Exploring hydrological system performance for alpine low flows in local and continental prediction systems

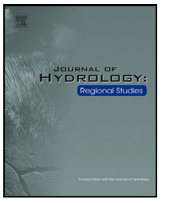
Annie Y.-Y. Chang, Maria-Helena Ramos, Shaun Harrigan, Christel Prudhomme, François Tilmant, Daniela Domeisen, Massimiliano Zappa

Journal of Hydrology: Regional Studies 56 (2024) 102056

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

journal homepage: www.elsevier.com/locate/ejrh



Exploring hydrological system performance for alpine low flows in local and continental prediction systems

Annie Y.-Y. Chang [a,b,c,*], Maria-Helena Ramos [d], Shaun Harrigan [c], Christel Prudhomme [c], François Tilmant [d], Daniela I.V. Domeisen [e,b], Massimiliano Zappa [a]

[a] Swiss Federal Institute WSL, Birmensdorf, Switzerland
[b] Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[c] European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK
[d] Université Paris-Saclay, INRAE, UR HYCAR, Antony, France
[e] University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

Keywords:
Low flows
EFAS
GR6J
PREVAH
Continental system
Local system
European alps

ABSTRACT

Study region: The European alpine space.

Study focus: Despite the advancements in hydrological modeling, there is a lack of comparative studies quantifying the performance differences between local and large-scale models, particularly in the context of drought and low flow assessment. This study addresses this gap by designing a framework to evaluate the European Flood Awareness System (EFAS), a continental system, in simulating low flow events across 101 alpine stations from 1999–2018, with detailed comparisons to two local systems — GR6J in France and PREVAH in Switzerland — at 34 of these stations.

New hydrological insights for the region: The results show that the local systems, PREVAH and GR6J, generally outperform the continental system EFAS for almost all stations. The performance gap between the systems increases as low flow conditions intensify, highlighting the importance of local systems for extreme low flow events. Despite EFAS not being specifically set up for low flows, it has shown an overall acceptable performance compared to the local hydrological systems, especially at locations that are not calibrated within the EFAS system, indicating its potential in ungauged areas. This study lays a foundation for understanding how a continental hydrological system like EFAS can complement local systems or fill the gap when a local system is unavailable to provide reliable predictions of low flow conditions.

1. Introduction

The European Alps have experienced a notable shift in hydrological patterns, particularly a change in the frequency and severity of low flow events since the end of the 20th century (Brunner et al., 2023; Bard et al., 2015). Such events exert critical pressure on ecosystems and present challenges for cross-border and local water management in the alpine region, as many economic sectors rely on water availability, including hydropower production, navigation and transportation, agriculture, and tourism. As a result, there is an increasing need for decision-makers to have access to accurate simulations and early warnings tailored to local low flow conditions.

* Corresponding author at: Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.
E-mail address: ychang@ethz.ch (A.Y.-Y. Chang).

https://doi.org/10.1016/j.ejrh.2024.102056
Received 7 July 2024; Received in revised form 31 October 2024; Accepted 2 November 2024

The landscape of hydrological modeling has evolved significantly from the first widely used rainfall-runoff model (the *Rational Method*) to conceptual models and physics-based models (Beven, 2012). The field of hydrology has also embraced the advancements of machine learning (ML) and artificial intelligence (AI) with ongoing developments to incorporate ML and AI into hydrological modeling, especially using long short-term memory (LSTM) network (Kratzert et al., 2018) as well as hybrid (AI + process-informed) systems (Slater et al., 2023; Ng et al., 2023; Wei et al., 2024).

Physics-based hydrological models are typically informed by static physiogeographical input and time-varying meteorological observations or weather—climate model outputs to provide simulations of streamflow at different time scales (typically from hourly to annual) and spatial resolutions (from lumped models at the catchment scale to fully-distributed pixel-based models) (Beven, 2012). Unlike meteorological or climate models, which are needed at global scales to provide boundary conditions for regional models and which rely heavily on global information, most hydrological models are set up locally at the catchment scale or regional level. The development of global hydrological models began in the late 1980s, aiming to map the global water balance at large scales. However, these early models lacked the accuracy required for local-scale water management and decision-making. Over time, improvements in meteorological and hydrological modeling, particularly the integration of ensemble prediction systems (EPS), advancements in precipitation forecasting, and increased computational power, led to the development of numerous continental and global hydrological systems. Several reviews provide valuable insights into the development and application of these large-scale hydrological models. For instance, Sood and Smakhtin (2015) examined the evolution of 12 global hydrological models (GHMs), highlighting structural differences, limitations in model uncertainty, data scarcity, integration with remote sensing data and spatial resolution, while Bierkens et al. (2015) provides an overview of 14 global-scale and 4 continental-scale models, illustrating the breadth of hydrological forecasting tools available. Emerton et al. (2016) investigated six state-of-the-art operational large-scale flood forecasting systems (two global and four continental) and analyzed their capability to produce coarse-scale discharge forecasts in the medium-range. In a comparison study, Crochemore et al. (2020) found comparable performance between the continental model E-HYPE and local model GR6J when forecasting seasonal streamflow anomalies. In another comparison study with local applications, Kolling Neto et al. (2023) showed that with simple corrections, weekly streamflow forecasts from continental-scale models can achieve competitive skill as those issued by local operators for Brazilian hydropower plants. Today, systems such as the World-Wide HYPE (WWH) model (Arheimer et al., 2019), the Global Flood Awareness System (GloFAS) (Alfieri et al., 2013; Harrigan et al., 2020), the NOAA Water National Water Model WRF-hydro (Johnson et al., 2023), and the African Flood and Drought Monitor (AFDM) (University of Southampton, 2024), simulate streamflows in both gauged and ungauged catchments operationally at various centers or agencies.

These advancements highlight how large-scale modeling systems offer several advantages, such as providing extensive coverage of hydrological conditions over large areas and offering longer forecast lead times, making them particularly useful for early warning systems. Their wide coverage, especially in ungauged areas, has the potential to significantly enhance the effective implementation of initiatives like Early Warnings for All (World Meteorological Organization, 2023, 2024), which aims to deliver early warning systems globally, particularly in regions lacking local infrastructure. However, their coarse spatial resolution often limits their accuracy, especially in complex terrains like the European Alps (Brunner et al., 2019). In contrast, local models can perform data assimilation more easily to improve accuracy, run faster with higher-resolution setups, and provide a more in-depth understanding of local-scale hydrological processes and responses to meteorological forcings and human activities. However, local systems may lack the flexibility to incorporate Earth observations or global datasets and may not have the capacity to cover large spatial domains (Emerton et al., 2016).

Although different model setups (e.g., local vs. global models) offer distinct advantages and limitations, making the selection of an appropriate model important, Addor and Melsen (2019) highlights a strong tendency towards legacy-driven model selection. This reflects a preference among institutions and modelers for familiar or previously implemented models, even when alternative models might offer comparable or competitive performance or be more suitable for specific project purposes. The reluctance to adopt new systems is driven by familiarity and previous use, as well as the challenges in conducting comprehensive comparative analyses of model outputs. In fact, despite the advancements in hydrological modeling, there is a lack of comparative studies that quantify the performance differences between local and large-scale models, particularly in the context of drought risk assessment. This prevalence of legacy in model selection can lead to missed opportunities to improve forecasting systems, thus (Addor and Melsen, 2019) calls for more model intercomparison studies as they can provide guidance for model selection and improve model adequacy. Both Addor and Melsen (2019) and Emerton et al. (2016) advocate for a more flexible multi-model setup as the way forward to further advance hydrological forecasting. Such approaches could combine the strengths of local and large-scale models, providing more comprehensive and accurate hydrological forecasts for a variety of applications.

This paper presents a comparative analysis of three hydrological systems operating at both continental and local scales — the European Flood Awareness System (EFAS) covering Europe and the local systems GR6J and Precipitation-Runoff-Evapotranspiration HRU Model (PREVAH) operating in France and Switzerland, respectively. Our particular focus is on evaluating and contrasting the modeling systems' capabilities in simulating low flow conditions for the European alpine region (Fig. 1). The aim is to quantify the differences in performance between the continental-scale EFAS system and the locally set-up GR6J and PREVAH systems. Identifying the strengths and limitations of each modeling system will help us understand how they can complement each other in developing a robust low flow forecasting system for the alpine region, thereby contributing to a path towards better assisting regional and local water resource management in their decision-making. By conducting a detailed comparative analysis of model simulations, this study establishes an effective framework for evaluating different local and continental hydrological systems. As outlined in Fig. 1, this workflow enhances our understanding of the different systems' performance under low flow conditions and provides a structured methodology for future studies to evaluate and select appropriate model systems for hydrological simulations across various geographical settings.

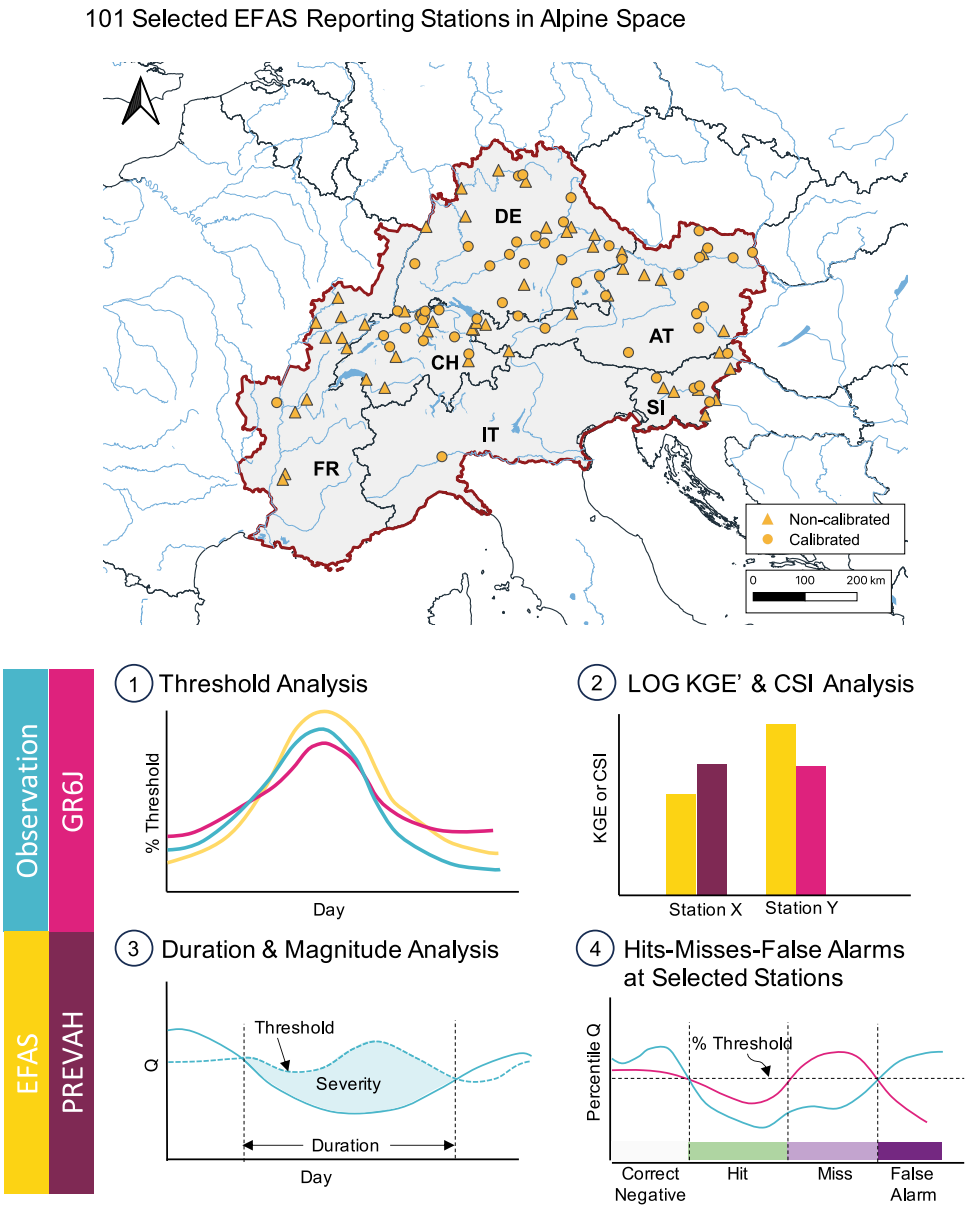


Fig. 1. Illustration of the study framework. The study area, outlined by the red line on the map, encompasses the European alpine regions of Italy (IT), France (FR), Switzerland (CH), Germany (DE), Austria (AT), and Slovenia (SI). 101 EFAS reporting stations are selected, distinguished by the calibration status within EFAS. This study carries out (1) a threshold analysis to determine the low flow event thresholds, (2) log KGE' and CSI analyses to evaluate the system performance, (3) duration and magnitude analyses, and (4) case studies at two selected stations with a focus on hits, misses and false alarms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Modeling system overview

Continental system - EFAS

In Europe, the European Flood Awareness System (EFAS, https://confluence.ecmwf.int/display/CEMS/European+Flood+Awareness+System), part of the European Union (EU) Copernicus Emergency Management Service (CEMS), has been operated by the Joint Research Centre (JRC) and four operational centers: the hydrological data collection center (Ghenova Digital), the meteorological data collection center (KISTERS AG and Deutscher Wetterdienst), the dissemination Centre (Swedish Meteorological and Hydrological Institute, Rijkswaterstaat and Slovak Hydro-Meteorological Institute), and the computational center (European Centre for Medium-Range Weather Forecasts, ECMWF), since 2012 to provide hydrological simulations and forecast twice a day at

6-hourly and kilometer scales over European river basins. EFAS has originally been designed to support and complement national flood forecasting services, in particular at trans-boundary catchments (Thielen et al., 2009), and has expanded gradually over the years including to cover flash-floods (Alfieri et al., 2013) and to simulate long-range low flow predictions (Arnal et al., 2018). Currently, EFAS delivers an extensive assessment of flood risk and issues flood notifications to the EFAS partners with a forecast lead time of up to 15 days. Based on an ensemble of state-of-the-art meteorological forecasts, EFAS hydrological simulations are obtained from the LISFLOOD model — a continental-scale, fully-distributed, physics-based hydrological model (De Roo et al., 2000). For simplicity purposes, hereafter the term "EFAS" is used to denote both the LISFLOOD model and its generated streamflow data product. The EFAS v4.0 is the version investigated in this study. It was calibrated at 1137 stations in 215 different catchments, using data from nearly 3000 river gauging stations. Calibration is performed on catchments larger than 500 km^2, with at least four years of discharge data within the calibration period (1990–2017), and a preference for 6-hourly data. Calibration in EFAS v4.0 is performed on 14 parameters using an evolutionary optimization algorithm (EA). The objective function for EFAS calibration is the modified Kling-Gupta Efficiency (KGE', see Section 3.4), tailored towards better flood simulation (Kling et al., 2012). Approximately 50% of the calibration stations achieve a KGE' value greater than 0.75, indicating high system performance. The accuracy of EFAS outputs and the system's performance are evaluated regularly, showing consistent improvements (CEMS, 2023).

The EFAS hydrological model runs continuously in time and, therefore, despite being initially developed to capture flood events, EFAS has also been investigated with respect to its ability to simulate low flow situations and capture large-scale droughts at sub-seasonal to seasonal time scales covering the range from two weeks to 3 months. Arnal et al. (2018) investigated the propagation of skill from meteorological seasonal forecasts to EFAS' seasonal forecasts. Their findings indicate that EFAS shows skill for the first month of lead time for low flow conditions. Performance analyses demonstrated the ability of EFAS seasonal outlooks to capture low flow anomalies with the example of capturing the drought that happened in Sweden in 2017 (Arnal et al., 2018). Wetterhall and Di Giuseppe (2018) found that EFAS, when forced with the merged seasonal forecast system (SEAM), shows an improved prediction of low flow conditions, with the example of the 2003 extreme low flow event. Despite these efforts, the potential of using EFAS for low flow simulation remains under-investigated, especially in the context of the changing hydrological landscape of the European Alps, as well as its gap and complementary added value to operational models already locally set up and running for low flow prediction and risk-based decision-making.

Local systems - PREVAH & GR6J

The two "local systems" we have selected to compare with EFAS are the GR6J and PREVAH models. The GR6J model is one of the models included in the operational low flow forecasting tool PREMHYCE (Nicolle et al., 2020; Tilmant et al., 2023) covering France, managed by France's National Research Institute for Agriculture, Food and Environment (INRAE). The model is a lumped, conceptual model, calibrated against local observations over the period of 1963–2020, and designed to improve accuracy in low flow prediction, while maintaining simplicity in capturing the rainfall–runoff relationship, in particular, in medium-size catchments (Nicolle et al., 2014, 2020; Pushpalatha et al., 2011). PREVAH, operated by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) in Switzerland, is a semi-distributed, process-based model (Gurtz et al., 1999; Viviroli et al., 2009). It is specifically designed for high-resolution simulations in mountainous areas, making it adept at handling the complex hydrological dynamics of the Swiss Alps. The model PREVAH, calibrated over the period of 1984–1996, is operational and responsible for providing hydrological outlooks on the open platform drought.ch (Zappa et al., 2014; Speich et al., 2015). Unlike EFAS, which is used for flood monitoring and forecasting, both local systems are currently used operationally for low flow and drought monitoring and predictions. With their local focus by design, for operations in their respective countries, the local systems are expected to provide a more targeted calibration or a more in-depth understanding of the local-scale hydrological processes than EFAS and therefore better accuracy.

Configuration comparison

Besides the different geographical coverage, the modeling systems analyzed in this study also differ from one another in terms of model structure, temporal and spatial resolution, calibration strategy, meteorological input datasets, and other configurations (see Table 1 for a comparison), which can all contribute to differences in their performance. In this study, we do not focus on the comparisons of every aspect of the systems. However, bearing these differences among the three systems in mind, we take a pragmatic approach to compare their ability to reproduce observed patterns in streamflow during low flow periods, with a focus on understanding the performance gaps between the local and the continental systems.

3. Methodology

The methodology to evaluate model performance used in this study follows the workflow outlined in Fig. 1. We first select stations based on the availability of observational streamflow data and evaluate EFAS' ability to capture their overall hydrological behavior. At those stations in France and in Switzerland where local systems are available, we start by conducting a percentile analysis to determine the low flow event thresholds and investigate the duration and magnitude of these events in EFAS and the respective local systems. We use different verification metrics to evaluate the performance of each system by comparing model simulations with observations. We conclude the study with two case studies on one station in the French dataset and one station in the Swiss dataset for a closer look at how the systems perform in terms of hits, misses, and false alarms.

Table 1
Summary of the main features of the EFAS, PREVAH and GR6J systems.

	EFAS	PREVAH	GR6J
Operated by	CEMS-Flood	WSL	INRAE
Version	4.0	Extracted on 16.03.22	airGR 1.7.0
Type	Full-distributed, physics-based	Semi-distributed, process-based	Lumped, process-based
Temp. Resol.	6-h	Daily	Daily
Spatial Resol.	5 km × 5 km	500 m × 500 m	Catchment-scale
Coverage	Europe	Switzerland	France
Calibration metric	modified KGE'	NSE	KGE, NSE
Meteo inputs	European Meteorological Observations (EMO5) (Thiemig et al., 2022)	MeteoSwiss (Speich et al., 2015)	Météo-France SAFRAN (Quintana-Seguí et al., 2008; Vidal et al., 2010)

3.1. Data and station selection

To perform system comparisons between EFAS and the two local systems, PREVAH and GR6J, we use historical streamflow simulation data from these models along with observational streamflow data. The selected study period is from 1999 to 2018.

We extract historical streamflow simulations at EFAS *reporting stations*, which are gauged locations with streamflow measurements in the CEMS-Flood database. While all model calibrations in EFAS are carried out at reporting stations, not all reporting stations are calibrated. This depends on the quality of the observation data and the catchment area as described in the section above. Our selection of stations is primarily dictated by the availability of continuous observational streamflow data with no missing data for longer than seven consecutive days. There are more than 500 EFAS reporting stations within the European Alpine space, and our selection process results in a total of 101 stations: 1 in Italy (IT), 12 in France (FR), 22 in Switzerland (CH), 29 in Germany (DE), 26 in Austria (AT), and 11 in Slovakia (SI). Among these, 73 stations are calibrated while 28 stations are not calibrated within EFAS. Focusing on the region where we have the local systems, one out of the 12 stations in France and 15 out of the 22 stations in Switzerland are calibrated in EFAS. Finally, we note that all stations selected in France (Switzerland) are calibrated in the GR6J (PREVAH) system.

To retrieve daily observational streamflow data at these reporting stations, we use several databases, including the Alpine Drought Observatory (Žun et al., 2022), the Global Runoff Data Centre (GRDC) (Global Runoff Data Centre (GRDC), 2023), and the HydroPortail in France (Dufeu et al., 2022). In addition to EFAS historical streamflow data, we extract daily streamflow simulations from the local systems PREVAH and GR6J for the selected stations in Switzerland and France, respectively. The EFAS data are publicly available on the Copernicus Climate Change Service (C3S) Climate Data Store (CDS), and the GR6J and PREVAH data are available upon request.

For EFAS, we first transform the 6-hourly dataset into a mean daily streamflow time series. For all data, we apply a 7-day moving average to reduce random variations in the data. This approach improves the day-to-day consistency of the time series, allowing for clearer detection of extended low flow periods for the evaluation of the system (Caillouet et al., 2017). The moving window is set to 7 days to match the weekly streamflow cycle as many catchments are influenced by hydropower production regulations (Meile et al., 2011).

3.2. Study area

The study area encompasses the European Alpine space (outlined by red border in Fig. 1), featuring a diverse range of catchments in terms of size, elevation, and other hydrological characteristics. This region provides an ideal setting to evaluate the performance of continental and local hydrological models across varying conditions. Tables S1 to S3 in the Supplement provide summaries of the main characteristics of the selected catchments in the study area. The catchment size ranges from 75 to 101,800 km^2, with the selected catchments in Switzerland showing a wide range, from 1150 to 35 800 km^2, and catchments in France up to 2,038 km^2. Given the wide range of the upstream areas of the catchment, for comparison purposes, daily streamflow is converted to specific discharge normalized by area and has a unit of mm. With the study area located in a mountainous region, the stations vary in height with median elevations starting from 98 m up to 1078 m. The highest station is the station Martinsbruck on river Inn, at the border of Switzerland and Austria. This elevation diversity contributes to diverse climatic conditions and results in different hydrological regimes within the study area, including nival (snow-dominated) and pluvial (rainfall-dominated) systems, which are discussed in more detail in the section below.

3.3. Climatology and low flow thresholds

A streamflow climatology is computed to gain insights into the typical patterns and variations of streamflow over time, as well as to establish criteria defining low flow periods. To create this climatological distribution, we first select, for each day of the year within the 20-year period (1999–2018), all the (7-day smoothed) daily streamflow data within a 31-day window that is centered on

that day, spanning 15 days before and after the central day (Brunner et al., 2022). By comparing the climatologies of the catchments in France and Switzerland, we observe a distinct difference in hydrological regimes. Catchments in Switzerland typically exhibit a nival regime (snow-fed), where streamflows are predominantly driven by snow and ice melt. This nival regime is characterized by a bell-shaped hydrograph, with streamflow increasing from early in the year, peaking around June or July due to snowmelt, and then gradually decreasing towards year-end (an example can be found in Fig. 5a). On the other hand, catchments in France in our sample are more typically influenced by a pluvial regime (rain-fed), where rainfall is the primary driver of streamflow. These catchments experience higher flows during winter and fall, attributed to increased rainfall, and lower flows during spring and summer, reflecting the seasonal precipitation patterns (an example can be found in Fig. 5b).

We compute low flow thresholds for each day of the year based on specific percentiles of the climatological distribution, focusing on streamflows within the lower 25th, 15th, and 5th percentiles of this distribution (Hisdal et al., 2024). These thresholds are used to identify and characterize low flow occurrences. For any given day, if the streamflow falls below the threshold, it is marked as a low flow day. Each day is analyzed individually without applying any persistence criterion or pooling of events. We distinguish between observational climatology, derived from historical observed data, and model climatology, derived from simulations, each providing distinct criteria for defining low flow events. Defining low flow events using the respective climatology allows us to capture the onset of these events and the deficit relative to each system (either the model simulation or the observations). With this approach, we can focus on evaluating the models' systematic errors, like their tendency to overestimate or underestimate low flows. This method effectively isolates the biases of the models but may not fully account for random errors or the exact matching of low flow magnitudes.

3.4. Verification metrics

To evaluate model performance, we employ two established goodness-of-fit measures. The first one is the log modified Kling-Gupta Efficiency (KGE') for the entire time series. The second metric is the critical success index (CSI) for low flow events only.

Modified Kling–Gupta efficiency (KGE')

Introduced by Gupta et al. (2009), KGE is a widely accepted hydrological model evaluation metric that addresses some limitations associated with the traditional Nash–SutcliffeEfficiency (NSE) (Schaefli and Gupta, 2007). Kling et al. (2012) introduced a modified version of this index (KGE') to prevent cross-correlation between the bias and the variability ratios. The KGE' is calculated using the following formula:

$$KGE' = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{1}$$

$$\beta = \frac{\mu_s}{\mu_o}, \tag{2}$$

$$\gamma = \frac{\sigma_s/\mu_s}{\sigma_o/\mu_o}, \tag{3}$$

where r is the Pearson correlation coefficient, and hence r evaluates the linear correlation between the observed and simulated flows and provides an indication of how well the pattern of the simulated sequence of flows matches the pattern of the observed flow over time. β is the bias ratio of the mean simulated flow to the mean observed flow, and it assesses the model's accuracy in estimating the flow magnitude. Lastly, γ is the ratio between the coefficient of variances of the simulated flow and that of the observed flow, and it ensures that the model adequately represents the range of flow variations. The KGE' provides a comprehensive assessment by simultaneously considering correlation, bias, and variability in the model outputs compared to observed data.

Studies have suggested the use of log-transformed streamflow data to compute the verification metric when evaluating low flow conditions (Krause et al., 2005; Pechlivanidis et al., 2014). Here we apply a logarithmic transformation to both observed and simulated flow data before calculating the KGE'. This transformation amplifies the differences in smaller values, making it more effective than the regular KGE' in low flow analysis by emphasizing the lower range of the flow spectrum. For this reason, only log KGE' is applied in this study.

The ideal value for each of the KGE' components is 1 and the range of log KGE' values is between $-\infty$ to 1. A value of 1 indicates a perfect match between the logarithms of the observed and simulated data, indicating an optimal model performance. The further away from a value of 1, the bigger the discrepancy is between the observed and simulated data. We consider a log KGE' value greater than -0.41 to be acceptable and indicative of model performance better than the mean flow benchmark, as suggested by Knoben et al. (2019). While using the mean flow as a benchmark with a KGE' value of -0.41 may not represent a particularly challenging threshold, Knoben et al. (2019) demonstrated that negative KGE values do not necessarily indicate worse model performance than the mean flow, despite the common use of zero as the threshold for acceptable performance. Knoben et al. (2019) shows that using a mean flow benchmark yields a KGE' value of approximately -0.41 ($KGE'_{clima} = 1 - \sqrt{(1+0+1)} = -0.41$), where the correlation coefficient is zero, $\mu_s = \mu_o$, and the standard deviation of mean flow is also zero. These assumptions are valid in the case of the log transformation. Additionally, the same benchmark was employed to evaluate GloFAS in Harrigan et al. (2020), providing further precedent for its use in this study.

Critical success index (CSI)

The Critical Success Index (CSI) (Schaefer, 1990) is another commonly used metric to evaluate the accuracy of a model to detect events. Different from KGE' or log KGE', which evaluates the continuous time series, the CSI was developed to evaluate the predictive power of dichotomous events. In our case, we use the CSI to assess the low flow event identification ability of a simulation with reference to observations, where events are grouped into four combinations of simulations and observations, based on the frequency of simulation and occurrences in the observations. The four combinations are:

- "hit" - event simulated to occur, and also observed;
- "false alarm" - event simulated to occur, but not observed;
- "miss" - event simulated not to occur, but observed;
- "correct negative" - event simulated not to occur, and also not observed.

The CSI is defined as:

$$CSI = \frac{\text{hits}}{\text{hits} + \text{false alarms} + \text{misses}} \tag{4}$$

The value of CSI ranges from 0 to 1, with 1 indicating perfect simulation/predictive power, and 0 indicating no power. A key strength of the CSI is that it accounts for both false alarms and misses, providing a balanced evaluation. In addition, the CSI disregards the correct simulation of non-events (or "correct negatives"), which takes up the majority of the occurrences. Hence, the CSI is a suitable metric for investigating rare and extreme events. We consider a CSI of 0.5 as an acceptable value because it implies that the number of hits equals the combined number of misses and false alarms. The CSI values are computed for thresholds at 25th, 15th, and 5th percentiles.

To complement CSI and to have a deeper understanding of what contributes to a high or low CSI value at a given station, we also compute the probability of detection (POD) and the false alarm rate (FAR):

$$POD = \frac{\text{hits}}{\text{hits} + \text{misses}} \tag{5}$$

$$FAR = \frac{\text{false alarms}}{\text{hits} + \text{false alarms}} \tag{6}$$

3.5. Duration and magnitude

Duration and magnitude are important aspects when evaluating low flow events. The duration of a low flow event is the number of days between the first day and the last day where the streamflow is below the threshold. Based on duration, we categorize events into six categories: shorter than 3 days, between 3 days and a week, longer than a week but shorter than 2 weeks, longer than 2 weeks but shorter than 3 weeks, longer than 3 weeks but shorter than 4 weeks, and longer than 4 weeks. These duration categories are selected considering the usefulness of the information to decision-makers; for example, events that are shorter than 3 days can be considered occurrences that do not have an impact on the major operational decisions.

For each low flow event, we also compute the magnitude, which is the total amount of water in deficit (also known as severity) relative to the threshold, divided by the duration of the event. For comparison purposes across different catchments, we compute low flow severity in units of mm, and magnitude in units of mm/day

3.6. Case study analyses

We examine two distinct stations: Rheinhalle in Switzerland and L'Ognon à Pesmes in France (see red arrows in Fig. 2a). Their main characteristics are summarized in Table 2. Both stations are representative of the typical flow regimes in their region, as described earlier, that is, a nival regime (governed by snow melt) for station Rheinhalle in Switzerland (see Fig. 5a), and a pluvial regime (governed by rainfall events) for station L'Ognon à Pesmes in France (see Fig. 5b). We choose the station Rheinhalle for its trans-national location at the border of Switzerland, Germany, and France and its important economic role in transportation and navigation on the River Rhine. The station L'Ognon à Pesmes represents a small catchment station, with its area aligning closely with the 25th percentile of catchment sizes within the study.

4. Results

4.1. Overall performance assessment with log KGE'

We first evaluate the entire time series of simulated flows (1999–2018) at all 101 stations using log KGE' to capture a holistic view of EFAS' ability to simulate the observed hydrological behavior, with more weight on low flows. The spatial distribution of log KGE' values across the study region (see Fig. 2a) shows a heterogeneous pattern of model performance. While most locations exhibit high log KGE' values (with a median of 0.81) indicating very good model fit, a few stations show poor performance, suggesting challenges in simulating flows at these specific catchments.

Notably, four stations display log KGE' values lower than −0.41 (colored in orange in Fig. 2a), suggesting that the model is performing more poorly than a simple mean flow prediction at these catchments (Knoben et al., 2019). These four stations are:

Table 2
Summary of the main characteristics of the selected stations Rheinhalle (Switzerland) and L'Ognon à Pesmes (France).

	Rheinhalle	L'Ognon à Pesmes
EFAS ID	#353	#3228
Country	Switzerland	France
River	Rhine	Ognon
Area (km^2)	35,897	2,038
Station elevation (m)	246	185
Mean annual specific discharge Q_A (mm)	932	305
25th percentile annual specific discharge (mm)	834	265
15th percentile annual specific discharge (mm)	809	243
5th percentile annual specific discharge (mm)	722	208

- **Neudorf bei Ilz** on river Ilzbach (EFAS ID #2139, AT, area: 150 km^2, elevation: 278 m),
- **Imbach** on river Krems (EFAS ID #2141, AT, area: 375 km^2, elevation: 231 m),
- **Neuenstadt** on river Brettach (EFAS ID #2152, DE, area: 75 km^2, elevation: 159 m),
- **Martinsbruck** on river Inn (EFAS ID #2564, CH, area: 2025 km^2, elevation: 1076 m).

These stations are not calibrated within EFAS and are geographically dispersed, suggesting that poor performance is not confined to a particular area. The strongly negative log KGE' values are likely due to the log transformation, as no KGE' values without the log transformation fall below the −0.41 threshold (not shown).

Visual inspection of Fig. 2a shows no consistent relationship between log KGE' values and catchment position along rivers or EFAS calibration status. Stations in close proximity tend to exhibit similar performance irrespective of their calibration (e.g. two clusters of stations with high performance in south-eastern Germany and northern Switzerland and a cluster of stations with reduced model performance in eastern Austria), implying that large-scale processes or features like lakes and dams may influence EFAS performance more than calibration at individual stations. Additionally, regional landscape and geological characteristics could further impact the model's performance by affecting processes that might not be well-represented in the system, thus contributing to the observed variations in model accuracy (Horton et al., 2022).

Fig. 2b takes a closer look at the 34 stations located in France and Switzerland, comparing their log KGE' values in EFAS with those in the local systems GR6J and PREVAH. Generally, the local systems outperform EFAS; however, at a few stations, EFAS performs comparably or even better than the local models, irrespective of its calibration status. These stations include:

- **Thun** on river Aare (EFAS ID #339, CH, area: 2650 km^2, elevation: 548 m),
- **Diepoldsau** on river Rhine (EFAS ID #354, CH, area: 6275 km^2, elevation: 410 m),
- **Soyans** on river Roubion (EFAS ID #2069, FR, area: 317 km^2, elevation: 200 m),
- **Saint-Hippolyte** on river Dessoubre (EFAS ID #3237, FR, area: 550 km^2, elevation: 380 m).

The median of the log KGE' differences calculated at each station indicates that EFAS's performance is approximately 12.5% poorer than that of the local systems. At six stations, the local systems outperform EFAS by more than 50%, particularly at non-calibrated EFAS stations, highlighting the impact of calibration on model performance. Analysis of KGE' components (Figure S1 in the Supplement) suggests that biases in mean flow (Beta component) and variability (Gamma component) contribute to EFAS's lower log KGE' values at these stations. Overestimation of mean flow and streamflow variability leads to discrepancies between simulated and observed flows.

Grouping the stations by EFAS calibration status (Fig. 2c) shows greater variability in log KGE' among non-calibrated EFAS stations compared to the local systems, suggesting that EFAS's performance is more sensitive to catchment characteristics when not calibrated. Nevertheless, some non-calibrated EFAS stations achieve log KGE' values comparable to those of the local systems (refer to panel (b)), indicating EFAS's potential utility in ungauged or data-sparse regions. Furthermore, although all local system stations are calibrated, their log KGE' values are generally higher at EFAS-calibrated locations than at non-calibrated ones. This implies that factors other than calibration — such as catchment characteristics — are influencing performance at these stations within the local systems.

4.2. Low flow detection assessment with CSI

To assess the performance of the different systems in simulating low flows, we analyze the CSI values at the selected stations in France and Switzerland for EFAS and the local systems at low flow thresholds of the 25th, 15th, and 5th percentiles (see Fig. 3). As expected, model performance deteriorates for all systems as the threshold lowers from the from 25th to 5th percentiles, with more stations falling below the CSI benchmark of 0.5, indicating decreased ability to simulate extreme low flow events. This trend may also be influenced by fewer available data points at lower thresholds, potentially leading to less stable CSI statistics.

Consistent with the log KGE' analysis (Fig. 2), the CSI analysis shows that local systems outperform EFAS at most stations. However, at certain stations, EFAS performs much more poorly than the local system based on the log KGE' evaluation, exhibiting negative values, but the performance gap is less pronounced in the CSI metric. This is likely due to that CSI assesses the model's ability to capture the occurrence of events rather than focusing on flow bias and variability (see Section 4.1 for the breakdown of KGE' components). These stations include:

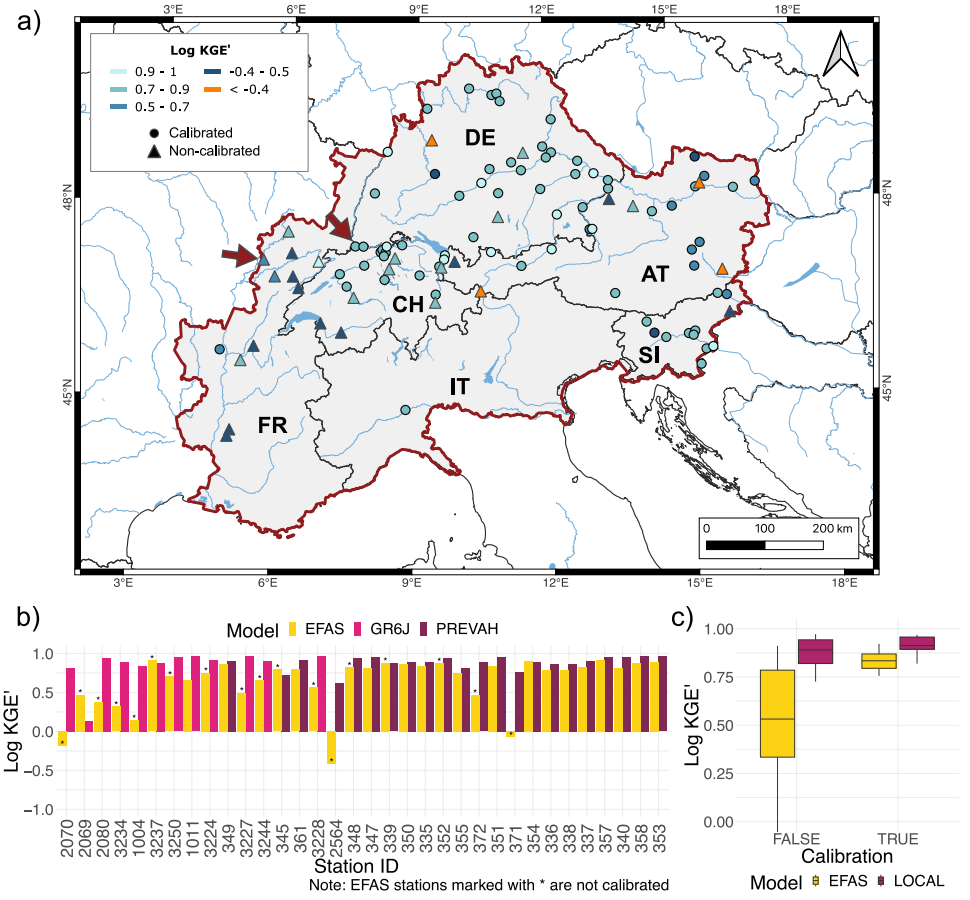


Fig. 2. Log KGE' analyses from 1999–2018: (a) Log KGE' values of EFAS simulation at all 101 selected stations in the Alpine space. Stations with a local system available for further comparisons are entirely located in France and Switzerland. Calibrated EFAS stations are denoted by dots (73), and non-calibrated EFAS stations are denoted by triangles (28). The red outline defines the boundary of the alpine space. Red arrows indicate the two case study stations: Rheinhalle in Switzerland and L'Ognon à Pesmes in France. (b) Log KGE' of the EFAS simulation at 34 selected stations in Switzerland and France, in comparison with those of the local systems PREVAH and GR6J, respectively. (c) Log KGE' comparison for the same 34 stations in panel (b) but grouped based on EFAS calibration status. 16 stations are calibrated (labeled "TRUE") and 18 are non-calibrated (labeled "FALSE") in EFAS. Note that all stations at these locations in the local systems are calibrated. Outliers are not shown. Log KGE' range: $-\infty$ to 1. Perfect score: 1. Acceptable Log KGE' values: > -0.41 (Knoben et al., 2019). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- **Porte du Scex** on river Rhone (EFAS ID #371, CH, area: 5300 km^2, elevation: 377 m),
- **Souspierre** on river Jabron (EFAS ID #2070, FR, area: 75 km^2, elevation: 298 m),
- **Martinsbruck** on river Inn (EFAS ID #2564, CH, area: 2025 km^2, elevation: 1076 m).

As the low flow threshold lowers from 25th to 5th percentiles, EFAS's performance deteriorates at a slightly higher rate than that of the local systems, resulting in a larger performance gap for more intense low flow conditions. The increasing performance difference suggests the need for a local system with a more detailed calibration strategy and more precise meteorological forcing data when it comes to capturing more intense low flow events. Notably, the local systems tend to also perform poorly at stations where EFAS has poor performance (e.g. stations #345, #2564, #355, #372, and #371). Fig. 3 further indicates that model performance does not systematically vary with catchment size, nor do the differences in performance between the models.

Grouping stations by EFAS calibration status (Fig. 3b, d, and f) shows that at the 25th percentile threshold, EFAS can achieve a median CSI value just above 0.5, even for the non-calibrated stations. At a lower threshold of 15th percentile, only the calibrated stations have a median CSI value of 0.5, and at the 5th percentile threshold, neither calibrated nor non-calibrated EFAS stations reach a median CSI of 0.5. Nevertheless, at all thresholds, at least 25% (top IQR) of EFAS stations have CSI values within the range of the local systems, suggesting EFAS's potential utility even at non-calibrated stations.

For further analysis, we select the 15th percentile threshold to define low flow events, as model errors are expected to change systematically with threshold levels. Complementary results for POD and FAR are provided in Figures S2 and S3 in the Supplement. The generally higher CSI values of the local systems result from higher POD and lower FAR. While EFAS does not outperform the local systems, it achieves POD and FAR levels within the range of the local systems, even at non-calibrated stations.

Fig. 3. Panels (a, c, e) - CSI values of EFAS simulation at 34 selected stations in Switzerland and France, in comparison with those of the local systems PREVAH and GR6J, respectively. Panels (b, d, f) - CSI comparisons at the same 34 stations, but grouped based on EFAS calibration status. 16 stations are calibrated (labeled "TRUE") and 18 are non-calibrated (labeled "FALSE"). CSI range: 0 to 1. Perfect score: 1. Top panel: 25th percentile threshold, middle panel: 15th percentile threshold, bottom panel: 5th percentile threshold.

4.3. Event duration & magnitude

Capturing the duration and magnitude of low flow events is crucial for accurate hydrological evaluations and decision-making. In Fig. 4, we compare the number of occurrences and mean magnitudes of events with different durations for EFAS, local systems, and observations.

Stations in Switzerland (Fig. 4a): EFAS simulates more events than observed in most duration categories other than events shorter than 3 days and longer than 4 weeks. EFAS better simulates shorter events (less than one week) compared to PREVAH but is less accurate than PREVAH for longer durations, where PREVAH aligns more closely with observations. Notably, at the 15th percentile threshold, PREVAH captures long events up to 154 days (observed: 158 days), while EFAS's longest simulated event is 90 days. Regarding magnitudes, excluding events shorter than 3 days, EFAS matches observed magnitudes for events less than 2 weeks. However, for events longer than 2 weeks, EFAS begins to overestimate magnitudes, with discrepancies increasing as event duration increases. PREVAH exhibits similar behavior but starts overestimating at the 3-week duration. Compared to EFAS, the discrepancies between PREVAH and observations are smaller (except for 2-week events) and remain relatively consistent across all durations.

Stations in France (Fig. 4b): The difference between EFAS and GR6J is more pronounced. EFAS simulates more short events (less than 2 weeks) than both GR6J and observations. For 3-week events, both systems simulate similar numbers as observed. For events longer than 4 weeks, GR6J closely matches observations, while EFAS underestimates the occurrences (about half of observed long events). The lack of calibration in EFAS at these stations in France may contribute to its lower performance. GR6J effectively captures the number of longer events, with its longest event at 165 days (observed: 156 days), whereas EFAS's longest event is 82 days. EFAS's high number of short events results from false alarms and misses that fragment long events into shorter ones (see example in Fig. 7). Examining magnitudes at stations in France, EFAS overestimates magnitudes for events up to 2 weeks compared to observations. EFAS simulates events with relatively consistent magnitudes for events lasting more than one week, but the observed magnitudes increase significantly from 2 to 3 weeks, leading to a large discrepancy in the 3-week (15–21 days) category. Compared to EFAS, GR6J maintains a more consistent gap with observations across all event durations.

4.4. A closer look: Case studies for selected stations

For the two selected stations, Rheinhalle in Switzerland and L'Ognon à Pesmes in France, Fig. 5 shows the differences between the climatology-based low flow event threshold at 25th, 15th, and 5th percentiles of the EFAS system (yellow lines) and the local systems (dark magenta for PREVAH in panel a and light magenta lines for GR6J in panel b), against the observed climatology in the blue background.

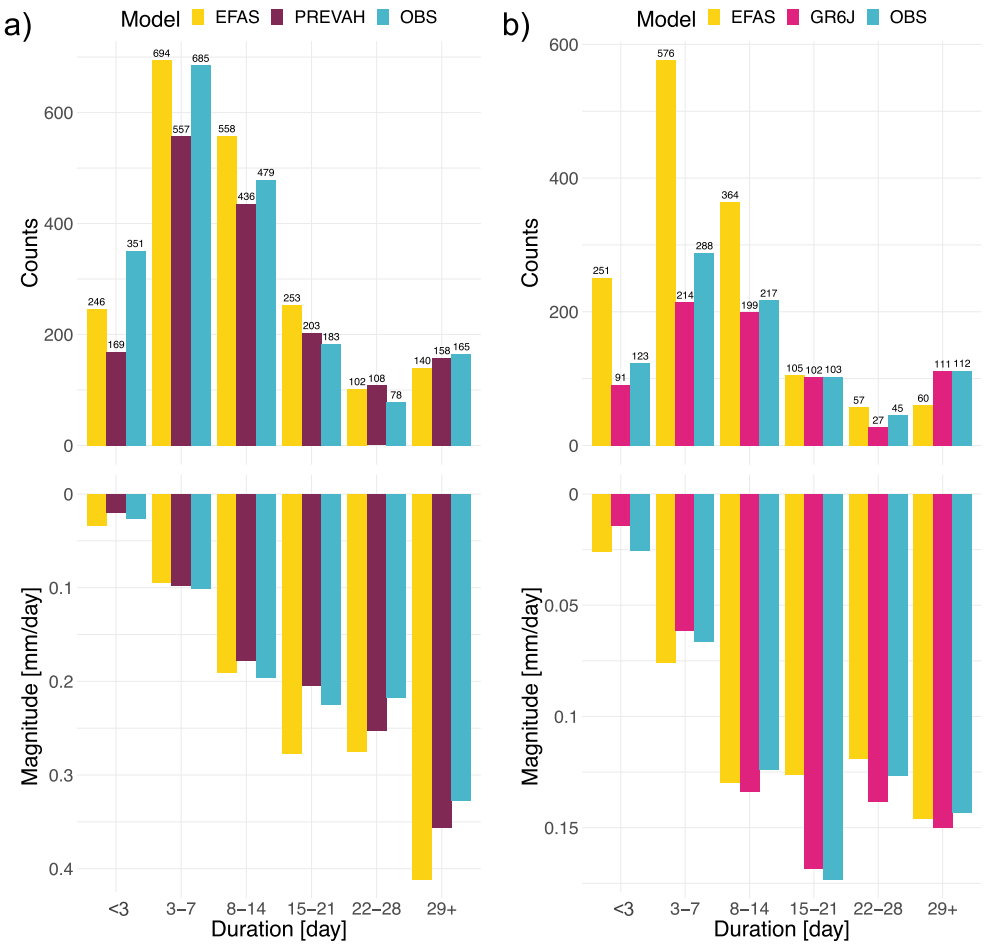


Fig. 4. Comparison of occurrences for events of different durations from 1999–2018, and the corresponding mean magnitudes in units of mm/day, the specific discharge normalized by area. (a) catchments in Switzerland and the local system is PREVAH. (b) catchments in France and the local system is GR6J. Low flow threshold: 15th percentile. Note the different scales in the two sub-figures.

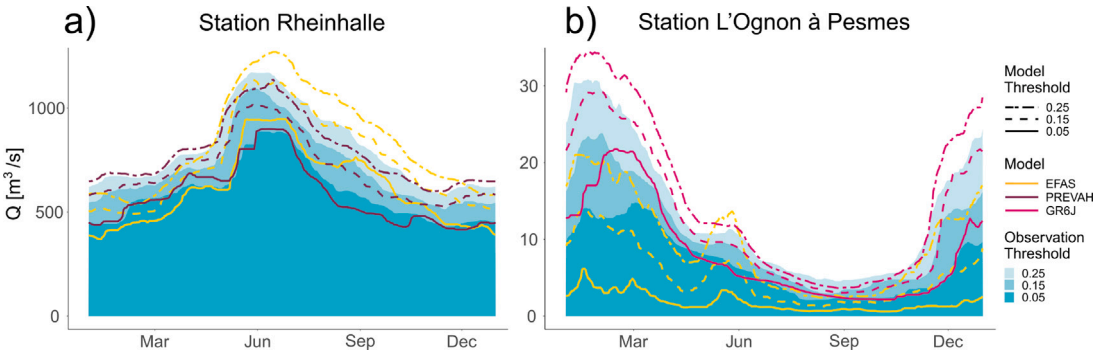


Fig. 5. 25th, 15th, and 5th percentile climatological low flow thresholds in panel (a) at Station Rheinhalle (35,897 km^2) for observations in shaded blue background, EFAS in yellow lines and PREVAH in dark magenta lines; in panel (b) at station L'Ognon à Pesmes (2,038 km^2) for observations in shaded blue background, EFAS in yellow lines and GR6J in light magenta lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

At both stations — Rheinhalle and L'Ognon à Pesmes — EFAS and the local systems capture the general patterns of observed streamflow but exhibit contrasting tendencies in simulating low flows. At Rheinhalle, EFAS underestimates flows in winter and early spring and overestimates low flows in summer and autumn, while PREVAH aligns more closely with observations, showing

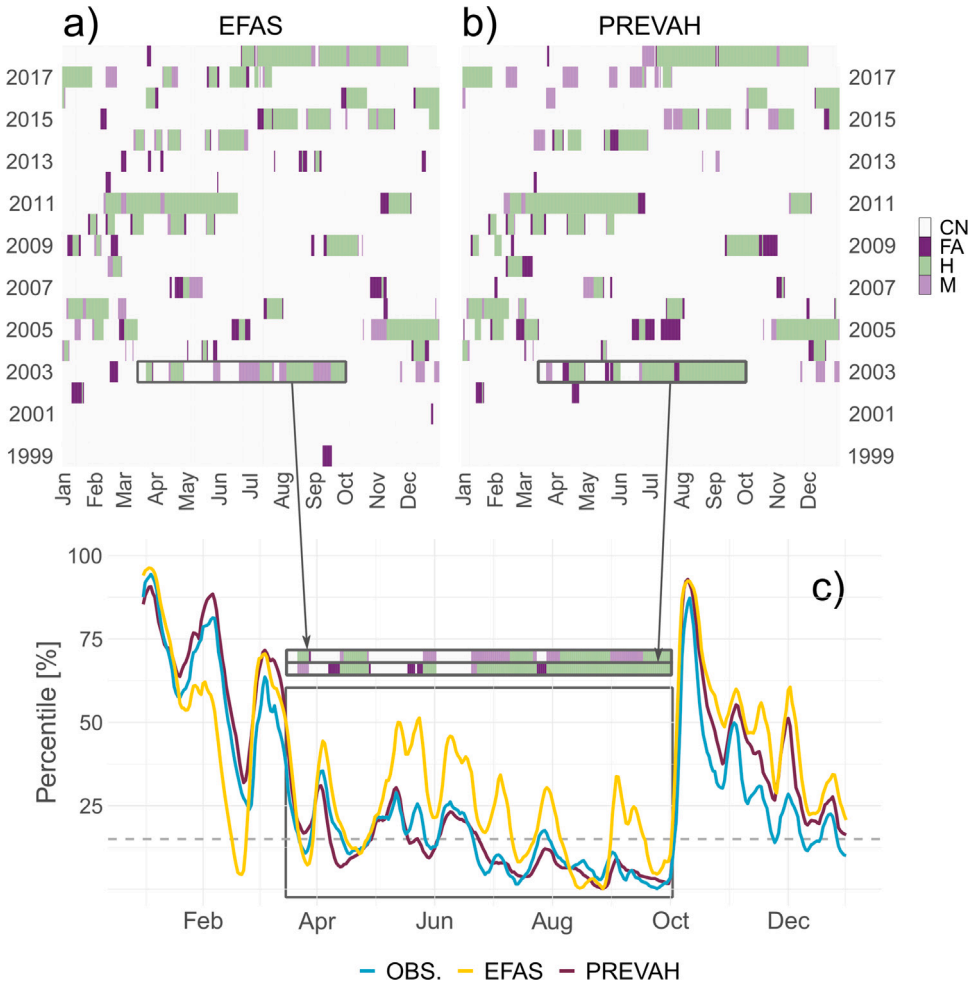


Fig. 6. Case study at station Rheinhalle (35,897 km^2, Switzerland). (a) and (b) mark the false alarms (FA, dark purple), hits (H, green), misses (M, light purple), and correct negatives (CN, light gray) for each day of the year from 1999–2018, for EFAS and PREVAH, respectively. Low flow threshold: 15th percentile. (c) Climatological percentile of daily streamflows for observations and model simulations in 2003. The gray dashed line indicates the low flow threshold at the 15th percentile. The boxed area highlights the long-lasting summer low flow event. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the opposite pattern. At L'Ognon à Pesmes, EFAS thresholds are generally below observations with notable deviations with an exceptionally high peak in late spring/early summer for the 25th percentile threshold, whereas GR6J thresholds are generally above observations except for the 5th percentile during mid-summer to late autumn. These differences highlight the contrasting responses of EFAS and the local models to the hydrological regimes at each station, reflecting their varying abilities to simulate low flow dynamics.

Figs. 6 and 7, depict the daily classification of hits (green), misses (light purple), false alarms (dark purple), or correct negatives (light gray) for EFAS in panel a and for the local system in panel b throughout the study period (1999–2018). These visuals effectively illustrate the models' abilities to capture observed low flow periods.

At Rheinhalle (Fig. 6), both EFAS and PREVAH generally simulate low flows well, with more hits than misses or false alarms, and effectively capture wet or drought-free periods (prevalence of correct negatives). However, they exhibit different distributions of misses and false alarms without a consistent seasonal pattern, likely due to the moving threshold used to define low flows (by construction, exactly 15% of the periods are classified as low flow/droughts). During the severe heatwave and drought of summer 2003 (Fig. 6c), EFAS fails to simulate most low flow days, displaying a "flashy" behavior and missing the long-lasting low flow event. In contrast, PREVAH captures the event's duration with a slight (3-day) delay at the onset but has more misses towards the end of the highlighted summer period.

At L'Ognon à Pesmes (Fig. 7), the differences between EFAS and the local system are more pronounced than at Rheinhalle station. EFAS generates more false alarms and misses compared to GR6J. Similar to the comparison at Rheinhalle, EFAS simulates the long-lasting low flow period more poorly compared to the local system. Since the analysis is based on relative low flow conditions, the systems are not penalized for simulating flows lower than the observations, as long as they accurately capture the timing when

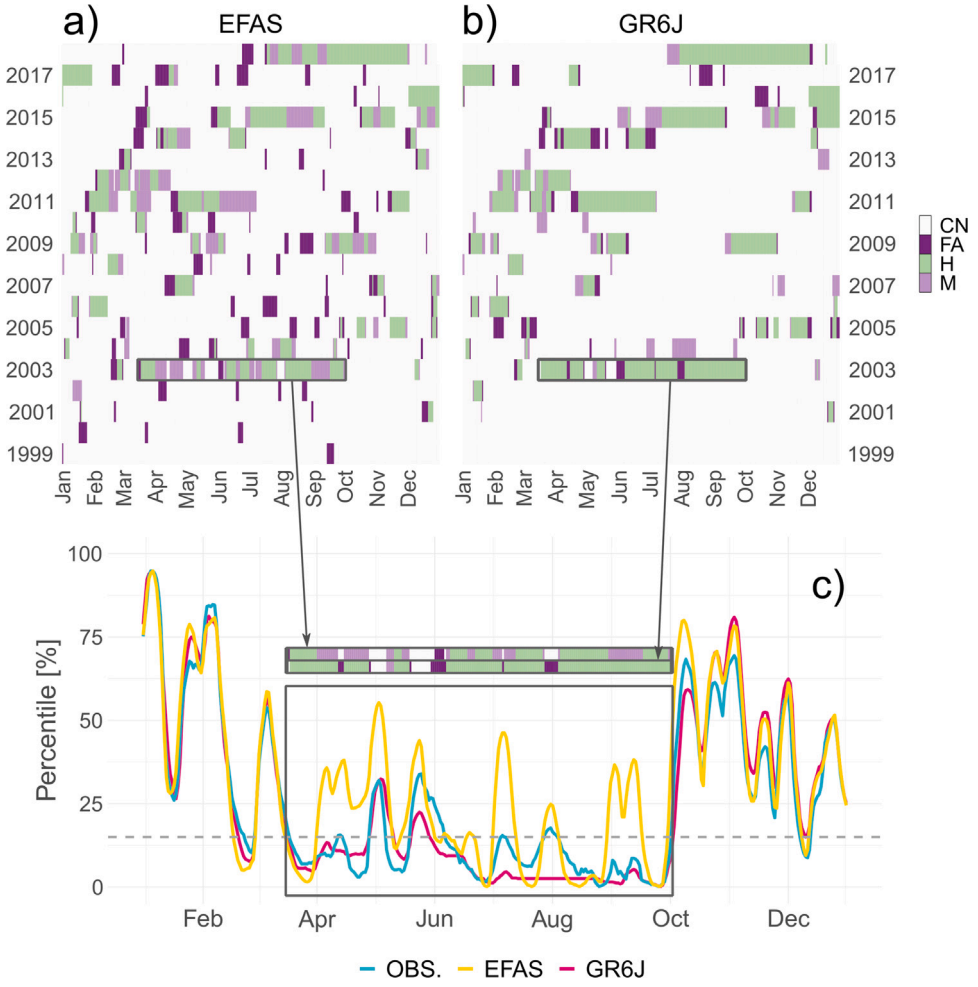


Fig. 7. Same as Fig. 6 but for station L'Ognon à Pesmes (2,038 km^2, France). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

observations fall below their threshold. During the summer period of 2003, while both models capture the timing of observed peaks, EFAS has more difficulty in capturing the correct flow percentile magnitudes, resulting in more misses. Both systems perform better in autumn and winter than in spring and summer; during the latter seasons, EFAS overestimates flow percentiles, whereas GR6J underestimates them.

Interestingly, despite the different climatologies and hydrological regimes, both catchments exhibit similar patterns in low flow occurrences and model performance. This suggests that certain challenges in simulating low flow events are consistent across different hydrological settings.

4.5. Specific analysis of dry years

In the previous section, we investigate how the systems behave during the long-lasting low flow event in 2003 at two selected stations. To gain a broader understanding of system performance during dry years across all stations, we now present an analysis focusing on the driest years of our study period for all selected catchments. For each year from 1999 to 2018, we assess the total number of days when observed daily streamflow fell below the 15th percentile threshold at each station. Based on this analysis (Figure S4 in the Supplement), we select the years 2003, 2011, and 2018 for an in-depth examination, as they exhibited the most extensive low flow periods across the majority of stations. The results in Fig. 8 show that for the three selected dry years, the local systems generally simulate fewer false alarms, more hits, and fewer misses than EFAS (Fig. 8).

To further understand the models' behavior during dry years, we analyze the standard deviation of streamflow percentiles, as shown in Figures S5 and S6 in the Supplementary Material. In the three selected dry years (2003, 2011, and 2018), EFAS exhibits a higher standard deviation of streamflow percentiles compared to observations and the local systems. In Switzerland, this increased variability is most notable in 2003, while in France, EFAS shows higher standard deviations in all three years, especially in 2003

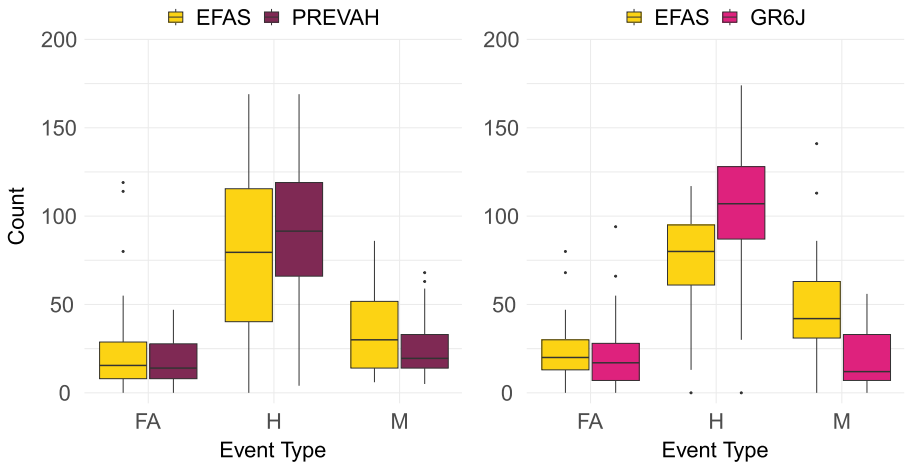


Fig. 8. Number of false alarms (FA), hits (H), and misses (M) associated with the EFAS simulations in comparison to the two local systems PREVAH (left panel) and GR6J (right panel), in the three selected dry years of 2003, 2011, and 2018 combined. Low flow threshold: 15th percentile.

and 2018. This analysis explains why the differences between local and EFAS models are larger in the catchments in France than in Switzerland. The higher variability indicates EFAS's "flashy" behavior, particularly in pluvial catchments during dry months, where the model simulates sudden fluctuations in streamflow not observed in reality, which could be attributed to precipitation input or representation of the soil water storage capacity in the model.

5. Discussion

Building upon our findings in Section 4, we further analyze the performance differences between the continental system (EFAS) and local systems, focusing on catchment size, event duration, calibration status, and hydrological regime type.

5.1. Large vs. small catchments

In general, one expects model performance to improve as the catchment size increases, partly because smaller catchments are less likely to have a rain gauge within the catchment, reducing the accuracy of precipitation forcing (Michelon et al., 2021). In the case of EFAS, the system has been shown to perform better in larger catchments for floods (Alfieri et al., 2014). However, from our analysis, there is no clear link between catchment size and system performance in terms of its overall ability to simulate flows as measured through the log KGE' criteria (Fig. 2), and more specifically for low flow conditions in Fig. 3. There is a slight indication that the performance gap between EFAS and the local systems decreases with increasing catchment size, suggesting the need for a more locally set up system for smaller catchments while a continental system like EFAS can provide reasonably good simulations for larger catchments. Nevertheless, the model performance gap does not change at a consistent level for the catchments investigated, suggesting that catchment size is likely not a dominant factor influencing the model performance gap.

5.2. Long vs. short low flow events

Overall, EFAS tends to correctly capture short low flow events at the expense of displaying more false alarms compared to local systems. It is, however, less capable of simulating long-lasting events, compared to the local systems, due to its flashy behavior as shown in Section 4.4. Although we only present results for two catchments in detail, with the additional dry year analysis in Section 4.5, the flayshy behavior of these two catchments is representative of the selection stations, especially in pluvial catchments. This flashy behavior could be attributed to differences in the implementation or the calibration of the degree-day snow melt model used by EFAS, which only considers air temperature without accounting for the energy balance. While the two local systems also implement the degree-day approach (Viviroli et al., 2009; Valery, 2010), EFAS might misclassify precipitation as snow during winter (when it is actually rain) and melt it more slowly (due to the lack of energy storage), thus extending the melting period into late summer. However, this is not the case in the pluvial catchments in France, and yet EFAS exhibits similar flashy behavior where the model responses sensitively to precipitation and exaggerates streamflow peaks. Other potential reasons for this behavior include the forcing data, which may produce small peaks not present in locally observed precipitation data, or the routing equation in the model, which dominates the shape of the hydrograph recession curve at the end of the model structure.

From a water management perspective, a 2-month low flow event can trigger a very different operation strategy than a 2-week low flow event, making it important to improve EFAS' flashy low flow behavior in the future. Having said that, we should also note that, if for this hypothetical 2-month event, EFAS predicts two one-month events with a small gap of missed days in between with

its seasonal outlook, the system can still provide valuable information for decision-makers, especially when local systems are not set up to provide extended range probabilistic forecasts of low flows, as EFAS could provide.

The ability to simulate the severity or the magnitude of low flows in EFAS varies with event duration but without a clear pattern. Although EFAS does not simulate event magnitudes as accurately as the local systems in most stations considered in this study, future studies could explore applying a correction factor derived from duration-dependent magnitude analysis to EFAS outputs. This adjustment could enhance the confidence level of decision-makers relying on EFAS.

5.3. Calibrated vs. non-calibrated stations

While all stations in the local systems are calibrated, 18 of the 34 stations are not calibrated in EFAS. The lack of calibration at these stations on the one hand contributes to the performance gap between EFAS and the local systems. On the other hand, the overall acceptable performance of EFAS, with performance values within the range of the values obtained with local systems, is a strong indication that EFAS can provide useful information in ungauged catchments or regions with poor or insufficient streamflow data for low flow predictions, especially since they are indistinguishable in Figs. 2 and 3. Stations displaying the lowest CSI values within EFAS also show similarly low CSI values in local systems. Notably, the majority of these stations with the lowest CSI values within EFAS are non-calibrated. This mirrored low performance suggests that the predictive limitations might be due to the complexities of the catchments' response during low flow periods, rather than being inherent to the scale of the model, whether it is a continental or a local modeling system.

5.4. Nival vs. pluvial regimes

All selected stations in Switzerland follow a nival regime, while those in France follow a pluvial regime. Two representative case studies of each regime are analyzed in detail in this paper. The results show that EFAS is able to capture the climatological patterns of both regimes. However, the discrepancies between EFAS simulations and observations vary depending on the regime. In general, in the nival catchment, the EFAS system is drier than observed in winter and spring, but wetter in summer and autumn (e.g. Fig. 5a). In the pluvial catchment, EFAS is drier than observed throughout the year, but has the smallest discrepancy in the summer months (e.g. Fig. 5b). The overestimation of streamflow in EFAS for a nival regime could be attributed to the degree-day snow melt model as described in Section 5.2. For the discrepancy in streamflow simulation of EFAS in pluvial regimes, one possible cause might be the forcings used in the system, which may not accurately reflect observed precipitation and evapotranspiration conditions. The EFAS Meteorological Observation maps, available on the CEMS-flood confluence website, show that the rain gauge network in France (pluvial catchments) is sparse compared to Switzerland (nival catchments) (CEMS, 2023).

6. Conclusions

This study provides a comprehensive comparison of EFAS, as an example a continental hydrological system, with two locally calibrated systems, GR6J in France and PREVAH in Switzerland, to assess their ability to simulate low flows in the European Alps. Analyzing 34 catchments over a 20-year period, we focus on overall performance and the simulation of low flow periods using threshold-based events defined via time-varying climatological percentiles. Our findings show that the local systems generally outperform EFAS across almost all stations, with the performance gap becoming more pronounced during more intense low flow events. Specifically, EFAS tends to simulate lower streamflows than observations in late winter and early spring but overestimates and exhibits a flashier behavior during summer, leading to shorter simulated low flow events compared to observed long-lasting conditions, especially for pluvial catchments.

Despite being outperformed, the performance of EFAS is not out of range relative to the local systems, with about a 12.5% reduction in performance based on the log KGE' metric. Notably, stations with poor performance in EFAS also exhibit lower performance in the local systems, and the performance difference between EFAS and local systems at non-calibrated EFAS locations is comparable to that at calibrated locations. This suggests that EFAS has the potential to provide low flow simulations in ungauged or non-calibrated catchments, highlighting its utility as a continental-scale system for low flow prediction. Our work contributes to a better understanding of the low flow simulation capabilities of continental-scale systems and presents an evaluation framework that can be applied to other regions where performance analysis is needed before operational implementation.

While our study offers valuable insights and a practical evaluation framework, it is important to acknowledge several limitations that need to be considered.

- **Low flow threshold**: Low flow events are identified using time-varying thresholds based on climatology from 1999–2018. Different time periods may produce different thresholds. While we expect the general conclusions to hold, future studies could carry out sensitivity tests using other time periods or thresholds.
- **Verification metrics**: The log KGE' and CSI metrics used have known limitations. Log KGE' has issues with zero flows, flow units, and mean flow dependency, as discussed by Santos et al. (2018). CSI, being dichotomous, does not penalize discrepancies in deficit magnitude but focuses on timing. Future work could explore alternative transformations or metrics.
- **Data size**: The dataset here is relatively small, with two local systems and 34 stations within them. Expanding the analysis to other alpine regions would provide a broader perspective. However, GR6J and PREVAH are state-of-the-art, locally calibrated systems that in operation for low flow/drought monitoring, and their use as benchmarks strengthens our findings. Despite the small number of stations, the dataset includes a range of catchment sizes, elevations, and hydrological regimes, ensuring robust results.

Although local systems demonstrate higher accuracy in simulating low flows — particularly during prolonged and severe low flow events — continental systems like EFAS offer unique advantages that can significantly contribute to hydrological modeling and water resource management. EFAS provides continuous pan-European coverage, including transboundary catchments, and delivers seasonal streamflow predictions using state-of-the-art probabilistic forecasts. These features offer opportunities to extend EFAS's application from flood to drought prediction, enhancing early warning systems and contributing to initiatives like the WMO's Early Warnings for All.

Moreover, EFAS can supply gridded pan-European historical simulation datasets needed for training ML models, which benefit from extensive data coverage. Integrating EFAS outputs with local systems or ML techniques presents an interesting pathway for future research and operational developments. Such integration could enhance operational decision-making and low flow prediction capabilities, especially in regions lacking well-established local models. By combining the strengths of both continental-scale and local systems, the hydrological community can work towards more accurate and reliable predictions.

CRediT authorship contribution statement

Annie Y.-Y. Chang: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Maria-Helena Ramos:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Shaun Harrigan:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Christel Prudhomme:** Writing – review & editing, Conceptualization. **François Tilmant:** Writing – review & editing, Data curation. **Daniela I.V. Domeisen:** Writing – review & editing, Supervision. **Massimiliano Zappa:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors AC, MZ were supported by the MaleFix project, which is part of the WSL Program Extremes. We thank Konrad Bogner from WSL, the PI of the MaleFix project for his input. AC contribution is also part of the Interreg Alpine Space Programme project ADO (Alpine Space Observatory; grant no. ASP940), which in Switzerland has been financed via agreements with the Federal Office for Spatial Development ARE, Switzerland and the Cantons of Ticino and Thurgau. The authors MHR and FT were supported by the CIPRHES project, funded by the Agence Nationale de la Recherche (ANR), Switzerland in France. MHR contribution was part of her WSL-supported fellowship. Support from the Swiss National Science Foundation, Switzerland through project PP00P2_198896 to D.D. is gratefully acknowledged. We also thank Cinzia Mazzetti and Maliko Tanguy from the Hydrological Monitoring and Forecast Team at ECMWF for sharing their valuable knowledge on EFAS.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejrh.2024.102056.

Data availability

Data will be made available on request.

References

Addor, N., Melsen, L.A., 2019. Legacy, rather than adequacy, drives the selection of hydrological models. Water Resour. Res. 55, 378–390. http://dx.doi.org/10.1029/2018WR022958.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., Pappenberger, F., 2013. Glofas-global ensemble streamflow forecasting and flood early warning. Hydrol. Earth Syst. Sci. 17 (3), 1161–1175. http://dx.doi.org/10.5194/hess-17-1161-2013, URL https://hess.copernicus.org/articles/17/1161/2013/.

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., Salamon, P., 2014. Evaluation of ensemble streamflow predictions in Europe. J. Hydrol. 517, 913–922. http://dx.doi.org/10.1016/j.jhydrol.2014.06.035, URL https://www.sciencedirect.com/science/article/pii/S0022169414004958.

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J.C.M., Hasan, A., Pineda, L., 2019. Global catchment modelling using World-Wide HYPE (WWH), open data and stepwise parameter estimation. Hydrol. Earth Syst. Sci. http://dx.doi.org/10.5194/hess-2019-111.

Arnal, L., Cloke, H.L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., Pappenberger, F., 2018. Skilful seasonal forecasts of streamflow over Europe? Hydrol. Earth Syst. Sci. 22 (4), 2057–2072. http://dx.doi.org/10.5194/hess-22-2057-2018, URL https://hess.copernicus.org/articles/22/2057/2018/.

Bard, A., Renard, B., Lang, M., Giuntoli, I., Korck, J., Koboltschnig, G., Janža, M., d'Amico, M., Volken, D., 2015. Trends in the hydrologic regime of Alpine rivers. J. Hydrol. 529, 1823–1837. http://dx.doi.org/10.1016/j.jhydrol.2015.07.052, URL https://www.sciencedirect.com/science/article/pii/S0022169415005582.

Beven, K.J., 2012. Rainfall-Runoff Modelling : The Primer. Wiley-Blackwell, p. 457.

Bierkens, M.F.P., Bell, V.A., Burek, P., Chaney, N., Condon, L.E., David, C.H., de Roo, A., Döll, P., Drost, N., Famiglietti, J.S., Flörke, M., Gochis, D.J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R.M., Reager, J.T., Samaniego, L., Sudicky, E., Sutanudjaja, E.H., van de Giesen, N., Winsemius, H., Wood, E.F., 2015. Hyper-resolution global hydrological modelling: what is next? Hydrol. Process. 29 (2), 310–320. http://dx.doi.org/10.1002/hyp.10391, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.10391 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.10391.

Brunner, M.I., Götte, J., Schlemper, C., Van Loon, A.F., 2023. Hydrological drought generation processes and severity are changing in the alps. Geophys. Res. Lett. 50 (2), e2022GL101776. http://dx.doi.org/10.1029/2022GL101776.

Brunner, M.I., Van Loon, A.F., Stahl, K., 2022. Moderate and severe hydrological droughts in Europe differ in their hydrometeorological drivers. Water Resour. Res. 58 (10), e2022WR032871. http://dx.doi.org/10.1029/2022WR032871.

Brunner, M.I., Zappa, M., Stähli, M., 2019. Scale matters: Effects of temporal and spatial data resolution on water scarcity assessments. Adv. Water Resour. 123, 134–144. http://dx.doi.org/10.1016/j.advwatres.2018.11.013.

Caillouet, L., Vidal, J.-P., Sauquet, E., Devers, A., Graff, B., 2017. Ensemble reconstruction of spatio-temporal extreme low-flow events in France since 1871. Hydrol. Earth Syst. Sci. 21 (6), 2923–2951. http://dx.doi.org/10.5194/hess-21-2923-2017, URL https://hess.copernicus.org/articles/21/2923/2017/.

CEMS, 2023. EFAS versioning system - EFAS v4.0. https://confluence.ecmwf.int/display/CEMS/EFAS+v4.0. (Accessed 05 March 2024).

Crochemore, L., Ramos, M.-H., Pechlivanidis, I., 2020. Can continental models convey useful seasonal hydrologic information at the catchment scale? Water Resour. Res. 56, http://dx.doi.org/10.1029/2019WR025700.

De Roo, A.P.J., Wesseling, C.G., Van Deursen, W.P.A., 2000. Physically based river basin modelling within a GIS: the LISFLOOD model. Hydrol. Process. 14 (11–12), 1981–1992. http://dx.doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F.

Dufeu, E., Mougin, F., Foray, A., Baillon, M., Lamblin, R., Hebrard, F., Chaleon, C., Romon, S., Cobos, L., Gouin, P., Audouy, J.-N., Martin, R., Poligot-Pitsch, S., 2022. Finalisation de l'opération HYDRO 3 de modernisation du système d'information national des données hydrométriques. LHB 108 (1), 2099317. http://dx.doi.org/10.1080/27678490.2022.2099317.

Emerton, R., Stephens, E., Pappenberger, F., Pagano, T., Weerts, A., Wood, A., Salamon, P., Brown, J., Hjerdt, N., Donnelly, C., Baugh, C., Cloke, H., 2016. Continental and global scale flood forecasting systems. Wiley Interdiscipl. Rev.: Water 3, http://dx.doi.org/10.1002/wat2.1137.

Global Runoff Data Centre (GRDC), 2023. Global runoff database. URL https://www.bafg.de/GRDC/EN/Home/homepage_node.html.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1), 80–91. http://dx.doi.org/10.1016/j.jhydrol.2009.08.003, URL https://www.sciencedirect.com/science/article/pii/S0022169409004843.

Gurtz, J., Baltensweiler, A., Lang, H., 1999. Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountain basins. Hydrol. Process. - Hydrol. Process. 13, 2751–2768. http://dx.doi.org/10.1002/(SICI)1099-1085(19991215)13:173.3.CO;2-F.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., Pappenberger, F., 2020. Glofas-ERA5 operational global river discharge reanalysis 1979–present. Earth Syst. Sci. Data 12 (3), 2043–2060. http://dx.doi.org/10.5194/essd-12-2043-2020, URL https://essd.copernicus.org/articles/12/2043/2020/.

Hisdal, H., Tallaksen, L.M., Gauster, T., Bloomfield, J.P., Parry, S., Prudhomme, C., Wanders, N., 2024. Chapter 5 - hydrological drought characteristics. In: Tallaksen, L.M., van Lanen, H.A. (Eds.), Hydrological Drought, second ed. Elsevier, pp. 157–231. http://dx.doi.org/10.1016/B978-0-12-819082-1.00006-0, URL https://www.sciencedirect.com/science/article/pii/B9780128190821000060.

Horton, P., Schaefli, B., Kauzlaric, M., 2022. Why do we have so many different hydrological models? A review based on the case of Switzerland. WIREs Water 9 (1), e1574. http://dx.doi.org/10.1002/wat2.1574, arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wat2.1574 URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wat2.1574.

Johnson, J.M., Fang, S., Sankarasubramanian, A., Rad, A.M., Kindl da Cunha, L., Jennings, K.S., Clarke, K.C., Mazrooei, A., Yeghiazarian, L., 2023. Comprehensive analysis of the NOAA national water model: A call for heterogeneous formulations and diagnostic model selection. J. Geophys. Res.: Atmos. 128 (24), e2023JD038534. http://dx.doi.org/10.1029/2023JD038534, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023JD038534.

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424–425, 264–277. http://dx.doi.org/10.1016/j.jhydrol.2012.01.011, URL https://www.sciencedirect.com/science/article/pii/S0022169412000431.

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and kling–gupta efficiency scores. Hydrol. Earth Syst. Sci. 23 (10), 4323–4331. http://dx.doi.org/10.5194/hess-23-4323-2019, URL https://hess.copernicus.org/articles/23/4323/2019/.

Kolling Neto, A., Siqueira, V.A., Gama, C.H.d., Paiva, R.C.D.d., Fan, F.M., Collischonn, W., Silveira, R., Paranhos, C.S.A., Freitas, C., 2023. Advancing medium-range streamflow forecasting for large hydropower reservoirs in Brazil by means of continental-scale hydrological modeling. Water 15 (9), http://dx.doi.org/10.3390/w15091693, URL https://www.mdpi.com/2073-4441/15/9/1693.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall – runoff modelling using long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022. http://dx.doi.org/10.5194/hess-22-6005-2018, URL https://www.hydrol-earth-syst-sci.net/22/6005/2018/.

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97. http://dx.doi.org/10.5194/adgeo-5-89-2005, URL https://adgeo.copernicus.org/articles/5/89/2005/.

Meile, T., Boillat, J.L., Schleiss, A.J., 2011. Hydropeaking indicators for characterization of the Upper-Rhone River in Switzerland. Aquat. Sci. 73, 171–182. http://dx.doi.org/10.1007/s00027-010-0154-7.

Michelon, A., Benoit, L., Beria, H., Ceperley, N., Schaefli, B., 2021. Benefits from high-density rain gauge observations for hydrological response analysis in a small alpine catchment. Hydrol. Earth Syst. Sci. 25 (4), 2301–2325. http://dx.doi.org/10.5194/hess-25-2301-2021, URL https://hess.copernicus.org/articles/25/2301/2021/.

Ng, K., Huang, Y., Koo, C., Chong, K., El-Shafie, A., Najah Ahmed, A., 2023. A review of hybrid deep learning applications for streamflow forecasting. J. Hydrol. 625, 130141. http://dx.doi.org/10.1016/j.jhydrol.2023.130141, URL https://www.sciencedirect.com/science/article/pii/S0022169423010831.

Nicolle, P., Besson, F., Delaigue, O., Etchevers, P., François, D., Lay, M.L., Perrin, C., Rousset, F., Thiéry, D., Tilmant, F., Magand, C., Leurent, T., Jacob, É., 2020. PREMHYCE: An operational tool for low-flow forecasting. Proceedings of the International Association of Hydrological Sciences 383, 381–389. http://dx.doi.org/10.5194/piahs-383-381-2020.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Lay, M.L., Besson, F., Soubeyroux, J.M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., Morice, E., 2014. Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. Hydrol. Earth Syst. Sci. 18, 2829–2857. http://dx.doi.org/10.5194/hess-18-2829-2014.

Pechlivanidis, I.G., Jackson, B., McMillan, H., Gupta, H., 2014. Use of an entropy-based metric in multiobjective calibration to improve model performance. Water Resour. Res. 50 (10), 8066–8083. http://dx.doi.org/10.1002/2013WR014537.

Pushpalatha, R., Perrin, C., Moine, N.L., Mathevet, T., Andréassian, V., 2011. A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. J. Hydrol. 411, 66–76. http://dx.doi.org/10.1016/j.jhydrol.2011.09.034.

Quintana-Seguí, P., Moigne, P.L., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., Morel, S., 2008. Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. J. Appl. Meteorol. Climatol. 47 (1), 92–107. http://dx.doi.org/10.1175/2007JAMC1636.1, URL https://journals.ametsoc.org/view/journals/apme/47/1/2007jamc1636.1.xml.

Santos, L., Thirel, G., Perrin, C., 2018. Technical note: Pitfalls in using log-transformed flows within the KGE criterion. Hydrol. Earth Syst. Sci. 22 (8), 4583–4591. http://dx.doi.org/10.5194/hess-22-4583-2018, URL https://hess.copernicus.org/articles/22/4583/2018/.

Schaefer, J.T., 1990. The critical success index as an indicator of warning skill. Weather Forecast. 5 (4), 570–575. http://dx.doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21 (15), 2075–2080. http://dx.doi.org/10.1002/hyp.6825, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.6825 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.6825.

Slater, L.J., Arnal, L., Boucher, M.A., Chang, A.Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R.L., Wood, A., Zappa, M., 2023. Hybrid forecasting: blending climate predictions with AI models. http://dx.doi.org/10.5194/hess-27-1865-2023.

Sood, A., Smakhtin, V., 2015. Global hydrological models: a review. Hydrol. Sci. J. 60 (4), 549–565. http://dx.doi.org/10.1080/02626667.2014.950580, URL https://doi.org/10.1080/02626667.2014.950580.

Speich, M.J., Bernhard, L., Teuling, A.J., Zappa, M., 2015. Application of bivariate mapping for hydrological classification and analysis of temporal change and scale effects in Switzerland. J. Hydrol. 523, 804–821. http://dx.doi.org/10.1016/j.jhydrol.2015.01.086.

Thielen, J., Bartholmes, J., Ramos, M.-H., de Roo, A., 2009. The European flood alert system – Part 1: Concept and development. Hydrol. Earth Syst. Sci. 13 (2), 125–140. http://dx.doi.org/10.5194/hess-13-125-2009, URL https://hess.copernicus.org/articles/13/125/2009/.

Thiemig, V., Gomes, G.N., Skøien, J.O., Ziese, M., Rauthe-Schöch, A., Rustemeier, E., Rehfeldt, K., Walawender, J.P., Kolbe, C., Pichon, D., Schweim, C., Salamon, P., 2022. EMO-5: a high-resolution multi-variable gridded meteorological dataset for Europe. Earth System Science Data 14 (7), 3249–3272. http://dx.doi.org/10.5194/essd-14-3249-2022, URL https://essd.copernicus.org/articles/14/3249/2022/.

Tilmant, F., Bourgin, F., François, D., Le Lay, M., Perrin, C., Rousset, F., Vergnes, J.-P., Willemet, J.-M., Magand, C., Morel, M., 2023. PREMHYCE: an operational tool for low-flow forecasting. Sci. Eaux Territ. (42), 17–21. http://dx.doi.org/10.20870/Revue-SET.2023.42.7297, URL https://revue-set.fr/article/view/7297.

University of Southampton, 2024. African flood and drought monitor (AFDM). URL https://hydrology.soton.ac.uk/apps/afdm/. (Accessed 16 June 2024).

Valery, A., 2010. Mod'elisation Pr'ecipitations–D'ebit Sous Influence Nivale: 'Elaboration D'un Module Neige et 'Evaluation Sur 380 Bassins Versants (Ph.D. thesis). AgroParisTech.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.-M., 2010. A 50-year high-resolution atmospheric reanalysis over France with the safran system. Int. J. Climatol. 30 (11), 1627–1644. http://dx.doi.org/10.1002/joc.2003, arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.2003 URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.2003.

Viviroli, D., Zappa, M., Gurtz, J., Weingartner, R., 2009. An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. Environ. Model. Softw. 24, 1209–1222. http://dx.doi.org/10.1016/j.envsoft.2009.04.001.

Wei, Y., Wang, R., Feng, P., 2024. Improving hydrological modeling with hybrid models: A comparative study of different mechanisms for coupling deep learning models with process-based models. Water Resour. Manage. http://dx.doi.org/10.1007/s11269-024-03780-5.

Wetterhall, F., Di Giuseppe, F., 2018. The benefit of seamless forecasts for hydrological predictions over Europe. Hydrol. Earth Syst. Sci. 22 (6), 3409–3420. http://dx.doi.org/10.5194/hess-22-3409-2018, URL https://hess.copernicus.org/articles/22/3409/2018/.

World Meteorological Organization, 2023. Early warnings for all: Executive action plan 2023–2027. URL https://library.wmo.int/records/item/58209-early-warnings-for-all. (Accessed 05 March 2024).

World Meteorological Organization, 2024. WMO and the early warnings for all initiative. URL https://wmo.int/activities/early-warnings-all/wmo-and-early-warnings-all-initiative. (Accessed 16 June 2024).

Zappa, M., Bernhard, L., Spirig, C., Pfaundler, M., Stahl, K., Kruse, S., Seidl, I., Stähli, M., 2014. A prototype platform for water resources monitoring and early recognition of critical droughts in Switzerland. IAHS-AISH Proc. Rep. 364 (March 2012), 492–498.

Žun, M., Bernhard, L., Crouzat, E., Haslinger, K., Hribernik, M., Klemenčič, S., Kristan, B., Sušnik, A., Vlahović, Ž., Zappa, M., 2022. Alpine drought observatory. URL https://www.alpine-space.eu/wp-content/uploads/2022/11/ADO_d.t.3.1.1-report-on-drought-impacts-in-the-alps.pdf.