data Products (MeteoSwiss 2021b).

### c. Data preparation

To evaluate and compare the effect of the different pre- and postprocessing methods, only the mutual period covered by all datasets is considered for this study, which covers the period from 18 March 2018 to 22 May 2022, providing 391 separate forecasts. The daily hydrological outputs are first aggregated to weekly values, resulting in a total of four weekly values for the monthly forecast period. The weekly values are then converted to classes of “low,” “medium,” or “high” according to the class thresholds set by the climatology means of the same week for each variable. This procedure of tercile computation is applied to the raw forecasts, the preprocessed forecasts, and the reference simulations. For the two sets of forecasts (raw and preprocessed), each class is reported as a probability value, the percentage of ensemble members in that class out of 51 members. In contrast, the reference simulations are reported without a probability value as the reference simulations are forced with observational data (only one member). The reference simulations are considered a proxy for the true hydrological condition.

Similarly, for each IWR, the weekly ensemble mean of the 51 ensemble members is computed, resulting in seven weekly index values, one for each of the seven WR types (excluding the no-regime type). The PREVAH data are then matched with the WR data by forecast initialization dates, resulting in a combined dataset with coherent forecast initialization dates between the hydrological and WR data. Finally, we randomize the combined dataset and then split the data by 75% and 25% for training and testing purposes, respectively. Within the 98 initializations of the testing data, the number of reference simulations in each tercile has a median ranging from 26 to 41 across all catchments.

### 3. Study setup

The postprocessing method is described in detail in this section as it is the main focus of this study, whereas details on the computation of the preprocessed meteorological data can be found in Monhart et al. (2018) and details on the WR indices can be found in Osman et al. (2023) and Büeler et al. (2021). For the hydrological model (PREVAH) setup, refer to Viviroli et al. (2009) and Brunner et al. (2019a).

Figure 4 provides an overview of the different variables and processing techniques (hereby referred to as “cases”) investigated in this study. The four hydrological variables are total runoff ( $Q$ ), baseflow (BF), soil moisture (SM), and snowmelt (SMELT). We analyze six different cases (A–F) with different combinations of processing techniques. All cases first undergo a tercile aggregation to convert forecast values into tercile classes as described in section 2. Then, preprocessing is applied to cases B, D, and F, and postprocessing is applied to cases C–F, within which the two cases E and F include the IWRs.

#### a. Postprocessing

In a similar setup to Bogner et al. (2022), both raw and preprocessed tercile forecasts are postprocessed to correct the error between forecast and reference simulation by utilizing an ML algorithm. Different from the previous study, we carry out postprocessing either without or with the additional WR data. For an ML algorithm, the inputs used to make predictions are called “features,” and the outcomes are called “labels.” We use a total of four different types of postprocessing techniques depending on the input features (see Table 4): case C, postprocessing raw forecasts; case D, postprocessing preprocessed forecasts (Pre + Post); case E, postprocessing raw forecasts with WR data (Post  $\times$  WR); and case F, postprocessing preprocessed forecasts with WR data

		Tercile Aggregation	Quantile Mapping of Meteo. Input	Machine Learning	WR Index Inclusion
A	No Processing	X			
B	Pre-Process	X	X		
C	Post-Process	X		X	
D	Pre + Post	X	X	X	
E	Post x WR	X		X	X
F	Pre + Post x WR	X	X	X	X

■ Total Runoff (Q)    ■ Baseflow (BF)  
■ Soil Moisture (SM)    ■ Snowmelt (SMELT)

FIG. 4. Overview of the six processing cases (A–F) considered in this study. All four variables of total runoff, baseflow, soil moisture, and snowmelt undergo the same processes. The color code for the variables will be followed throughout this study.

(Pre + Post × WR). In the cases without WR data (C and D), there is only one feature, which is the tercile forecast, whereas in the case with WR data (E and F), there are eight features—one tercile forecast plus seven IWRs. In all cases, the label is always the reference simulation, meaning the models are trained to match the reference simulation and afterward tested and evaluated with respect to the reference simulation. This type of machine learning is classified as supervised learning.

Five ML algorithms are considered based on similar applications (Bogner et al. 2019): random forest (RF), support vector machine (SVM), neural network (NN), Gaussian process (GP), and gradient boosting machine (GBM). We carry out a screening phase to select the “best” algorithm among the five considered algorithms. GP is selected based on the overall accuracy results (see Tables S1 and S2) and the finding of the previous study of Bogner et al. (2022).

The basic principle of GP is to derive a mean function by fitting the training data through the random functions retrieved from the distribution of all possible functions (Knagg 2019). The hyperparameter (i.e., the parameter the user can calibrate in an ML model) we specify in this study is the sigma ( $\sigma$ ) value associated with the kernel, and it is used to define the shape of the function. A radial basis function (RBF) kernel is chosen. A more detailed explanation of the GP algorithm is provided in the supplemental material and in the work of Bogner et al. (2022).

The models then undergo hyperparameter tuning and training. One model per catchment per lead time is set up for each case, which yields a total of 1228 models per variable per case. To tune the hyperparameter  $\sigma$ , user-defined manual grids are specified, and the tuning process is performed with a fivefold cross-validation process repeated three times. To reduce the computational time and to prevent the model from reaching a

local minimum, instead of specifying a grid with a wide range at small intervals, the tuning process is carried out in three steps:

- Step 1: 75 catchments are randomly selected to determine the ranges of the hyperparameters.
- Step 2: Within the ranges determined in step 1, models are tuned with a grid at a coarse interval of 0.1.
- Step 3: Each model is further tuned with a finer grid at 0.02 interval for a range of  $\pm 0.1$  around the best hyperparameters from step 2.

The optimal parameters are selected by maximizing the model overall accuracy, which is the average agreement rate between the prediction and the reference value over the cross-validation iterations (Kuhn 2021). The average training time for both step 2 and step 3 is 8 h CPU time for all catchments combined per variable. Adding WR indices only slightly prolongs the training time, and there is no significant difference in training time among the variables. Once all models are trained, they are tested with the testing data. In the testing phase, the reference simulations are not provided to the trained model. Instead, the ML models perform predictions with the input features and are evaluated against the reference simulations.

#### b. Verification measures

Different verification measures are considered to analyze the impact of different processing techniques on forecast predictability. As the study focuses on tercile forecasts, only categorical measures should be applied. The first measure we choose is the overall accuracy, which is the proportion of the correct predictions in all classes out of the total number of predictions (Kolachian and Saghaian 2021). It can be expressed by Eq. (1), where  $N$  is the total number of predictions and  $X_i$  is the

number of correct predictions in each tercile  $i$ . For example, if a classifier makes 10 predictions, 9 of which are correct, the overall accuracy is 0.9 or 90%. An overall accuracy of 1 denotes a perfect forecast, and a value of 0 indicates no predictive skill. Note that verification is performed against the reference simulations forced with interpolated observational meteorological information instead of point hydrological observations as they are not available at the resolution for the 300 catchments used in this study that cover the entire Switzerland. This approach allows us to study catchments with sizes relevant for regional planning purposes in Switzerland. Case A (raw forecast without pre- or postprocessing) is considered the “base case” and the overall accuracy differences are computed with respect to case A to assess the impact of different cases: a positive difference indicates an improvement, and a negative difference indicates a reduction in overall accuracy:

$$\text{overall accuracy} = \frac{\sum_{i=1}^3 X_i}{N}. \quad (1)$$

The ranked probability score (RPS) is chosen as the second verification measure as it is the most commonly used metric for multicategorical forecasts, e.g., terciles. RPS measures the difference between the distribution of forecasts and the distribution of observations (the reference simulation in this case) over the three classes, capturing the balance or imbalance of predictive power among the classes (Weigel et al. 2007). It is a combined measure of a forecasting system’s accuracy, liability, sharpness, and resolution. The ranked probability skill score (RPSS) is the corresponding skill score that determines the improvement of the forecast with respect to climatology. By definition, climatology has an equal distribution of 0.33 in each class. The RPSS ranges between 1 and  $-\infty$  with a positive value indicating an improvement compared to climatology.

#### 4. Results

This section focuses on the results from total runoff, with a summary of other variables. For figures of baseflow, soil moisture, and snowmelt, refer to the supplemental document.

##### a. Base case: Raw forecast

Before assessing the impact of different processing techniques and the added value of WR data, it is important to understand of the hydrological model’s predictability without pre- or postprocessing. Figures 5 and 6 demonstrate the magnitude and spatial variability of the overall accuracy and RPSS of the raw forecast for total runoff. The forecast skill varies greatly in space and decreases with lead time. In general, the two measures agree with each other, with higher forecast skill observed in the Jura Mountains, the Swiss Plateau, and the Southern Alps. The Alps display poor predictability starting in week 1 and throughout the forecast horizon. The low forecast skill (e.g., negative RPSS) is a result of the more complex runoff generation mechanism in these areas with higher ice/glacier coverage.

##### b. Forecast skill improvements

Figures 7 and 8 demonstrate the overall accuracy and RPSS achieved by the six cases for total runoff in all catchments and from week-1 to week-4 lead times. The two verification measures show a similar pattern, with overall accuracy displaying a larger variability. The median value indicates that the accuracy is above this value for half the catchments assessed in this study, and the variability reflects the consistency of skill in space. In week 1, cases B (preprocessing only) and D (pre- and postprocessing) show the highest improvement scoring the highest median value of overall accuracy and RPSS over all catchments, while other methods struggle to achieve a median value of skill that is higher than the raw forecast. However, the variability of skill in space has been reduced, especially for the RPSS for all cases. Moving to week 2, the differences in performance among the five processing cases start to level off. Case B (preprocessing only) still shows better performance with the highest median of overall accuracy and RPSS, but the RPSS variability is significantly lower in all cases with ML postprocessing (C–F). From week 3, the benefit of including WRs starts to surface with case F (pre- and postprocessing with WR) scoring the highest median in both measures, and in week 4, both cases with WRs, cases E (postprocessing with WR) and F (pre- and postprocessing with WR), outperform the other cases with higher median skill and reduced or similar variability. The RPSS results clearly indicate that using an ML postprocessing technique can better improve the skill consistency in space compared to the stand-alone preprocessing technique. This result suggests that the bias correction of meteorological inputs has a higher dependency on the geographical location of the catchment, which is likely linked to the station density in the region where observation data are used for quantile mapping. One possible reason for the higher skill consistency in space produced by the postprocessing cases is that each catchment is trained individually with its own set of parameters to maximize error correction.

The results from Figs. 7 and 8 are mapped in Figs. 9 and 10 in the form of overall accuracy and RPSS difference between each case and the raw forecast (A, base case) to show the spatial variability of the degree of absolute improvement. Each row is associated with one case, and every column is associated with one week of lead time. The stronger the opacity, the larger the difference is in either spectrum of forecast skill improvement or reduction. The main observations are summarized as follows:

- Among all cases and lead times, the most significant improvements are observed in several catchments in the Alps, whereas the Southern Alps show only small improvements. In regions north of the Alps, improvements are achieved with an increasing number of catchments with increasing lead time.
- Case D (pre- and postprocessing) is expected to exhibit results that combine the ones of case B (preprocessing only) and case C (postprocess without WR); however, this is not observed in all catchments in the Jura and Plateau regions in weeks 2 and 3, where the effect of preprocessing disappears when combined with postprocessing in some catchments.
- Case E (postprocessing with WR) and case F (pre- and postprocessing with WR) exhibit very similar spatial patterns. In



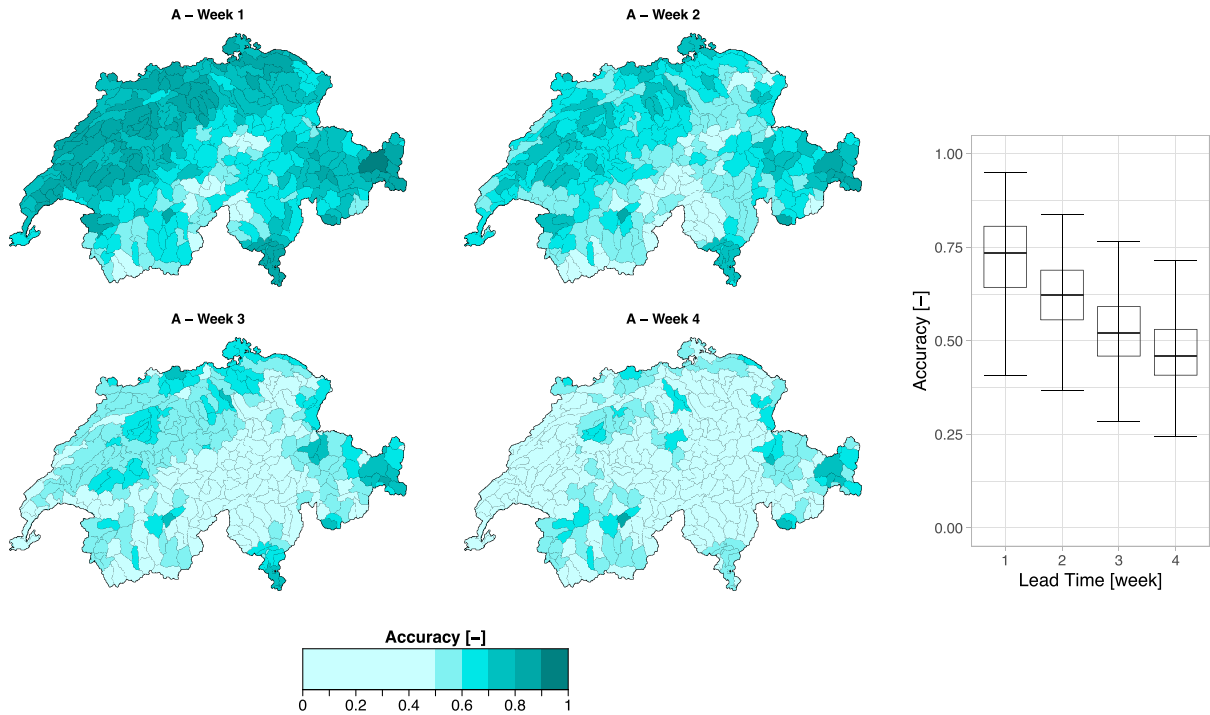


FIG. 5. Overall accuracy of base case (A), that is, the raw forecast with no pre- or postprocessing technique applied. Map view on the left demonstrates the spatial distribution, and the boxplot on the right displays the magnitude variability across catchments. Accuracy ranges from 0 to 1, with 1 denoting a perfect value. Variable: total runoff.

week 1, these two cases exhibit reduced forecast skill in catchments in the Jura and Plateau regions. However, compared to the cases without WR data, cases E and F can improve more catchments in week 4.

The proportions of catchments with an improved forecast skill (a positive overall accuracy/RPSS difference) for the five processing cases (B–F) are summarized in Table 2. The number of improved catchments increases with lead time for all cases

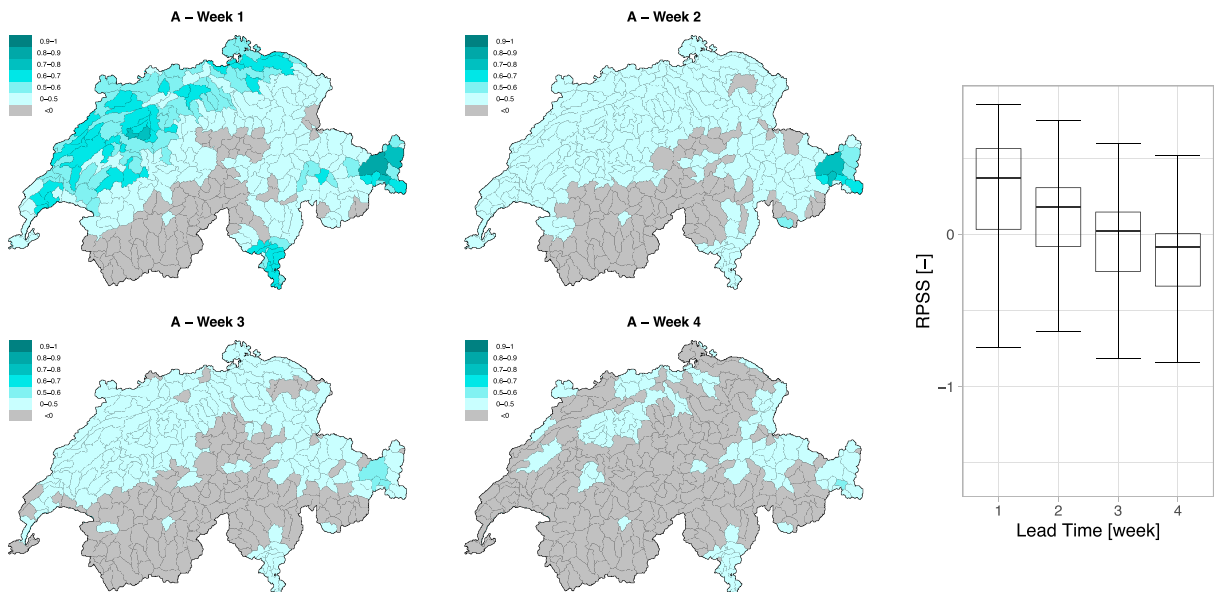


FIG. 6. As in Fig. 5, but for RPSS. An RPSS of 1 denotes a perfect skill while a negative RPSS indicates that the model performance is poorer than climatology.

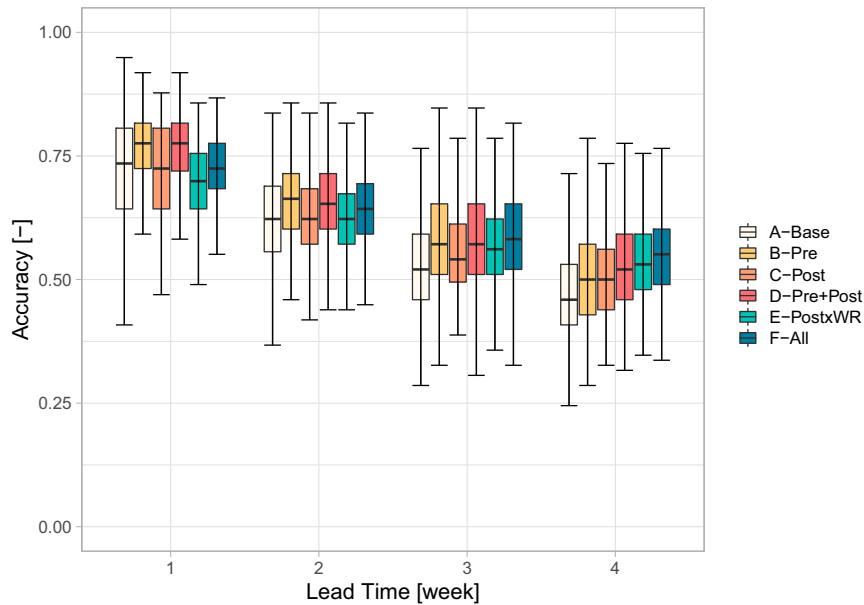


FIG. 7. Overall accuracy of all cases investigated. Outliers are not shown here. Variable: total runoff.

except for cases B (preprocessing only) and D (pre- and postprocessing) when evaluating overall accuracy. For the two cases, E (postprocessing with WR) and case F (pre- and postprocessing with WR), where WRs are included, the proportion of improved catchments follows a similar trend (with a marginal difference in weeks 3 and 4), most likely driven by the WR data. Provided adding WR data does not always increase the number of improved catchments (e.g., comparing case D with case F in weeks 1 and 2 based on overall accuracy), it

indicates that the improvement is not achieved by simply adding any additional piece of information, but the relevant information within the WR at the right time scale.

### c. Best practice

Our results agree with the findings in Bogner et al. (2022) that different processing techniques exhibit varying performances depending on the catchment and lead time. To optimize predictability across all catchments, we implement a multimodel

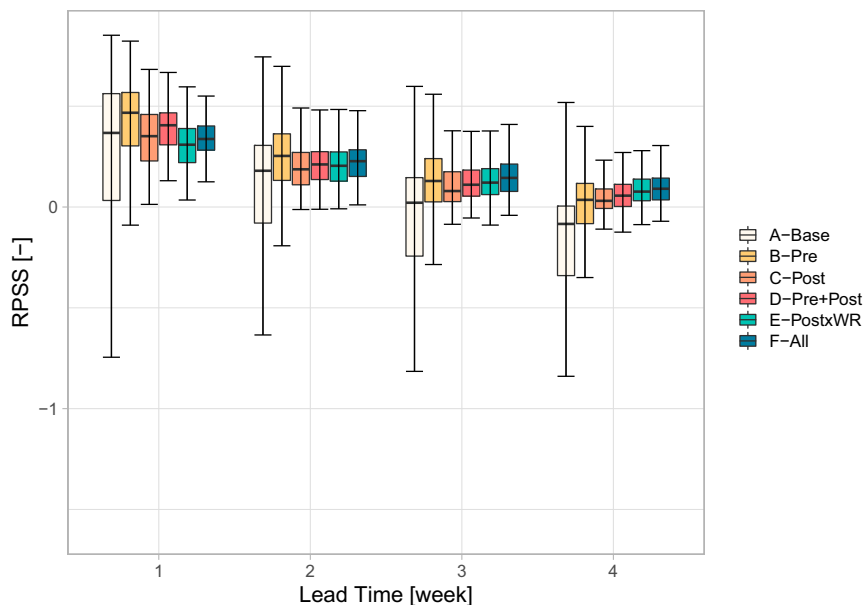


FIG. 8. As in Fig. 7, but for RPSS.

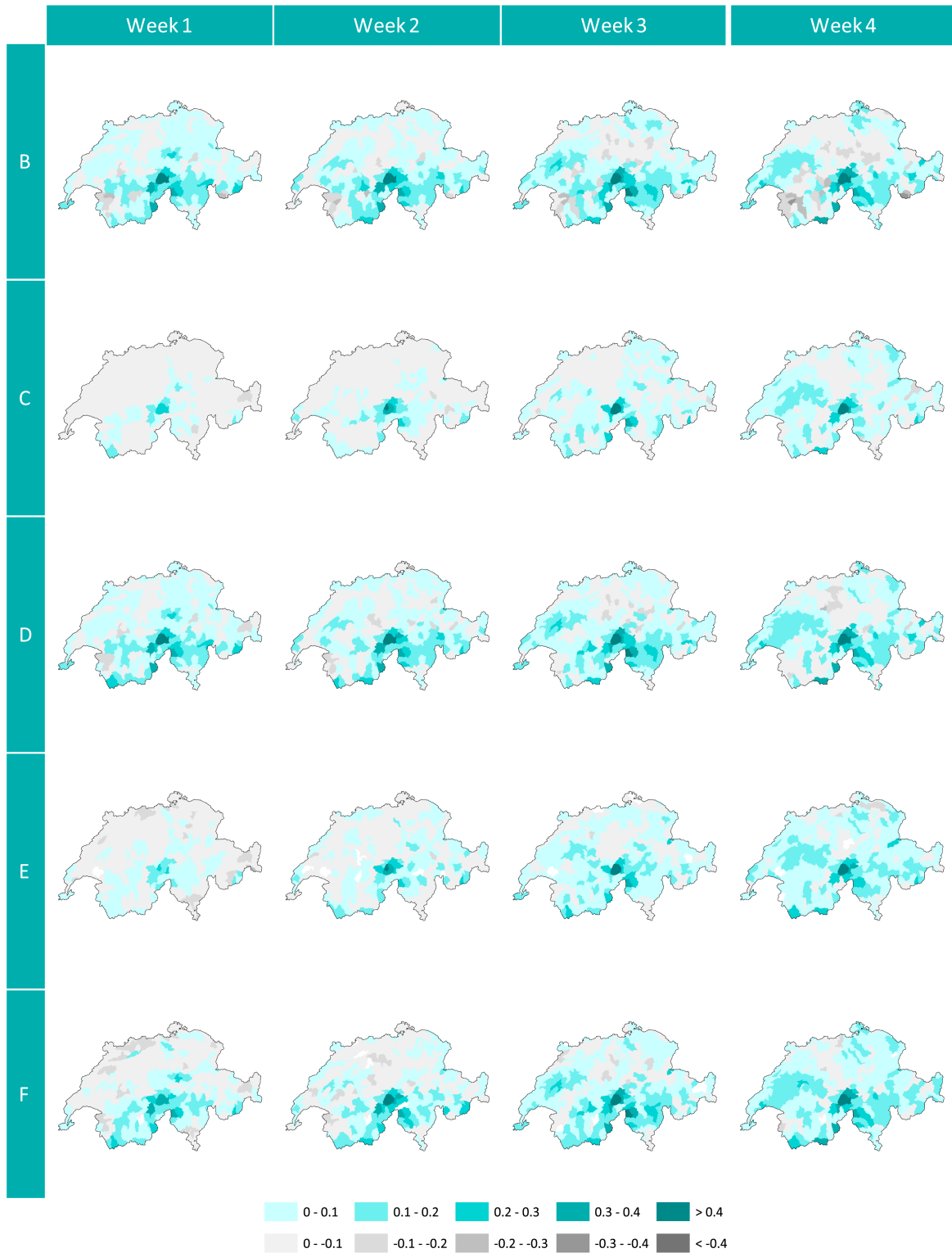


FIG. 9. Maps of overall accuracy differences of all processing cases (B–F) with respect to the raw forecast (A, base case), displaying the spatial distribution of forecast improvement achieved in four lead time weeks. Most improvements are observed in the Alpine region. Refer to Fig. 4 for information on the different cases. Variable: total runoff.

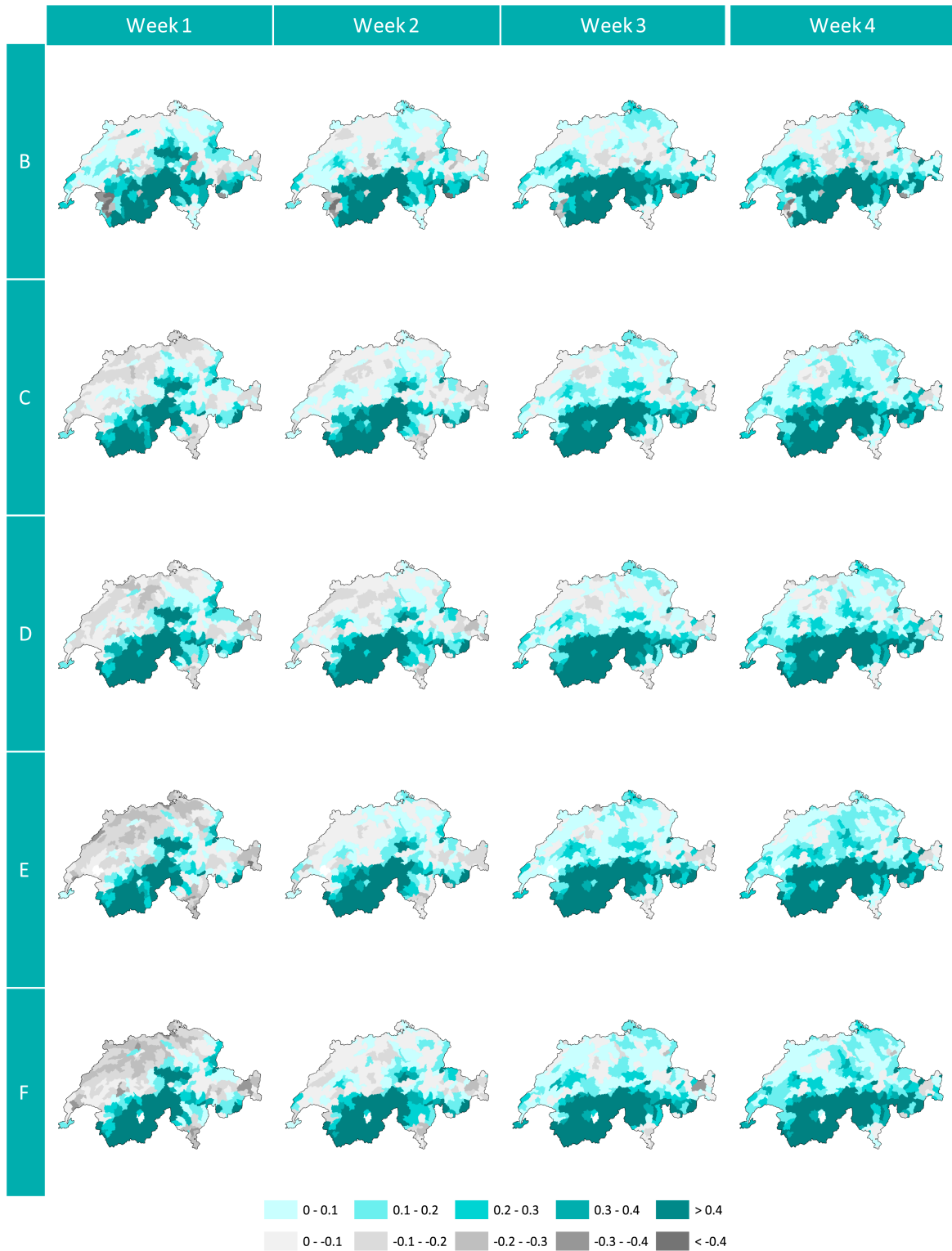


FIG. 10. As in Fig. 9, but for RPSS.

TABLE 2. Proportion of catchments where the forecast performance can be improved by each of the proposed processing techniques based on overall accuracy and RPSS. Bold values correspond to the highest proportion for each lead time week (per column). Variable: total runoff.

	Overall accuracy				RPSS			
	WK1	WK2	WK3	WK4	WK1	WK2	WK3	WK4
B, Pre	<b>67%</b>	<b>62%</b>	65%	57%	<b>65%</b>	<b>65%</b>	72%	72%
C, Post	17%	26%	48%	61%	48%	52%	68%	80%
D, Pre + Post	66%	59%	66%	63%	52%	52%	69%	79%
E, Post × WR	29%	40%	64%	76%	41%	57%	<b>77%</b>	84%
F, Pre + PostxWR	41%	47%	<b>68%</b>	<b>78%</b>	43%	52%	<b>77%</b>	<b>85%</b>

ensemble approach to identify the optimal processing technique (“best practice”) for each catchment, based on the technique that produces the highest skill score. This approach enables us to capitalize on the strengths of different processing techniques. As the RPSS is a more commonly used metric in the field of hydrology than overall accuracy, we choose the RPSS as the skill score to determine the best practices. Figure 11 maps out the best practices for total runoff. The proportions of catchments where each technique is deemed to be the best practice are summarized in Table 3. The map demonstrates that the choice of best practice changes with lead time in most catchments for the skill score considered, as expected. However, some catchments perform best with the same processing techniques (mostly preprocessing in the Plateau region) throughout the one-month time horizon. It is also worth noting that some catchments (about 3% of the catchments for total runoff) do not respond to any form of processing and the raw forecast remains the best option for all lead times, reflecting the skill of the raw forecast. By week 4, the two methods with WR data are deemed the best practice for most catchments, with a combined proportion of 62% based on the RPSS.

#### d. Additional hydrological variables

The main findings are listed here for the additional hydrological variables—baseflow, soil moisture, and snowmelt. Note that soil moisture and snowmelt results are based on the selected periods of March–October and February–June, respectively, where most fluctuations take place for these two variables making the selected periods most relevant for decision-makers. As it is relatively less challenging to predict soil moisture and snowmelt outside of these two time windows, the model performance would appear more skillful if year-round results were presented, which can be misleading and make it difficult to see the true effects of the different processing techniques. Furthermore, note that there are regions in Switzerland where snowmelt is not a key factor for water resources management.

##### 1) BASEFLOW

Refer to Figs. S5–S11 and Table S4.

- In week 1, fewer catchments are improved by postprocessing for baseflow than for any other variables.

- In weeks 1–2, cases B (preprocessing only) and D (pre- and postprocessing) have the best performance when evaluating overall accuracy.
- In weeks 1–3, case B (preprocessing only) outperforms the other methods when evaluating the RPSS.
- Comparing cases E and F with cases C and D, respectively, the inclusion of WR data can have a negative effect in weeks 1 and 2, but by week 4, the two cases with WR (E and F) outperform the other cases.

##### 2) SOIL MOISTURE

Refer to Figs. S12–S18 and Table S5.

- Case B (preprocessing only) is outperformed by other cases in weeks 2–4 with the other four cases having similar performance.
- There is no clear advantage of including WR data.
- Spatially, most improvements are located in the western and northern parts of Switzerland based on overall accuracy, which are regions where most of the croplands are located, bringing added value to the agriculture sector. However, according to the RPSS results, improvements are more evenly distributed in the study area starting in week 2.

##### 3) SNOWMELT

Refer to Figs. S19–S25 and Table S6.

- Case B (preprocessing only) outperforms the other processing methods based on the median values, but it produces larger spatial variability than other methods (based on RPSS).
- Case C (postprocessing without WR) has the poorest performance compared to the other processing methods.
- The benefit of adding WR to the ML model is not obvious for snowmelt.

The outcomes of the best practices of all variables in terms of RPSS are shown in Fig. 12, together with those of the raw forecast (base case), to demonstrate the positive effects when the preferred pre- and postprocessing techniques are applied to each catchment. The associated medians are summarized in Table 4. As expected, the degree of improvement varies by variable. The predictability of soil moisture has the highest improvement potential. Total runoff and baseflow maintain approximately the same level of improvement across all

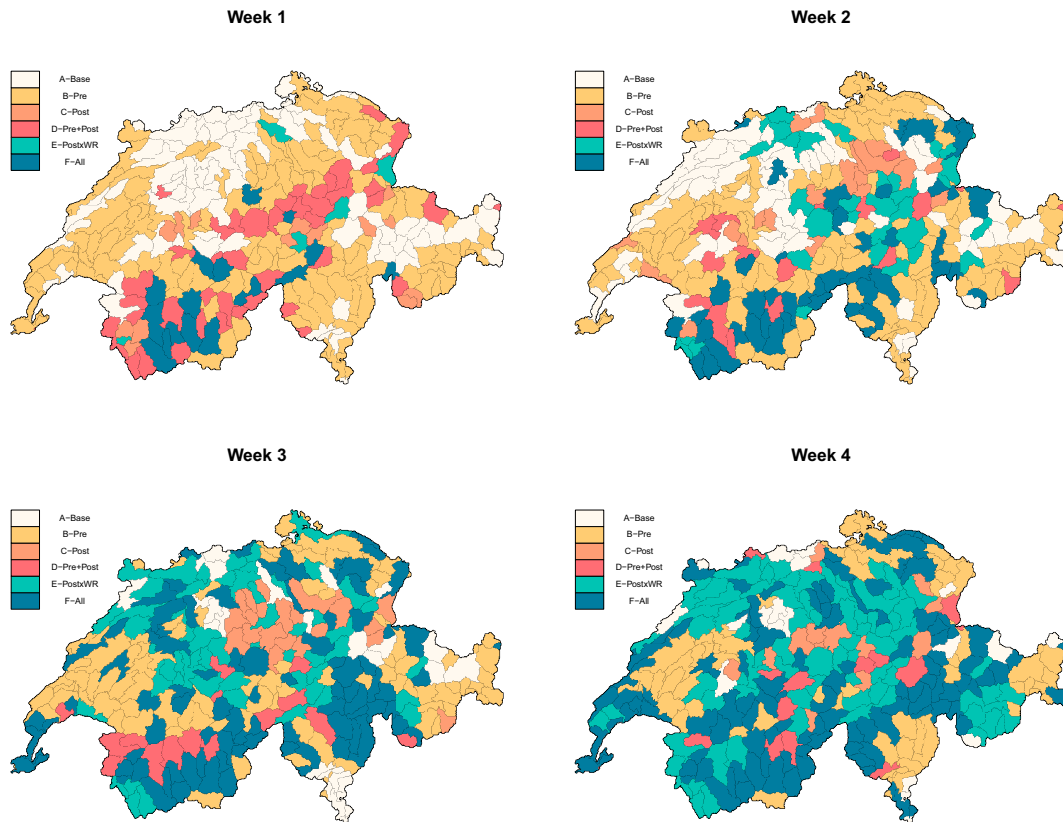


FIG. 11. Maps of best practices, which are determined by the processing techniques with the highest RPSS among all cases (cases A–F) for each catchment. Refer to Table 3 for the proportion of catchments where each case is deemed as the best practice. Variable: total runoff.

lead times, while snowmelt has a decreasing degree of improvement with lead time between weeks 1 and 3. The difference in response could be related to the variables' response time to changes in the system or their memory mechanism, which will be discussed in the next section.

## 5. Discussion

### a. Total runoff

Total runoff is an important hydrological variable for hydro-power production and hazard control applications. For this

TABLE 3. Proportion of catchments where each case is deemed to be the best practice based on the highest RPSS values among the six cases. Bold values correspond to the highest proportion for each lead time week (per column).

	WK1	WK2	WK3	WK4
A, No process	29%	23%	11%	9%
B, Pre	<b>45%</b>	<b>40%</b>	<b>31%</b>	20%
C, Post	5%	7%	7%	4%
D, Pre + Post	15%	6%	5%	5%
E, Post × WR	2%	10%	18%	25%
F, Pre + Post × WR	5%	14%	29%	<b>37%</b>

variable, case B (preprocessing only) shows promising results from weeks 1–3, whereas in week 4, postprocessing with WR information (cases E and F) shows the most improvement. This observation is an indicator that in early lead time, the meteorological inputs can be significantly improved by quantile mapping, which results in high hydrological forecast (PREVAH) skill, and leaves little room for improvement with postprocessing techniques. However, after week 3, the ability of preprocessing to improve the meteorological input quality starts to reduce, and postprocessing with additional WR information is needed. This result suggests that the physical consistency using meteorological models plays a crucial role at early lead times. At longer lead times, the WR approach can contribute to increased predictability through a statistical representation of the weather systems.

Spatially, the Jura and the Swiss Plateau regions have higher overall accuracy than other areas in the base case (raw forecast), which is mostly related to their lower elevations and flatter slopes. This high overall accuracy can be a reason that overall accuracy improvement is achieved in fewer catchments in northern Switzerland than in the rest of the country, especially in the first two weeks of lead time (see Figs. 9 and 10). In contrast, catchments with higher ice/glacier coverage tend to have lower

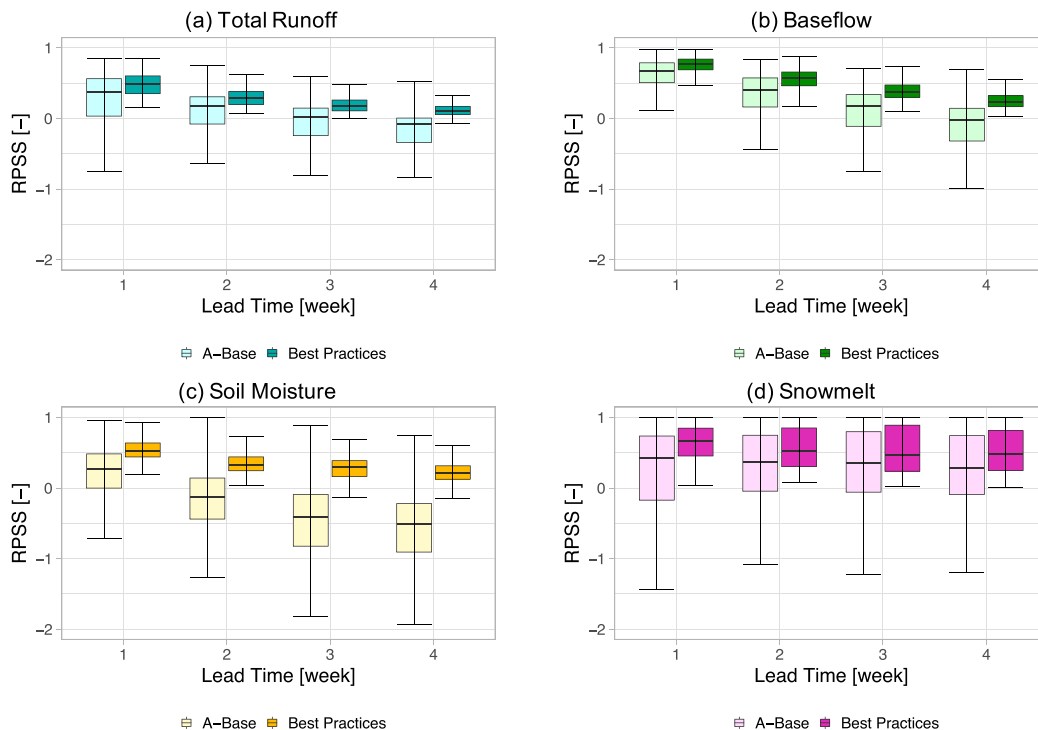


FIG. 12. RPSS results of all four variables when applying best practices compared with the ones of the base case (raw forecast) with no processing applied. A summary of the median values and the degree of improvement is provided in Table 4.

predictability in the base case, and these catchments are most responsive to pre- and postprocessing techniques. This confirms the findings of Monhart et al. (2019) on the high value of preprocessing of air temperature (and precipitation) in the seasons where cryosphere processes govern the accumulation and ablation of snow resources in Alpine areas. When plotting model performance improvement against catchment elevation and slope, a cluster is observed with a high degree of improvement in high elevation catchments with steep slopes for the two cases involving WR data for longer lead times. The increasing effect of WR data with lead time can be explained from two angles. First, in the first two weeks of lead time, preprocessing and postprocessing (without WR) are sufficient to improve overall accuracy in those catchments, but in week 3 and beyond, additional information that is not captured in the hydrological outputs is needed to accommodate the lower overall accuracy and higher variability,

which the WR data can provide. The second explanation is connected to the characteristic of WRs being “quasi-stationary, recurrent and persistent” (Osman et al. 2023), such that the WR data provide an opportunity for extended-range forecasts; therefore, their effect is more obvious at longer lead times. To support this theory, total runoff was directly predicted using WR indices only without PREVAH forecasts (see Figs. S26–S29), and the skill of weeks 2–4 is slightly higher than for week 1. In contrast to the deteriorating skill of the raw forecast with lead time, the WR-only prediction has a similar skill throughout the 4-week lead time. The same observation can also be made for the other hydrological variables.

In the recent study of Monhart et al. (2019), the benefits of the numerical weather prediction (NWP) approach on the sub-seasonal hydrometeorological ensemble predictions were investigated in small and medium-size mountainous catchments in

TABLE 4. Summary of the RPSS medians of the raw forecasts and the ones when best practices are applied. The “Diff.” column indicates the degree of improvement achieved by best practices. NP, no processing; BP, best practice.

Variable	Week 1			Week 2			Week 3			Week 4		
	NP	BP	Diff.	NP	BP	Diff.	NP	BP	Diff.	NP	BP	Diff.
Total runoff	0.37	0.49	+0.12	0.18	0.29	0.11	0.02	0.17	+0.13	-0.08	0.11	+0.19
Baseflow	0.67	0.77	+0.10	0.41	0.57	+0.16	0.17	0.37	+0.20	-0.03	0.23	+0.26
Soil moisture	0.27	0.53	+0.26	-0.12	0.32	+0.44	-0.41	0.29	+0.70	-0.51	0.21	+0.72
Snowmelt	0.42	0.66	+0.24	0.37	0.52	+0.15	0.35	0.47	+0.12	0.28	0.48	+0.20

Switzerland. Streamflow forecasts were generated using different preprocessing techniques. Like this study, the spatial variability of prediction skill is also observed. By preprocessing both temperature and precipitation, positive skill characterized by the continuous ranked probability skill score (CRPSS) could be extended from 5 days lead time to 15 days in the Verzasca catchment in southern Switzerland, and from 14 days lead time to 28 days (with several days of zero skill in between) in the Klöntal catchment in northeastern Switzerland. For the Thur catchment, with a high skill of the raw forecasts, the skill improvement by preprocessing was negligible. The results indicate that depending on the catchment, the effect of preprocessing varies accordingly. This effect has been observed with the “best practice” choice, where many catchments in the Swiss Plateau perform best with case B (preprocess) throughout the forecast horizon, but the best choice varies with time in other regions of Switzerland. With increasing lead times, when the effect of preprocessing diminishes, an approach like postprocessing with additional WR information is more promising to improve the forecast skill. This is also observed in terms of the “best practice” choice in Fig. 11.

As the computation of total runoff depends on the amount of precipitation, snowmelt in spring, and soil moisture storage, among other factors, its accuracy depends on the accuracy of other mentioned variables. Since the proposed postprocessing techniques treat each variable separately after the full model simulation is completed, there might be potential for further accuracy improvement if each variable influencing total runoff can be corrected at every time step prior to the computation of total runoff or in a multivariate postprocessing procedure. However, this approach would require a much more complex model architecture.

#### *b. Baseflow*

Baseflow, also called slow runoff, is the portion of water that percolates to the deeper part of the soil and contributes to groundwater flow; therefore, it is a very slow process. It has an impact on environmental flow for fish habitats and navigation in large rivers. Compared to total runoff and soil moisture, the raw forecast of baseflow has a much higher RPSS, which is mostly associated with the long memory mechanism, such that it is less sensitive to errors embodied in the meteorological forecast. On the other hand, the forecast skill of baseflow raw forecast among all catchments varies more with increasing lead time (the interquantile range gets larger with time), which could be related to the different soil types around the study area. When a best practice is applied to each catchment, case B (preprocessing only) dominates the study area up to week 3, and then the dominance switches to the two cases with WR data in week 4. To understand the spatial variability, it would be insightful to relate model performance results with a soil type map.

WR data only show added value to improve the predictability of baseflow in week 4, which can again be related to the long memory of baseflow. The inclusion of WR information only adds noise to the model at the early time horizon (before week 4); therefore, the additional WR information cannot have a

positive impact on the forecast skill. Perhaps a different type of large-scale weather pattern could be explored. For example, Rust et al. (2018, 2019) showed links between groundwater level variance, a much slower process, and the periodicity of the North Atlantic Oscillation and the east Atlantic pattern.

#### *c. Soil moisture*

Soil moisture experiences the greatest degree of skill improvement compared to other variables, and such improvement can be beneficial for its application in agriculture. When evaluating different techniques based on overall accuracy, case B (preprocessing only) demonstrates little improvements in weeks 1–3, and this could be linked to the finding of Orth and Seneviratne (2013) that realistic initial conditions are more important for soil moisture forecasts than accurate (meteorological) forcing forecast. This finding explains why preprocessing the meteorological input did not show great improvements. The predictability of soil moisture can be analyzed from the aspect of soil moisture memory, given the strong relationship between the forecast skill and memory demonstrated by Orth and Seneviratne (2013). Soil moisture has a long memory during persisting dry periods, but soil gets saturated and such memory is erased when precipitation occurs, resulting in the direct runoff, and hence responding fast to the precipitation signal. This mechanism leads to poor forecast quality in the base case, with the overall accuracy reduced to below 0.5 from week 3, and this low overall accuracy can be a reason for the high degree of improvement.

Although the magnitude of overall accuracy can be improved via a range of different techniques, the variability can hardly be narrowed, which is expected to be because the hydrological model already captures the spatial variability of soil type, and no new information could be learned or gained in the postprocessing procedure to narrow this variability. Another reason is that the WR data are at a continental scale and cannot cope with such high variability to help the model gain useful information to improve the tercile forecasts at the catchment scale.

From a drought perspective for agriculture application, soil moisture can be grouped with baseflow and improve the predictability of both variables in a bivariate model (Brunner et al. 2019b).

#### *d. Snowmelt*

Snowmelt is another variable worth investigating, given its importance for tourism (e.g., ski resorts) and flood generation. In different parts of the country, depending on altitude, aspect of slopes, and the local climate, snowmelt takes place at a different time, contributing to a large variability of model performance among catchments. Case B (preprocessing only) has the best performance, while the WR approach shows little added value for improving the model predictability. This finding agrees with the one in Jörg-Hess et al. (2015) on the importance of improving temperature predictability for cryospheric processes. It is worth noting that for snowmelt in this study, instead of analyzing results from the entire period, only the spring and summer seasons (February–June) when snowmelt occurs are



analyzed. The small sample size imposes a challenge for drawing conclusions on a seasonal basis.

#### e. Further findings

The responses of the four variables to the applied pre- and postprocessing methods have been discussed above. In addition, other findings that are not variable specific are discussed below.

Overall, the forecast skill of all cases decreases with lead time, which is driven mainly by the reducing skill of the meteorological inputs due to the chaotic nature of the system. When one processing technique is applied to the entire study area, due to the different catchment physical characteristics across space, the models exhibit large spatial variability in their skill even when pre- and/or postprocessing techniques are applied. By applying a best practice to each catchment, this spatial variability of forecast skill can be largely reduced. However, the choice of best practice (technique with the highest RPSS) can change with lead time, particularly in the majority of the catchments for total runoff and baseflow. Such choice can differ even among catchments sharing similar physical characteristics (only elevation and slope are considered here) for the same lead time, mainly during weeks 2 and 3, which is considered a transitioning period from short term (e.g., week 1) to medium term (e.g., week 4). Contrarily, the best practice remains the same throughout the forecast horizon in some areas, such as the flat lands in the Plateau region where concentrated agriculture activities are located, the high mountain areas where icemelt governs discharge, and in subcatchments with large water bodies. Further investigation on individual catchment level is required to understand if there is a connection between lead time dependency of the best practice (or the dominant effect of the pre- and postprocessing techniques) during this transitioning period and properties such as streamflow signatures, geographical domains and dominant runoff processes as suggested in the work of Pechlivanidis et al. (2020). Note that the best practice is determined simply by choosing the case with the highest RPSS, even when the RPSS values are very close, or the same in some cases. In the scenario where two cases have the same RPSS value, the less complex technique is chosen as the best practice (e.g., case C would be chosen over D). This selection process could have contributed to the patchy pattern in this transitioning period. Although the RPSS medians of all variables in week 4 are below 0.5 (see Table 4) when best practices are applied, and the usefulness of the tercile forecast at such skill level needs to be determined by taking into consideration of factors such as economic value, the degree of improvement has indicated the potential for such a multimodel best practice approach. Baseflow and soil moisture have the highest increase in RPSS median in week 4 (0.26 and 0.72, respectively), suggesting that such a best practice approach is suitable for drought prediction in longer lead times, which aligns with the finding of Fundel et al. (2013).

Results have shown some promising indication of the use of the WR data, mostly in weeks 3 and 4, with the degree of benefit varying across different variables. Several reasons might have contributed to the limited effect of adding WR indices in the postprocessing procedure:

- 1) The WR forecasts have their own skill, which also changes over time.
- 2) In most of the existing studies, WRs or weather patterns were identified specifically for the variable and region of interest based on known dependence or linkage. Differently, the WR data used here are derived independently of the purpose of this study.
- 3) The spatial variability of skill improvement can be due to the large difference in spatial resolution between the hydrological output and WR data (500 m versus continental scale). Furthermore, with the heterogeneous regions and the effect of high mountains in the Swiss Alps, it is expected that the WR data cannot yield the same amplitude and type of response in all catchments.
- 4) The inclusion of WR data helps the ML model consider the variability in weather more than the local meteorological data that is the input to the hydrological model. Therefore, the improvement from WR data is the greatest when the information contained in the hydrological output becomes insufficient for an ML algorithm to correlate the features to labels.
- 5) As shown in Figs. S3 and S4, WRs are associated with seasonality. However, no additional information is currently provided to the ML model to indicate this seasonality.

#### f. Limitations

Limited data availability is the main challenge of this study. With the current study setup, we are unable to extend the data prior to 2018 for two reasons. First, the meteorological reforecasts are not preprocessed on an operational basis by MeteoSwiss at the scale of entire Switzerland for resource reason. Only the forecasts themselves are preprocessed using the (raw) reforecasts to estimate the correction factors. Preprocessed reforecasts are thus so far only available for specific analyses and for periods therein. For example, Monhart et al. (2019), Jörg-Hess et al. (2015), Fundel et al. (2013), and Anghileri et al. (2019) provide an extensive discussion about the value of reforecasts in subseasonal runoff predictions in selected Swiss Alpine catchments. Second, WR data are not available operationally prior to 2018, and using operational data is an important aspect of this study. Without the preprocessed reforecast data, we are unable to extend the training period and potentially further improve postprocessing performance. Although meteorological reforecasts are indirectly included in this study in the estimation of the preprocessing correction factor, by excluding the preprocessed reforecasts, it also prevents this study from providing a more comprehensive assessment of the subseasonal forecasting system as reforecast data are commonly used to better understand the systematic errors. Given that using reforecast data is not possible for this study, limited by the availability of the preprocessed forecasts, readers should keep in mind that the assessment of the reforecast component is missing for a full investigation of the hydrological forecast. With a larger amount of WR data (e.g., reforecast data), the ML models could be trained to capture more connections between weather regimes and local hydrological variability. Furthermore, since the connection between the Swiss weather anomalies and the different

European weather regimes is stronger in some seasons than in others (e.g., AT exhibits a warm anomaly in winter that is much stronger than the rest of the year, see Figs. S2 and S3), the model performance might improve if the models are trained separately by season. However, this approach would further reduce the data size and potentially the model performance.

Currently, we train one ML model per catchment per lead time, which yields 1228 models for each variable for each postprocessing case. The option of implementing regional models should be explored as it will not only help reduce the number of models to be trained, but it also has many potential benefits. First, a regional model can be more robust as it applies the same or a similar number of parameters for a greater area, reducing the chance and degree of overfitting or overparameterization. Second, a regional model might perform better as it is trained with more data (lumped from multiple catchments). As a result, it might be able to learn about the catchment interaction, which cannot be learned at the catchment level. Finally, such regionalization could be executed according to different criteria, such as catchment physical properties (e.g., elevation or slope), the performance of the current catchment models, the best practice method, or the link between weather regime and local weather.

Furthermore, there are limitations associated with the ML algorithms that prevent the postprocessing from achieving a better performance. In cases C (postprocessing without WR) and D (pre- and postprocessing without WR), there are three scenarios where a model cannot be trained. First, a model cannot be trained when both the feature (forecast) and the label (reference simulation) only have one class (e.g., all forecasts and reference simulations have a tercile class of “low”). Another scenario is when the forecast has more classes than the reference (e.g., the forecast contains classes of “low” and “medium,” but the reference simulation only has one class of “medium”). A third scenario is mismatched classes (e.g., forecast contains classes of “low” and “medium,” but reference has classes of “medium” and “high”). In addition to missing classes, there is also the scenario of imbalanced classes where certain classes have more occurrences than the others, resulting in misclassification and low model performance (Sun et al. 2009).

While preprocessing specifically removes errors in the meteorological forcing, postprocessing with WRs aims to further remove hydrological errors by linking errors in catchment level hydrological forecasts with different WRs. Nevertheless, the type of errors that are reduced or removed by the ML model during postprocessing is not always clear, and such approach might not produce physically plausible or explainable results, which is another limitation of this study (Slater et al. 2023). However, the ML community has put a large amount of effort and resource in improving the model interpretability in recent years, and there are more advanced algorithms tackling this issue. To assess the added value of WRs, we have chosen a more conventional algorithm for this study.

In terms of the computation of RPSS, Manrique-Suñén et al. (2020) pointed out that the sample size can affect the

robustness of the skill score. As the testing dataset is limited in this study, the skill score might be prone to noise. One should keep in mind that the skill scores in this study are meant for comparison purposes among the different processing techniques, and they should be interpreted in relative values and not in absolute values.

## 6. Conclusions

This study examines the effectiveness of different pre- and postprocessing techniques in enhancing the predictability of sub-seasonal tercile forecasts for four hydrological variables, namely, total runoff, baseflow, soil moisture, and snowmelt. Specifically, we seek to assess the potential and added value of using European weather regimes as a postprocessing method. We observe predictability improvements for up to four weeks of lead time, with preprocessing demonstrating the most significant improvement in week 1 for most variables. From week 2 onward, postprocessing becomes more noticeable, although the degree of improvement varies with variables, catchments, and lead time.

Our results indicate that the inclusion of WR data, specifically the European weather regimes, has a positive impact on the forecast skill of hydrological variables in Switzerland. However, this impact of WRs differs across hydrological variables, catchments, and lead times. Generally, the benefit of including WR data increases with lead time, and high elevation catchments with steep slopes experience the most significant improvements with the two postprocessing methods involving WR data. For total runoff, there is a clear added value of WR in weeks 3 and 4. For baseflow, the effect of WR becomes obvious in week 4. For soil moisture and snowmelt, some catchments still experience improvements with WR, but overall, the added value of WR for these two variables is considered insignificant. When comparing the different variables, the WR approach can be identified as the “best practice” in more catchments for total runoff and baseflow than for soil moisture and snowmelt. This difference in response could be related to the strong seasonality of soil moisture and snowmelt. Furthermore, the spatial coverage associated with the snowmelt occurrence can help explain the low impact of WRs on snowmelt. To further understand the added value of WR information and how to apply it effectively, the next step is to analyze which of the seven WRs can improve the hydrological forecast skill the most.

In conclusion, advanced from the previous study of Bogner et al. (2022), this study shows the potential of incorporating European weather regime data in an ML-aided hybrid model to improve the skill of subseasonal tercile hydrological forecasts in Switzerland, especially in longer lead times, which has important implications for water resources management decision-makers. This finding lays a foundation for further inclusion of weather regime data in the forecasting of local hydrological events, such as extending the lead time beyond 4 weeks given the added value of WR becomes obvious in week 3 and above, extending the study area to the European domain as some parts of Europe might experience stronger

response to the weather regimes and produce more skillful forecast, or applying such an approach directly to streamflow and lake level predictions. Furthermore, we implement a multimodel ensemble approach by combining various processing techniques to provide insights into the most effective strategies for improving subseasonal hydrological forecasting in Switzerland. Overall, our study offers insights into the potential of including weather regimes in a hybrid setup to enhance forecast accuracy and highlights the importance of taking a holistic approach to hydrological forecasting, one that considers multiple variables and processing techniques.

**Acknowledgments.** This study is supported by the Malefix project, which is part of the WSL Program Extremes. Support from the Swiss National Science Foundation through projects PP00P2\_170523 and PP00P2\_198896 to A.C. and D.D. is gratefully acknowledged. A.C.'s contribution is also part of the Interreg Alpine Space Programme project ADO (Alpine Space Observatory; Grant ASP940), which in Switzerland has been financed via agreements with the Federal Office for Spatial Development ARE and the Cantons of Ticino and Thurgau. C.M.G.'s contribution was funded by the Helmholtz Association as part of the Young Investigator Group "Sub-seasonal Predictability: Understanding the Role of Diabatic Outflow" (SPREADOUT, Grant VH-NG-1243). ECMWF and Deutscher Wetterdienst are acknowledged for granting access to computing facilities and operational ensemble forecast data. We thank Marisol Osman from the LSDP group at KIT for providing the updated operational regime forecasts.

**Data availability statement.** Meteorological data used in this study are provided by MeteoSwiss in an operational mode.

## REFERENCES

- Anghileri, D., S. Monhart, C. Zhou, K. Bogner, A. Castelletti, P. Burlando, and M. Zappa, 2019: The value of subseasonal hydrometeorological forecasts to hydropower operations: How much does preprocessing matter? *Water Resour. Res.*, **55**, 10159–10178, <https://doi.org/10.1029/2019WR025280>.
- 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>.
- Bogner, K., K. Liechti, and M. Zappa, 2016: Post-processing of stream flows in Switzerland with an emphasis on low flows and floods. *Water*, **8**, 115, <https://doi.org/10.3390/w8040115>.
- , —, L. Bernhard, S. Monhart, and M. Zappa, 2018: Skill of hydrological extended range forecasts for water resources management in Switzerland. *Water Resour. Manage.*, **32**, 969–984, <https://doi.org/10.1007/s11269-017-1849-5>.
- , F. Pappenberger, and M. Zappa, 2019: Machine learning techniques for predicting the energy consumption/production and its uncertainties driven by meteorological observations and forecasts. *Sustainability*, **11**, 3328, <https://doi.org/10.3390/su11123328>.
- , A. Y.-Y. Chang, L. Bernhard, M. Zappa, S. Monhart, and C. Spirig, 2022: Tercile forecasts for extending the horizon of skillful hydrological predictions. *J. Hydrometeorol.*, **23**, 521–539, <https://doi.org/10.1175/JHM-D-21-0020.1>.
- Brunner, M. I., A. Björnson Gurung, M. Zappa, H. Zekollari, D. Farinotti, and M. Stähli, 2019a: Present and future water scarcity in Switzerland: Potential for alleviation through reservoirs and lakes. *Sci. Total Environ.*, **666**, 1033–1047, <https://doi.org/10.1016/j.scitotenv.2019.02.169>.
- , K. Liechti, and M. Zappa, 2019b: Extremeness of recent drought events in Switzerland: Dependence on variable and return period choice. *Nat. Hazards Earth Syst. Sci.*, **19**, 2311–2323, <https://doi.org/10.5194/nhess-19-2311-2019>.
- Büeler, D., L. Ferranti, L. Magnusson, J. F. Quinting, and C. M. Grams, 2021: Year-round sub-seasonal forecast skill for Atlantic-European weather regimes. *Quart. J. Roy. Meteor. Soc.*, **147**, 4283–4309, <https://doi.org/10.1002/qj.4178>.
- Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, <https://doi.org/10.1175/MWR2905.1>.
- Cohen, J., D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, 2019: S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdiscip. Rev.: Climate Change*, **10**, e00567, <https://doi.org/10.1002/wcc.567>.
- Delorit, J., E. C. Gonzalez Ortuya, and P. Block, 2017: Evaluation of model-based seasonal streamflow and water allocation forecasts for the Elqui Valley, Chile. *Hydrol. Earth Syst. Sci.*, **21**, 4711–4725, <https://doi.org/10.5194/hess-21-4711-2017>.
- Dobrynin, M., and Coauthors, 2018: Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophys. Res. Lett.*, **45**, 3605–3614, <https://doi.org/10.1002/2018GL077209>.
- Domeisen, D. I. V., A. H. Butler, K. Fröhlich, M. Bittner, W. A. Müller, and J. Baehr, 2015: Seasonal predictability over Europe arising from El Niño and stratospheric variability in the MPI-ESM seasonal prediction system. *J. Climate*, **28**, 256–271, <https://doi.org/10.1175/JCLI-D-14-00207.1>.
- Fundel, F., S. Jörg-Hess, and M. Zappa, 2013: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices. *Hydrol. Earth Syst. Sci.*, **17**, 395–407, <https://doi.org/10.5194/hess-17-395-2013>.
- Girons Lopez, M., L. Crochemore, and I. G. Pechlivanidis, 2021: Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden. *Hydrol. Earth Syst. Sci.*, **25**, 1189–1209, <https://doi.org/10.5194/hess-25-1189-2021>.
- Grams, C. M., R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli, 2017: Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nat. Climate Change*, **7**, 557–562, <https://doi.org/10.1038/nclimate3338>.
- , L. Magnusson, and L. Ferranti, 2020: How to make use of weather regimes in extended-range predictions for Europe. *ECMWF Newsletter*, No. 165, ECMWF, Reading, United Kingdom, 8 pp., <https://www.ecmwf.int/en/eLibrary/81199-how-make-use-weather-regimes-extended-range-predictions-europe>.
- Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, [https://doi.org/10.1175/1520-0493\(2004\)132<1434:ERIMFS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2).
- Heish, W. W., 2009: *Machine Learning Methods in the Environmental Sciences*. Cambridge University Press, 349 pp.
- Hwang, J., P. Orenstein, J. Cohen, K. Pfeiffer, and L. Mackey, 2019: Improving subseasonal forecasting in the western U.S.

- with machine learning. *KDD'19: Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, Association for Computing Machinery, 2325–2335, <https://doi.org/10.1145/3292500.3330674>.
- Jörg-Hess, S., S. B. Kempf, F. Fundel, and M. Zappa, 2015: The benefit of climatological and calibrated reforecast data for simulating hydrological droughts in Switzerland. *Meteor. Appl.*, **22**, 444–458, <https://doi.org/10.1002/met.1474>.
- Klok, E. J., K. Jasper, K. P. Roelofsma, J. Gurtz, and A. Badoux, 2001: Distributed hydrological modelling of a heavily glaciated Alpine river basin. *Hydrol. Sci. J.*, **46**, 553–570, <https://doi.org/10.1080/02626660109492850>.
- Knagg, O., 2019: An intuitive guide to Gaussian processes. Accessed 10 March 2020, <https://towardsdatascience.com/an-intuitive-guide-to-gaussian-processes-ec2f0b45c71d>.
- Kolachian, R., and B. Saghafian, 2021: Hydrological drought class early warning using support vector machines and rough sets. *Environ. Earth Sci.*, **80**, 390, <https://doi.org/10.1007/s12665-021-09536-3>.
- Kuhn, M., 2021: caret: Classification and regression training, version 6.0-88. R package, <https://CRAN.R-project.org/package=caret>.
- Lavaysse, C., J. Vogt, A. Toreti, M. L. Carrera, and F. Pappenberger, 2018: On the use of weather regimes to forecast meteorological drought over Europe. *Nat. Hazards Earth Syst. Sci.*, **18**, 3297–3309, <https://doi.org/10.5194/nhess-18-3297-2018>.
- Liu, L., C. Xiao, L. Du, P. Zhang, and G. Wang, 2019: Extended-range runoff forecasting using a one-way coupled climate-hydrological model: Case studies of the Yiluo and Beiji River basins in China. *Water*, **11**, 1150, <https://doi.org/10.3390/w11061150>.
- Madadgar, S., H. Moradkhani, and D. Garen, 2014: Towards improved post-processing of hydrologic forecast ensembles. *Hydrol. Processes*, **28**, 104–122, <https://doi.org/10.1002/hyp.9562>.
- Manrique-Suñén, A., N. Gonzalez-Reviriego, V. Torralba, N. Cortesi, and F. J. Doblas-Reyes, 2020: Choices in the verification of S2S forecasts and their implications for climate services. *Mon. Wea. Rev.*, **148**, 3995–4008, <https://doi.org/10.1175/MWR-D-20-0067.1>.
- MeteoSwiss, 2021a: Daily mean, minimum and maximum temperature: TabsD, TminD, TmaxD. MeteoSwiss Grid-Data Products Doc., 5 pp., [https://www.meteosuisse.admin.ch/dam/jcr:818a4d17-cb0c-4e8b-92c6-1a1bdf5348b7/ProdDoc\\_TabsD.pdf](https://www.meteosuisse.admin.ch/dam/jcr:818a4d17-cb0c-4e8b-92c6-1a1bdf5348b7/ProdDoc_TabsD.pdf).
- , 2021b: Daily precipitation (final analysis): RhiresD. MeteoSwiss Grid-Data Products Doc., 6 pp., [https://www.meteoswiss.admin.ch/dam/jcr:4f51f0f1-0fe3-48b5-9de0-15666327e63c/ProdDoc\\_RhiresD.pdf](https://www.meteoswiss.admin.ch/dam/jcr:4f51f0f1-0fe3-48b5-9de0-15666327e63c/ProdDoc_RhiresD.pdf).
- , 2021c: Monthly outlook. MeteoSwiss <https://www.meteoswiss.admin.ch/home/weather/forecasts/monthly-outlook.html>.
- Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A. Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.*, **123**, 7999–8016, <https://doi.org/10.1029/2017JD027923>.
- , M. Zappa, C. Spirig, C. Schär, and K. Bogner, 2019: Subseasonal hydrometeorological ensemble predictions in small- and medium-sized mountainous catchments: Benefits of the NWP approach. *Hydrol. Earth Syst. Sci.*, **23**, 493–513, <https://doi.org/10.5194/nhess-23-493-2019>.
- Neal, R., D. Fereday, R. Crocker, and R. E. Comer, 2016: A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteor. Appl.*, **23**, 389–400, <https://doi.org/10.1002/met.1563>.
- Orth, R., and S. I. Seneviratne, 2013: Predictability of soil moisture and streamflow on subseasonal timescales: A case study. *J. Geophys. Res. Atmos.*, **118**, 10963–10979, <https://doi.org/10.1002/jgrd.50846>.
- Osman, M., R. Beerli, D. Büeler, and C. M. Grams, 2023: Multi-model assessment of sub-seasonal predictive skill for year-round Atlantic–European weather regimes. *Quart. J. Roy. Meteor. Soc.*, **149**, 2386–2408, <https://doi.org/10.1002/qj.4512>.
- Papacharalampous, G., and H. Tyralis, 2022: A review of machine learning concepts and methods for addressing challenges in probabilistic hydrological post-processing and forecasting. *Front. Water*, **4**, 961954, <https://doi.org/10.3389/frwa.2022.961954>.
- Pechlivanidis, I. G., 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>.
- Rust, W., I. Holman, R. Corstanje, J. Bloomfield, and M. Cuthbert, 2018: A conceptual model for climatic teleconnection signal control on groundwater variability in Europe. *Earth Sci. Rev.*, **177**, 164–174, <https://doi.org/10.1016/j.earscirev.2017.09.017>.
- , —, J. Bloomfield, M. Cuthbert, and R. Corstanje, 2019: Understanding the potential of climate teleconnections to project future groundwater drought. *Hydrol. Earth Syst. Sci.*, **23**, 3233–3245, <https://doi.org/10.5194/hess-23-3233-2019>.
- Sahu, N., A. W. Robertson, R. Boer, S. Behera, D. G. DeWitt, K. Takara, M. Kumar, and R. B. Singh, 2017: Probabilistic seasonal streamflow forecasts of the Citarum River, Indonesia, based on general circulation models. *Stochastic Environ. Res. Risk Assess.*, **31**, 1747–1758, <https://doi.org/10.1007/s00477-016-1297-4>.
- Schlef, K. E., H. Moradkhani, and U. Lall, 2019: Atmospheric circulation patterns associated with extreme United States floods identified via machine learning. *Sci. Rep.*, **9**, 7171, <https://doi.org/10.1038/s41598-019-43496-w>.
- Slater, L., and Coauthors, 2023: Hybrid forecasting: Blending climate predictions with AI models. *Hydrol. Earth Syst. Sci.*, **27**, 1865–1889, <https://doi.org/10.5194/nhess-27-1865-2023>.
- Stephan, R., M. Erfurt, S. Terzi, M. Žun, B. Kristan, K. Haslinger, and K. Stahl, 2021: An inventory of Alpine drought impact reports to explore past droughts in a mountain region. *Nat. Hazards Earth Syst. Sci.*, **21**, 2485–2501, <https://doi.org/10.5194/nhess-21-2485-2021>.
- Sun, Y., A. K. C. Wong, and M. S. Kamel, 2009: Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.*, **23**, 687–719, <https://doi.org/10.1142/S0218001409007326>.
- Tippett, M. K., A. G. Barnston, and A. W. Robertson, 2007: Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate*, **20**, 2210–2228, <https://doi.org/10.1175/JCLI4108.1>.
- Vigaud, N., A. W. Robertson, and M. K. Tippett, 2018: Predictability of recurrent weather regimes over North America during winter from submonthly reforecasts. *Mon. Wea. Rev.*, **146**, 2559–2577, <https://doi.org/10.1175/MWR-D-18-0058.1>.
- Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>.
- , and Coauthors, 2019: Extended-range prediction. ECMWF Tech. Memo. 854, 60 pp., <https://www.ecmwf.int/node/19286>.
- Viviroli, D., M. Zappa, J. Gurtz, and R. Weingartner, 2009: An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. *Environ. Modell. Software*, **24**, 1209–1222, <https://doi.org/10.1016/j.envsoft.2009.04.001>.

- 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>.
- White, C. J., and Coauthors, 2017: Potential applications of Subseasonal-to-Seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, <https://doi.org/10.1002/met.1654>.
- , 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>.
- Yuan, X., C. Chen, X. Lei, Y. Yuan, and R. M. Adnan, 2018: Monthly runoff forecasting based on LSTM–ALO model. *Stochastic Environ. Res. Risk Assess.*, **32**, 2199–2212, <https://doi.org/10.1007/s00477-018-1560-y>.
- Zappa, M., L. Bernhard, C. Spirig, M. Pfändler, K. Stahl, S. Kruse, I. Seidl, and M. Stähli, 2014: A prototype platform for water resources monitoring and early recognition of critical droughts in Switzerland. *Proc. IAHS*, **364**, 492–498, <https://doi.org/10.5194/piahs-364-492-2014>.