https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

Exploring Hybrid Forecasting Frameworks for Subseasonal Low Flow Predictions in the European Alps

Annie Y.-Y. Chang[1, 2, 3,6], Shaun Harrigan[3], Maria-Helena Ramos[4], Massimiliano Zappa[1], Christian M. Grams[5,6], Daniela I.V. Domeisen[7, 2], and Konrad Bogner[1]

[1]Swiss Federal Institute WSL,Birmensdorf, Switzerland
[2]Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland
[3]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, United Kingdom
[4]Université Paris-Saclay, INRAE, UR HYCAR, Antony, France
[5]Institute of Meteorology and Climate Research Troposphere Research (IMKTRO), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[6]now at: Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich Airport, Switzerland
[7]University of Lausanne, Lausanne, Switzerland

Correspondence: Annie Y.-Y. Chang (ychang@ethz.ch)

Abstract.

Since the start of the 21st century, the European Alpine region has faced unprecedented low-flow conditions and drought events, severely impacting sectors dependent on reliable water availability, such as hydropower production, agriculture, and transportation. The growing frequency and severity of these low-flow conditions have led to a need for early warning systems. In this study, we present a novel machine learning (ML) aided hybrid forecasting framework designed to enhance sub-seasonal low-flow predictions in the European Alps. By harnessing the statistical power of ML and integrating diverse data sources, we trained 11 models using the Temporal Fusion Transformer (TFT) algorithm. These models incorporate features such as European Atlantic Weather Regimes (WR) for capturing large-scale atmospheric circulation patterns, in-situ streamflow observations for initial conditions, and process-based predictions from the European Flood Awareness System (EFAS). Our results show that the hybrid framework, even when using only WR data, outperforms climatology. The best results are achieved by combining observational data with process-based model data (raw EFAS output), underscoring the value of integrating diverse data sources. The models effectively capture initial condition persistence and correct biases in the raw EFAS output. Based on the Continuous Ranked Probability Skill Score (CRPSS), the best model effectively extends the skilful forecast horizon by 5 days on average across all stations during low flow periods. Furthermore, the interpretability of the TFT model provides valuable insights, identifying glacier coverage as a key catchment feature influencing model performance. Future research should further explore the connections between hydrological features and prediction skill, as well as the framework's applicability in ungauged areas and other regions.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

1 Introduction

Streamflow prediction in the European Alps presents a significant challenge due to the complex joint effects of climatic, hydrological, and geographical factors (Horton et al., 2022). The region's diverse topography, combined with distinct seasonal weather patterns, complicates accurate forecasting, which is crucial for water resource management, flood prevention, and hydropower generation. Additionally, hydrological patterns in the Alps have changed noticeably with climate change, with a marked increase in the intensity and frequency of low flow events since the late 20th century (Bard et al., 2015; Brunner et al., 2023). This trend places added pressure on ecosystems and poses significant challenges for both cross-border and local water management. The inventory study of Alpine drought impact reports highlights the need for increased preparedness to address the growing frequency, severity, and complexity of drought impacts in the region (Stephan et al., 2021). The study calls for a shift from reactive emergency responses to proactive prevention and preparedness actions. Historically, streamflow predictions in the Alps have relied heavily on hydrological models driven by meteorological forecasts. Yet, the region's high spatial and temporal variability in precipitation and snowmelt processes introduces substantial difficulties for accurate streamflow forecasting (Gurtz et al., 2003; Viviroli et al., 2009; Horton et al., 2022).

To address the uncertainties in hydrological models, various state-of-the-art post-processing techniques have been developed. These methods aim to bias correct model outputs to improve their accuracy and reliability. Techniques such as quantile mapping (QM), model output statistics (MOS), and ensemble model output statistics (EMOS) are widely used in the field of hydrometeorology (Gneiting et al., 2005; Wilks, 2011; Monhart et al., 2018). Quantile mapping, for instance, aligns the distribution of model forecasts with observed data, thereby reducing systematic errors that arise, for example, when using coarse-resolution forecasts in complex topography (Monhart et al., 2018). Alternatively, MOS and EMOS techniques use statistical models, typically regressions, to relate model outputs to observed data, fitting a parametric distribution to improve forecast accuracy. Furthermore, Bogner and Pappenberger (2011) demonstrated that error correction methods, particularly those using wavelet transformations, can significantly improve the accuracy of river discharge predictions by effectively accounting for forecast errors across multiple temporal scales. Nevertheless, many traditional post-processing methods often assume linear relationships between predictors and outcomes, which may limit their ability to fully capture the complex, nonlinear dynamics influencing streamflow.

Recent advances in machine learning (ML) have led to the development of new statistics-based streamflow prediction models. Google's end-to-end flood warning system is a good example of deploying artificial intelligence (AI) in operational hydrological forecasting (Nevo et al., 2022; Nearing et al., 2024). While traditional hydrological models and data-driven techniques have independently advanced streamflow prediction, they each have their own disadvantages and often fall short of capturing the full complexity of hydrological systems. This gap has prompted the development of hybrid forecasting frameworks, with the aim to combine the strengths of both approaches. By combining physical models with ML algorithms, these systems leverage the understanding of hydrological processes from physical models and the adaptive, statistical capabilities of ML, providing the opportunity to enhance prediction accuracy (Papacharalampous and Tyralis, 2022; Hauswirth et al., 2022; Slater et al., 2023; Ng et al., 2023; Wei et al., 2024).

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

In their review study, Slater et al. (2023) investigated over 20 published hybrid hydroclimatic forecasting systems, and categorized them by aspects such as predictands, type of data-driven and dynamical models, model combination methods, and forecasting horizons. This comprehensive analysis revealed a wide range of setups and applications for hybrid approaches. Several strengths of these systems were identified, including reduced computational costs and minimized biases, which are achieved by leveraging large datasets from diverse sources to enhance forecast accuracy. The review also highlighted key challenges and opportunities, such as the need for physically realistic results, the need for interpretability to boost forecast uptake by operational centres, and the development of seamless prediction frameworks across different time scales.

A barrier to the broader application of ML-based forecasting systems, especially at regional and continental scale, is the heterogeneous availability of hydrological data across regions. While some catchments have dense observational records and calibrated models, others are sparsely gauged or lack consistent datasets. This variability limits the transferability of hybrid approaches and remains a key obstacle to operational implementation at scale.

In response to this challenge, the present study introduces a hybrid forecasting framework designed to support flexible model configurations while enhancing sub-seasonal streamflow prediction, with a particular focus on low-flow conditions in the European Alps. Using interpretable AI, the study also seeks to identify the dominant drivers of forecast skill in Alpine catchments, ultimately offering water managers a practical framework for decision support.

This study builds on our previous work (Chang et al., 2024), which defined the catchment network of interest and established a baseline evaluation of low-flow prediction in the European Alps. Here, we extend that work by introducing a hybrid forecasting framework that integrates multiple data sources.

The framework combines three complementary data sources: (1) North Atlantic–European Weather Regimes (WR) data, which capture large-scale atmospheric circulation patterns; (2) in-situ streamflow observations, used for model training, initial conditions, and persistence information; and (3) process-based streamflow simulations from the European Flood Awareness System (EFAS), representing conventional physics-based hydrological modelling.

The following sections provide an overview of the study area (Sec. 2). In Sec. 3, we describe the three data sets used in this study, followed by the methodology in Sec. 4. In Secs. 5 and 6, we present the main results and discuss the interpretability of the models as well as the limitations of the study. We close with some concluding statements in Sec. 7.

2 Study area

Our study area, the European Alpine space, is defined by the red boundary in Figure 1, spanning across eight Alpine countries: France (FR), Monaco (MC) Switzerland (CH), Liechtenstein (LI) Italy (IT), Germany (DE), Austria (AT), and Slovenia (SI). Within the Alpine space, we select 101 stations where we have both observational streamflow data and process-based streamflow model data available. Station selection is mainly based the availability of continuous observational streamflow data, requiring that data gaps not exceed seven consecutive days. As a result, the study includes 12 stations in FR, 22 in CH, 1 in IT, 29 in DE, 26 in AT, and 11 in SI. These stations are distributed across four Alpine river basins: the Danube, Rhine,

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

Rhône, and Po. This study area, and the associated station network, is the same as that analysed in Chang et al. (2024). The metadata of the selected stations is provided as a supplementary CSV file.

Table 1 summarizes the primary characteristics of the 101 selected stations. Their catchment areas range from 75 to 101,800 km^2. Due to this substantial variation in upstream areas, there is a large variability in mean annual streamflow among the stations, ranging from 118 mm to 1843 mm, with a median of 774 mm per year. As the study area is situated in a mountainous region, the stations exhibit a wide range of elevations, with median heights spanning from 98 m to 1078 m above sea level. The Alpine space is home to many glaciers and the catchments have a glacier coverage from zero up to 13.7 %. There are many lakes and reservoirs in the Alpine space, which can take up as much as 5% of the catchment area. The diverse nature of these catchments results in complex hydrological behavior, presenting challenges for accurate hydrological modeling.

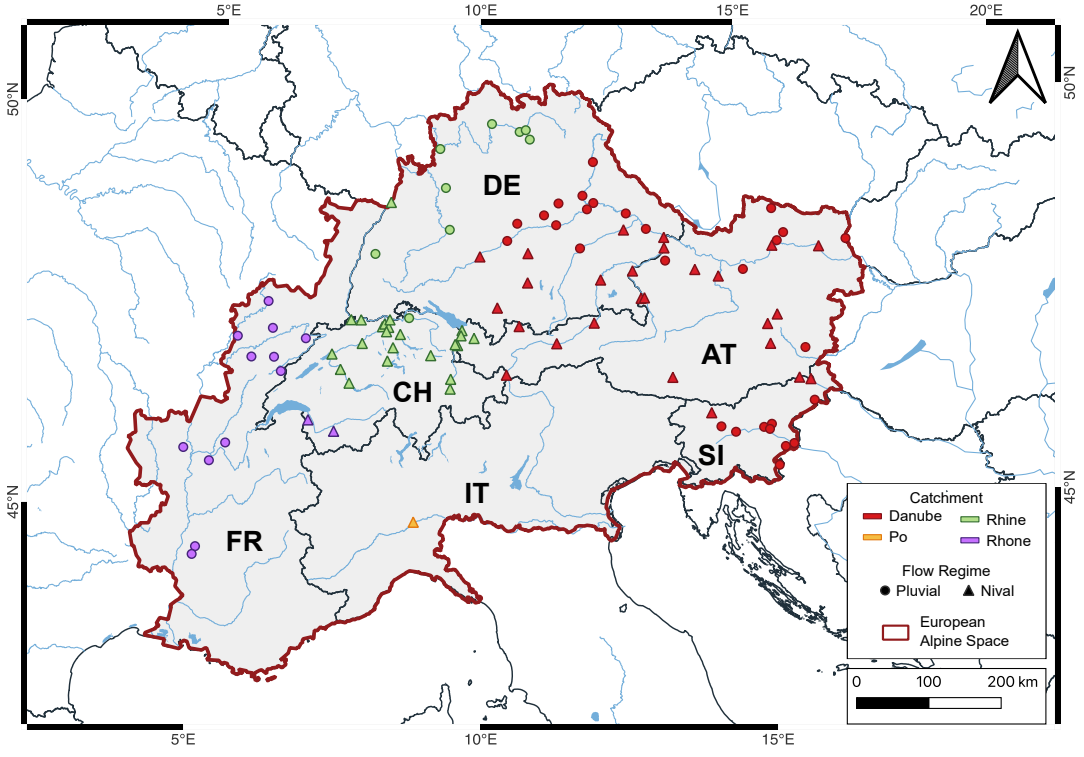

Figure 1. Map of the study area - the European Alpine space in France (FR), Monaco (MC) Switzerland (CH), Liechtenstein (LI) Italy (IT), Germany (DE), Austria (AT), and Slovenia (SI). 101 streamflow stations are selected for this study, including 54 on the River Danube, 32 on the River Rhine, 14 on the River Rhône, and 1 on the River Po. The round markers denote pluvial catchments, and the triangle markers denote nival catchments.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

	Min	25%	Median	75%	Max
Area (km^2)	75	1250	3175	9700	101800
Median elevation (m)	76	229	334	416	1076
Glacier coverage (%)	0	0	0	0.7	13.7
Lake coverage (%)	0	0	0.2	0.4	4.9
Mean annual streamflow Q_A (mm)	118	424	774	1061	1843
25^{th} percentile annual flow (mm)	92	364	680	961	1609

Table 1. Summary of the main characteristics of the 101 selected catchments.

2.1 Hydrological regimes

We identify two primary types of hydrological regimes in the study area: pluvial and nival regimes. A nival regime, where streamflow is primarily influenced by snow and ice melt, features a bell-shaped hydrograph. In nival catchments, streamflow builds from spring, reaches its maximum in early summer (typically June–July) (Figure 2a). By contrast, a pluvial regime is rainfall-driven. In pluvial catchments, streamflow is highest during the wetter autumn and winter months and lowest in spring and summer, responding to the annual rainfall cycle (Figure 2b). Two catchments illustrating these regimes from the same dataset were presented in Chang et al. (2024).

In Figure 2, each grey line represents the climatological streamflow averaged over 1999-2018 for a station. To construct the climatological distribution, we use a 31-day moving window centred on each day of the year, incorporating all (7-day smoothed) daily streamflow data from the 20-year period (1999–2018). This approach, which includes data from 15 days before and after each target day, helps smooth short-term variability and ensures a more stable estimate of climatological streamflow conditions for any given time of the year. A catchment is classified as nival if the mean climatological streamflow during the extended summer months (May - October) is higher than during the extended winter months (November - April). Vice versa, a catchment is classified as pluvial if the mean climatological streamflow is higher in the winter months as compared to summer. Based on this definition, there are 49 nival stations (mean elevation at 417 m) and 52 pluvial stations (mean elevation at 277 m). The dark black lines in Figures 2a and b represent the median streamflow of the nival and pluvial stations, respectively. Although some stations exhibit influences from both rainfall and snowmelt, leading to mixed hydrological regimes (e.g., nival-pluvial), we have classified each station as either nival or pluvial for this study. This decision is made because distinguishing these in-between regimes can be tricky and may introduce ambiguity in the analysis. Refer to Figure 1 for the spatial distribution of these two types of catchments.

3 Data

Our study uses three main types of data: meteorological weather regimes, streamflow observations, and process-based streamflow simulations. This section provides a description of each dataset.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

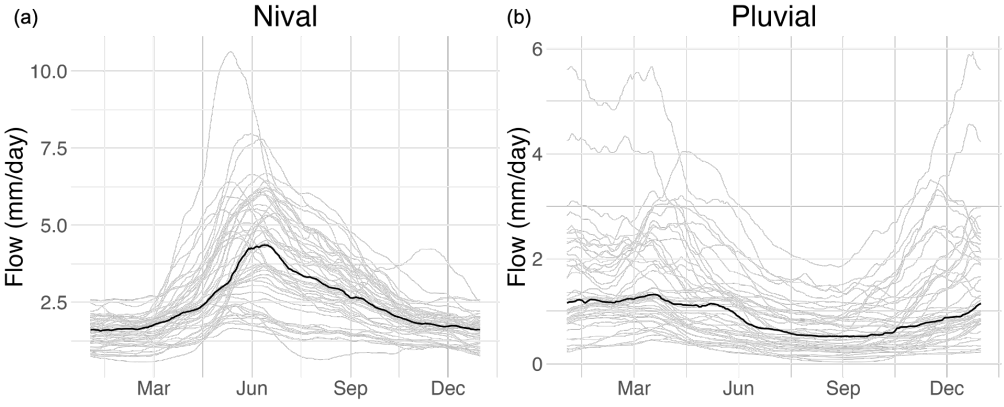

Figure 2. Panel (a) - streamflow climatology of all 49 stations that exhibit a nival flow regime. Panel (b) - streamflow climatology of all 52 stations that exhibit a pluvial flow regime. Medians are shown by the thick black lines. Climatology period: 1999-2018.

3.1 Observations

Observational streamflow data serve multiple purposes in this study. They provide target values for the ML models as training data and are independently used for verifying model performance as separate testing data. We also incorporate observational data as an input feature in the encoding period to provide initial condition information and recent past information to allow the model to harness the catchment memory (Sutanto and Lanen, 2022). Daily observational streamflow data were obtained from the same databases used in Chang et al. (2024), namely the Alpine Drought Observatory (Žun et al.), the Global Runoff Data Centre (GRDC) (Global Runoff Data Centre (GRDC), 2023), and the HydroPortail in France (Dufeu et al., 2022). Due to the considerable range in upstream catchment areas, daily streamflow data are converted from volumetric discharge (in m3/s), facilitating meaningful comparisons across stations and enhancing model training by standardizing input magnitudes and reducing scale-related biases.

3.2 Weather regimes

The European weather regimes (WRs) represent dominant large-scale atmospheric circulation patterns over the North Atlantic–European region, which influence weather variability across Europe. Chang et al. (2023) showed that European WRs have added value for post-processing sub-seasonal hydrological forecasts in Switzerland. In the present study, we extend the scope of the analysis to the European Alpine space and explore the potential of using the same set of European weather regime data to predict streamflow in a hybrid setup.

As in Chang et al. (2023), we employ the year-round classification of seven Atlantic-European weather regimes (WRs) introduced by Grams et al. (2017), in combination with the forecasting methodology described in Grams et al. (2020) and Büeler et al. (2021). The seven regimes are the Atlantic Trough regime (AT), Zonal regime (ZO), Scandinavian Trough (ScTr), Atlantic Ridge (AR), European Blocking (EUBL), Scandinavian Blocking (ScBL) and Greenland Blocking (GL) regimes.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

The large-scale circulation patterns associated with each regime are illustrated in Figure 1 of Büeler et al. (2021). The regime definition used in this study is based on ERA5 reanalysis data covering 1979–2019. For each ensemble member and every 6-hour forecast lead time (or the operational analysis), we calculate seven ”weather regime indices" (IWRs), which represent the normalized projection of the 10-day low-pass filtered normalized 500 hPa Geopotential Height (Z500) anomalies with respect to an ERA5 1979–2019 calendar day climatology onto the respective regime patterns. From the IWRs, we derive a WR life cycle (LC), enabling us to assign each time step to one of the regimes, while marking periods lacking regime characteristics as ”No Regime". The methodology for computing IWR and LC attribution follows the steps outlined by Büeler et al. (2021). In contrast to Chang et al. (2023), here we use a post-processed version of the regimes as described in Osman et al. (2023).

The typical temperature and precipitation anomalies (based on ERA5 data) in the European Alpine space associated with the seven WRs are demonstrated in Figures S1 and S2 in the Supplementary. To further explore the connection between WRs and local streamflow conditions, we analyse streamflow anomalies at the selected stations, grouped by their respective large river catchments (Danube, Rhine, and Rhône). The Po River catchment is excluded from the analysis as it contains only one station, which is insufficient to provide a representative assessment of streamflow patterns with respect to WRs for the catchment. Figure S3 in the Supplementary shows the streamflow anomalies standardised by the standard deviation of the climatology on days where one of the seven WRs or ”No Regime" is dominant. As we have seen a strong seasonality in the temperature and precipitation anomalies, in Figure S3 in the Supplementary we present the streamflow anomalies for every season separately. A positive anomaly indicates a higher-than-normal flow, while a negative standardised anomaly indicates a lower-than-normal flow. WRs with the strongest and most robust streamflow signal within each season are highlighted with a solid frame for the strongest positive anomaly or a dashed frame for the strongest negative anomaly. The ZO regime shows an overall low flow signal. Similarly, the AR regime is associated with substantial low-flow conditions in winter and spring, but shows high-flow signals in summer and a largely station-dependent signal in autumn. The AT regime, on the other hand, shows strong high-flow signals for autumn, winter, and spring. The pronounced and robust signals in streamflow as a function of flow regime further motivate the use of WR data in predictive systems.

Figure 3 illustrates the example of the ZO regime in spring in terms of the surface weather anomalies (panels a & b) alongside streamflow anomalies (panels c & d). The ZO regime is characterised by an anomalous low-pressure system over Greenland and an anomalous high pressure system in mid-latitudes extending over southern and central Europe. This configuration is associated with in a fairly zonal and strong jet stream. The high-pressure system covering our study area further leads to generally warm and dry conditions. On spring days dominated by the ZO regime, a majority (70) of the stations experience lower-than-normal flow conditions. However, the higher-than-normal temperatures accelerate snow and ice melt, leading nival stations to experience higher-than-normal flow conditions (Figure S3e in the Supplementary). The ZO regime is further associated with warmer and drier than normal conditions in summer and autumn (Figures S5 and S6 in the Supplementary), leading to lower-than-normal flow conditions at most stations during these seasons.

As a comparison example (Figures S4 and S7 in the Supplementary), the Atlantic Ridge (AR) regime is characterised by an anomalous high-pressure system over the Atlantic, hence the name ”Atlantic Ridge". This high-pressure system creates an anticyclonic circulation, advecting cold and dry air from the polar region. Consequently, northern and central Europe tend to

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

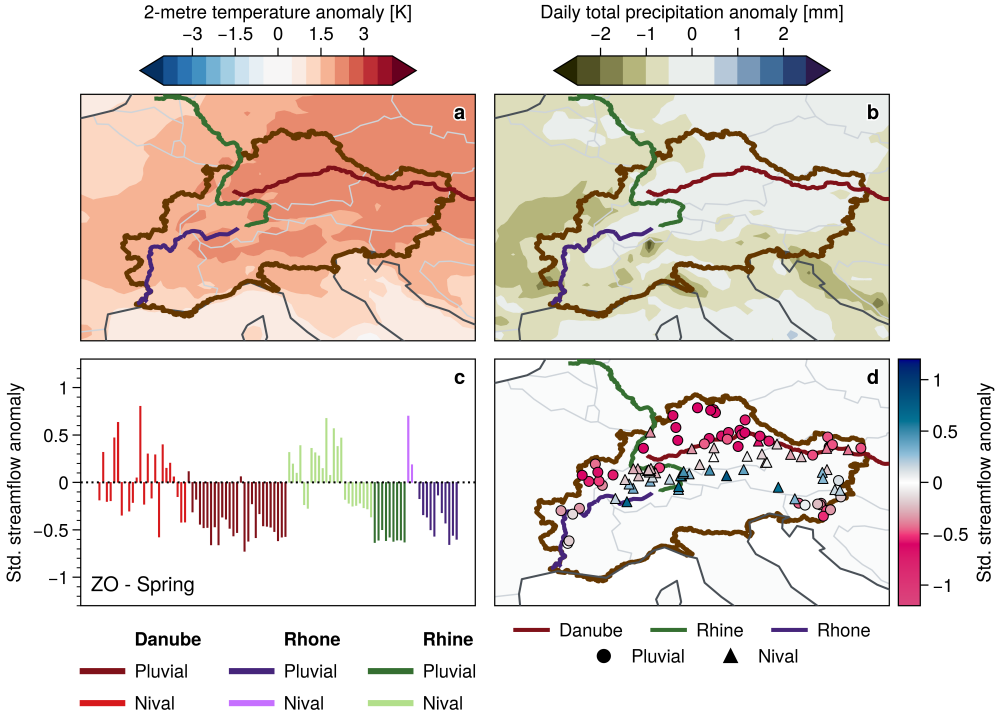

Figure 3. Composite of surface weather anomalies and streamflow anomalies during Zonal (ZO) regime days in spring. (a) surface temperature, (b) daily accumulated total precipitation, (c) standardized streamflow anomaly for different stations (repeated from Figure S3y in the Supplementary), and (d) standardized streamflow anomalies displayed on a map of the study area.

experience colder and drier conditions than normal. The reduced temperatures slow down the melting of snow and ice in spring, and the lack of precipitation further contributes to lower-than-normal flow conditions at the majority of stations (Figure S3m in the Supplementary). In winter, AR is the regime with the strongest lower-than-normal flow anomaly signal. These examples show how the analysis of WR behaviour can help to gain physical insights into how the inclusion of WR data sources can give predictive skill within a model.

Figure S8 in the Supplementary illustrates composite streamflow anomalies for days dominated by specific WRs, with varying lag days applied, highlighting the persistence and lasting impact of WRs on streamflow conditions.

3.3 Process-based information - EFAS

In this study, we explore process-based information using the streamflow data product from the European Flood Awareness System (EFAS, https://european-flood.emergency.copernicus.eu/en). This dataset is chosen for its comprehensive gridded coverage (5 km x 5 km for EFAS v4.0) over Europe, at multiple forecast horizons, and its operational application in real-time flood forecasting systems.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

Operational since 2012, EFAS is part of the Copernicus Emergency Management Service (CEMS, https://emergency.copernicus.eu/) and delivers hydrological forecasts and simulations twice daily at 6-hourly steps and kilometer-scale resolution for European river basins. EFAS delivers extensive flood risk assessments and issues flood notifications with a lead time of up to 15 days. Additionally, extended range (46-day) and seasonal (7-month) forecasts are available. These simulations are produced from the fully-distributed and physics-based hydrological model LISFLOOD (De Roo et al., 2000). Detailed model documentation can be accessed at https://ec-jrc.github.io/lisflood-model/Lisflood_Model.pdf. For simplicity,"EFAS" hereafter refers to both the LISFLOOD model and its streamflow data products.

The 101 stations selected for this study are all EFAS reporting stations. Calibration in EFAS is conducted at reporting stations, though not all stations are calibrated, depending on data quality and catchment area size (only catchments larger than 500 km^2 are calibrated). In total, the station network is composed of 73 calibrated EFAS station and 28 without calibrationss, same as in Chang et al. (2024).

We extract both EFAS reanalysis and EFAS reforecast datasets for this study. To ensure model consistency, both datasets used are based on EFAS v4.0, which has a spatial resolution of 5 km x 5 km. The performance of EFAS outputs is regularly evaluated and shows consistent improvements (CEMS, 2023; Bartholmes et al., 2009). The inter-model comparison study of Chang et al. (2024) analysed the strengths and weaknesses of EFAS in simulating low flows. The results show that EFAS output has potential for low flow applications but also highlights a "flashy" behaviour that limits its ability to capture prolonged low flow events. This study aims to evaluate whether the deployed ML models can effectively extract useful process-based information from EFAS while correcting for its biases and errors, particularly in the context of low flow periods.

Detailed findings and additional information on EFAS can be found in Chang et al. (2024). These EFAS datasets, available from 1999 to 2018, are transformed from 6-hourly data into a mean daily streamflow time series. This period defines the scope of our analysis.

4 Method

4.1 Temporal Fusion Transformer

The Transformer, a deep learning network model introduced by Vaswani et al. (2017), is a novel architecture primarily based on the attention mechanism. Unlike previous models, the Transformer eliminates the need for recurrence and convolutions, enabling highly parallelisable computation. This design allows the Transformer to efficiently model long-range dependencies by using self-attention, which calculates relationships between all elements in a sequence simultaneously, regardless of their distance. As a result, the Transformer excels at capturing both short-term and long-term dependencies more effectively than traditional sequential models like recurrent neural networks (RNNs) and long- short-term memory (LSTMs).

Built upon the Transformer architecture, the Temporal Fusion Transformer (TFT) is a state-of-the-art deep learning model specifically designed for time series forecasting. Introduced by Lim et al. (2021), the TFT combines the strengths of transformers and RNNs to effectively handle the complex nature of temporal data. The model's key innovation lies in its ability to capture both long-term dependencies and local patterns within time series data, making it particularly robust for multi-horizon

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

forecasting tasks. A notable advantage of the TFT over other popular models is its capability to produce probabilistic forecasts by predicting multiple quantiles, effectively generating a comprehensive distribution of possible outcomes. This built-in quantile forecasting enables the TFT to inherently account for uncertainties, providing users with valuable insights into potential outcomes.

The TFT architecture features a sequence-to-sequence structure, interpretable multi-head attention mechanism, variable selection network, and static covariate encoders. Similar to the original Transformer architecture introduced by Vaswani et al. (2017), the sequence-to-sequence structure combined with multi-head attention allows TFT to capture both local temporal patterns and long-range dependencies. The variable selection network dynamically weighs the importance of each input feature at every time step, thereby reducing overfitting and increasing the model's robustness when handling complex data. Additionally, static covariate encoders integrate context-specific information that remains constant over time, enabling the model to effectively differentiate between entities or stations. Moreover, the TFT excels at handling heterogeneous data sources by 'fusing' static, known, and observed variables through separate pathways before integrating them into the main architecture, further improving its forecasting capability across various applications.

Another primary strength of the TFT is its interpretability. Unlike traditional deep learning models, which often operate as black boxes, the TFT explicitly provides insights into how different input features and individual time steps influence predictions. This interpretability is facilitated primarily by the multi-head attention mechanism and the variable selection network, providing information on feature importance and temporal relationships within the data. Such interpretability is crucial for applications requiring transparent decision-making and enhances trust in the model's predictions.

The TFT has been applied across various domains with promising results. In finance, it has been used for predicting stock prices and market trends, leveraging its ability to process and integrate multiple data sources and indicators Hajek and Novotny (2024). In healthcare, the TFT has shown potential in predicting emergency department admission to optimize staffing levels Caldas and Soares (2023). The energy sector has also benefited from the TFT's capabilities, particularly in load forecasting and renewable energy production prediction, where accurate multi-horizon forecasts are essential for operational planning and management (Giacomazzi et al., 2023; Jenko and Costa, 2024) .

In this study, we deploy the TFT for streamflow time series prediction given its strengths in handling complex, multi-horizon forecasts and its ability to incorporate diverse data sources. Streamflow data often exhibit intricate temporal patterns influenced by various factors such as precipitation, temperature, and catchment characteristics. The attention mechanism of TFT models can effectively capture these long-term dependencies and localised variations, leading to more accurate and interpretable forecasts. Additionally, the ability to include static and dynamic input features allows the model to integrate essential hydrological and meteorological variables, enhancing the robustness of streamflow predictions. (Liu et al., 2024) showed that Transformer models are competitive against the most well-established streamflow prediction algorithm, LSTM (Kratzert et al., 2018).

Our forecasting horizon extends up to 32 days, aiming for the sub-seasonal scale (which is defined as beyond 2 weeks (White et al., 2022)). Spatially, one model is trained for all 101 stations to include as much diverse data as possible (Kratzert et al., 2024). The encoder period, which captures recent past data before the forecasting horizon, is set to 64 days. The target variable is the observed daily streamflow. The dataset is split into three periods: a training period from January 1991 to April 2012; a

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

validation period from May 2012 to June 2015; and a testing period from July 2015 to December 2018. Training runs for a maximum of 100 epochs, with 100 batches of encoder-decoder sequences randomly sampled from the time series data across all 101 stations in each epoch. Table S1 in the Supplementary details the specific architecture used.

An attention-based model like the TFT can be computationally intensive due to the explicit modelling of interactions between input-output elements (Caldas and Soares, 2023). In our case, training requires approximately one to two hours on a high-performance computing (HPC) cluster using a single core and 15 GB of RAM. The exact time varies depending on the number of input features, but this is considered computationally manageable.

4.2 Static Features

In the TFT, static features are time-invariant variables used to provide context and differentiate between entities (or stations, in our case), helping the model make more informed and accurate predictions across different time steps and stations. The static features we include are summarized in Table 2. These features are chosen for their roles in streamflow generation, particularly within the Alpine region. The features "Country", "Local River", and "Large Basin" (e.g. Rhine or Danube) provide spatial information, helping to contextualize the catchment within broader hydrological and geographical frameworks.

The catchment area is included for their impact on model performance (Poncelet et al., 2017; Harrigan et al., 2020; Andersson et al., 2015; Pappenberger et al., 2015). The challenge for prediction in smaller catchments arises due to greater hydrological variability, more localized precipitation patterns, and the influence of local features such as land use and soil type. Additionally, smaller catchments generally have quicker response times to precipitation events, necessitating higher-resolution data and more detailed modelling approaches.

Elevation is included because it strongly influences precipitation patterns and snow accumulation—particularly in Alpine regions—thus affecting the timing and magnitude of streamflow. In traditional dynamical models, this relationship is typically represented with a temperature lapse rate, linking elevation to temperature variations and thus influencing snow accumulation and melt processes (Andersson et al., 2015).

The flow regime is included to indicate the seasonality of high flow periods and whether a catchment is nival or pluvial. We believe this distinction has added value for predicting streamflow patterns, as rain-fed catchments respond quickly to precipitation events, while snow-fed catchments have delayed responses due to snowmelt processes.

Glacier coverage is included to provide insights into the contribution of ice melt to streamflow. Glaciers act as natural reservoirs, releasing water during warmer months, which can sustain streamflow during dry periods. The presence of glaciers can therefore delay maximum seasonal flow and moderate annual and monthly runoff variations (Fountain and Tangborn, 1985).

Lakes and reservoirs are included because they have a damping effect on streamflow (Quin and Destouni, 2018). They can store excess water during periods of high precipitation and release it during dry periods, thus smoothing out the variability in streamflow. This buffering capacity is essential for understanding the timing and magnitude of flow downstream.

By incorporating these static features, the model can better capture the complex interactions between various catchment characteristics and their influence on streamflow generation, leading to more accurate and reliable predictions.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

Static Feature	Description	Class	Value
Country	country name	category	FR, CH, IT, DE, AT, or SI
Local River	tributary name	categorical	54 river names
Large Basin	large catchment name	categorical	Danube, Rhine, Rhône, or Po
Area	catchment size in km^2	numeric	75 - 101800
Elevation	median elevation of catchment in m	numeric	76 - 1076
Flow Regime	type of hydrological regime	categorical	Pluvial or Nival
Glacier	percentage of catchment covered by glacier	numeric	0 - 13.7%
Lake	percentage of catchment covered by lake	numeric	0 - 4.9%

Table 2. Summary of static features applied in all models.

4.3 Varying Features

All models include the sine and cosine transformations of the day of the year as varying features. These periodic features help the model capture seasonal and weekly patterns in the data.

4.3.1 Training

The models are distinguished by their use of different key predictors: Weather Regime Indices (IWR), Weather Regime Life Cycle (WR_LC), Observations (OBS), and EFAS reanalysis data (EFAS). We take a perfect-model-like approach, using reanalysis data for training instead of forecast data to eliminate forecasting errors in the three model simulation datasets: IWR, WR_LC, and EFAS. To explore the contribution of these predictors to the hybrid framework, we train 11 models with various combinations. The simplest model uses WR_LC, representing the dominant weather regime. To increase data complexity, IWRs are incorporated to capture the full field of large-scale circulation over the Atlantic. Another model type uses observational data (OBS), available only at gauged locations. Finally, we introduce EFAS data, which provides process-based information, benefiting from the EFAS system's predictive capability. Refer to Table 3 for the different combinations.

Each model is trained using 11 different seeds to assess the model's stability and robustness. The use of multiple seeds ensures that the model performance is not dependent on a particular initialization and helps in evaluating the variability in the results. The model output consists of seven streamflow quantiles for each station and 32 days of lead time.

4.3.2 Testing/Forecasting

We first evaluate the trained model with reanalysis data the models have not yet seen in the training and validation period. To assess the true forecasting ability of the trained models in a quasi-operational setting, we replace the WR and EFAS reanalysis data with their respective reforecast datasets during the decoding period, but only for the testing phase (July 2015 to December 2018). Unlike the reanalysis data, which provides a single deterministic value, the reforecast datasets consist of 11 ensemble

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

Model Name	TFT	Static Features	WR_LC	IWR	OBS	EFAS
RAW EFAS						x
WR_LC	x	x	x			
IWR	x	x		x		
OBS	x	x			x	
WR_LC+OBS	x	x	x		x	
IWR+OBS	x	x		x	x	
EFAS	x	x				x
WR_LC+EFAS	x	x	x			x
IWR+EFAS	x	x		x		x
OBS+EFAS	x	x			x	x
WR_LC+OBS+EFAS	x	x	x		x	x
IWR+OBS+EFAS	x	x		x	x	x

Table 3. Summary of model compositions.

members. As a result, the TFT model generates 11 ensemble outputs for each of the 11 seeds, with each output consisting of 7 quantiles.

4.4 Verification

Modified Kling–Gupta efficiency Skill Score (KGESS')

The Kling–Gupta Efficiency (KGE) metric (Gupta et al., 2009) is widely used for hydrological model evaluation, and a revised form (KGE′) was later proposed by Kling et al. (2012) to reduce dependencies among its components. KGE' provides an integrated view of model performance across correlation, bias, and variability. A detailed description of KGE′ was provided in Chang et al. (2023). The KGE' skill score is calculated by comparing the KGE' of the forecast to that of a reference model. It is defined as:

$$\mathrm{KGESS} = \frac{\mathrm{KGE} - \mathrm{KGE}_{\mathrm{ref}}}{\mathrm{KGE}_{\mathrm{perf}} - \mathrm{KGE}_{\mathrm{ref}}} \tag{1}$$

The choice of benchmark can influence the measure of skill (Pappenberger et al., 2015), and here we use mean flow as the benchmark. Knoben et al. (2019) demonstrated that we can assume a KGE' value of -0.41 for mean flow (KGE_clima = 1 - sqrt(1+1+0) = -0.41, where $\mu_s = \mu_o$, correlation coefficient is 0, and the standard deviation of climatology is also zero). The same benchmark was applied in the assessment of GloFAS by Harrigan et al. (2020) and has since been employed in the evaluation of EFAS (Chang et al., 2024), supporting its application here. The highest KGESS' is 1, while a value of 0 indicates

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

that the forecast has no skill compared to the mean flow. Negative values indicate that the forecast is worse than the mean flow prediction.

4.4.1 Continuous Ranked Probability Skill Score

The Continuous Ranked Probability Skill Score (CRPSS) is a widely used metric for evaluating the accuracy of probabilistic forecasts. The CRPSS is derived from the Continuous Ranked Probability Score (CRPS), which measures the difference between the predicted and reference distributions (Hersbach, 2000).

Similar to the KGESS', CRPSS is then calculated by comparing the CRPS of the forecast to that of a benchmark forecast, a climatology model in this case. It is defined as:

$$\text{CRPSS} = 1 - \frac{\text{CRPS}_{\text{forecast}}}{\text{CRPS}_{\text{climatology}}} \tag{2}$$

The climatology model is derived from the observed streamflow over the testing period as our reference. Same as KGESS', a CRPSS value of 1 indicates a perfect forecast, while a value of 0 indicates that the forecast has no skill compared to climatology. Negative values indicate that the forecast is worse than the climatology.

5 Results

5.1 Reanalysis

In this section we present results obtained from the 11 trained TFT models with different input datasets (see Table 3) and evaluate their skill in modelling streamflow during low-flow days at the selected 101 stations across our study area. We begin by analysing the models' performance with reanalysis data inputs, using KGESS' (benchmarked against a mean flow model) to assess their skill. Although the reanalysis input data is deterministic, the TFT model produces probabilistic outputs in the form of seven quantiles. While the model's performance could theoretically be evaluated using ensemble verification scores like CRPSS, we opt to use KGESS' for comparing the trained models' reanalysis testing results to the raw deterministic EFAS output.

Figure 4 shows the KGESS' for low flow periods, averaged across all stations for each of the 11 simulation seeds. Results are presented for all 101 stations (pink) and only the non-calibrated stations (teal). Even the TFT model with only the dominant weather regime type as the key predictor (*WR_LC*) is skilful against mean flow. Changing the WR information from only the dominant type to the seven IWRs in the TFT model *IWR* results in an increase in skill. The skill of the WR TFT models demonstrates that the TFT algorithm is able to capture the connection between WRs and streamflow as shown in Section 3.2, and the models can benefit from the more detailed information of the full Z500 field provided by the IWRs. These results quantify the fundamental relevance of large-scale atmospheric processes in determining streamflow during low-flow situations.

When in-situ observational data are available, the TFT model *OBS*, which uses observed streamflow data in the encoder period as initial condition (can be considered as building a 'persistence-like' model), reaches a similar skill as the TFT model

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

IWR with a reduced KGESS' variability. The reduced variability from the different seeds indicates a more stable and robust model. Although the TFT model *IWR* and *OBS* have similar skill individually, when combining WR information with observed streamflow data, we observe a clear jump in skill for TFT models *WR_LC+OBS* and *IWR+OBS*.

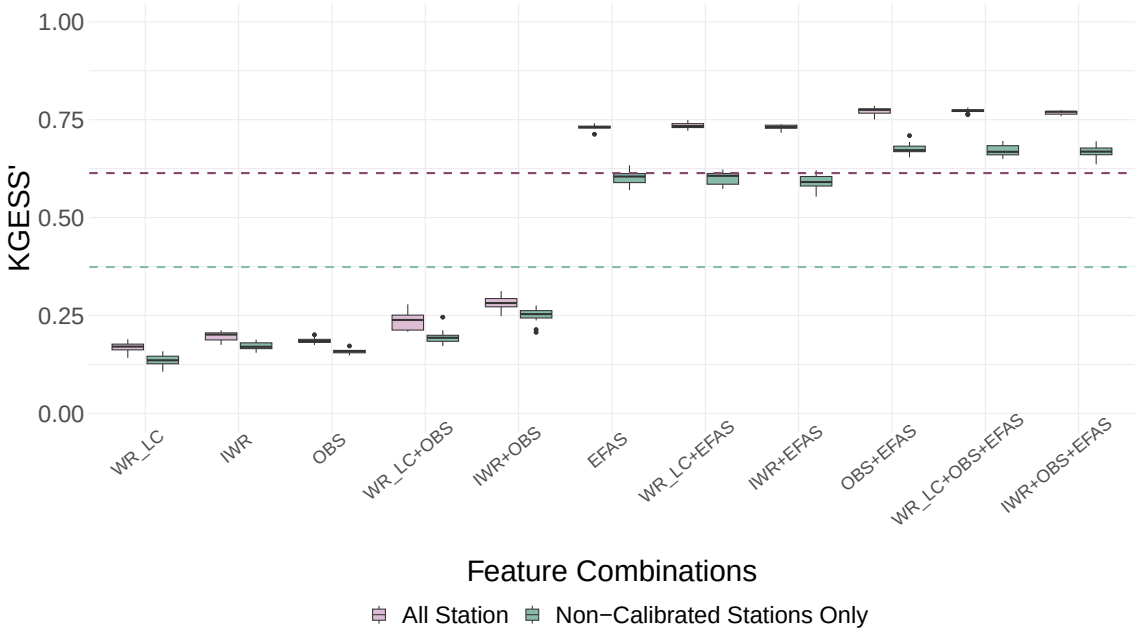

Figure 4. Comparison of KGESS' model performance with various feature combinations, evaluated using reanalysis data on low-flow days during the testing period (July 1, 2015 - December 28, 2018). Each model configuration includes 11 seeds. Refer to Table 3 for information on model setups. Dotted lines represent the reference KGESS' from raw EFAS output. Results averaged over all 101 stations are shown in light pink, while teal represents non-calibrated EFAS stations.

Among the trained models, the largest increase in skill occurs when process-based information—in this case, EFAS output—is provided as input. Specifically, the TFT model *EFAS* outperforms the raw EFAS output itself (pink dotted line in Figure 4). As shown previously, for models using weather regimes (WR), incorporating full-field information from the seven IWRs (TFT models *IWR* and *IWR+OBS*) provides higher skill than using WR_LC alone (TFT models *WR_LC+EFAS* vs. TFT model *IWR+EFAS*). However, once process-based information (EFAS) is included, no differences in skill improvement is observed from the choice of WR data (*WR_LC+EFAS* vs. *IWR+EFAS*). Adding large-scale circulation information becomes redundant in the presence of EFAS data.

Introducing observed streamflow data along with EFAS output (*OBS+EFAS*) further improves model skill, yet again, additional inclusion of WR information provides no extra benefit. This limited utility of WR data in models using EFAS likely arises because EFAS already captures most meteorological signals that WRs represent. Consequently, ML models can achieve optimal performance by selectively utilising relevant, non-redundant inputs, ensuring more efficient and precise forecasting. A similar pattern is evident when evaluating all initialisation dates as well as high-flow days specifically (not shown here).

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

For the non-calibrated EFAS stations (teal boxes in Figure 4), the same general evolution of model performance is observed as for all stations. For the TFT models using EFAS input, the proposed framework brings the performance of non-calibrated stations to a level comparable to the raw EFAS output averaged across all stations (pink dotted line).

5.2 Forecast mode

To assess the operational potential of our hybrid models, we select six out of the eleven trained models based on their KGESS scores from the previous section. The six selected models are the TFT models: *IWR*, *OBS*, *IWR+OBS*, *EFAS*, *OBS+EFAS*, and *WR_LC+OBS+EFAS*. Predictions are generated using reforecast data, incorporating uncertainties from the meteorological forecasts embedded in the WR and EFAS reforecast data. Figure 5 presents the CRPSS values for the raw EFAS reforecast (in dark blue) as a reference, alongside the six selected models, evaluated on low flow days between July 1, 2015, to December 31, 2018 (testing period).

The results demonstrate that the selected models, despite being trained with reanalysis data, consistently exhibit forecasting skill relative to climatology, achieving CRPSS values above zero. As shown in Figure 5a (all stations), the TFT model *IWR* (green line) maintains positive CRPSS across the entire 32-day forecast horizon. Although the TFT model *IWR* does not outperform the raw EFAS system, this result indicates that, in the absence of other forecasting models, using the WR-based model alone still provides substantial improvements over climatology. Additionally, the *IWR* model demonstrates notably lower variability across stations compared to *RAW EFAS*, reflecting greater consistency in its performance. When considering non-calibrated stations exclusively (Figure 5b), the TFT model *IWR* matches the median CRPSS of *RAW EFAS* for the first two weeks and surpasses it at longer lead times, although uncertainties associated with this improvement are relatively large. The rest of the TFT models reach a higher skill compared to the raw EFAS model. This improvement is largest at the beginning of the forecast period, with the skill differences between the trained models and the raw EFAS model reducing over time.

Figure 6 illustrates the performance of different TFT models at four selected stations, one from each major basin. These stations exhibit patterns consistent with those discussed for Figure 5, with one notable exception at Station Rheinhalle. At this station, the performance of the TFT models *EFAS* and *OBS+EFAS* is indistinguishable, suggesting that for large catchments, incorporating observed initial conditions provides limited additional value once process-based EFAS information is included. This result aligns with previous findings indicating that EFAS typically performs better in larger catchments, particularly for flood forecasting (Alfieri et al., 2014).

The TFT model *OBS* demonstrates higher skill than the raw EFAS model, and it even surpasses the TFT model *EFAS* during the first 5 days based on median station values, highlighting its ability to capture the persistence of low-flow conditions. This persistence effect is more pronounced at non-calibrated stations, where the *OBS* model maintains comparable skill to the *EFAS* model across the entire 32-day forecast horizon. However, when considering all stations, the forecasted streamflow data from EFAS improves the TFT model skill from around day 5 to two weeks ahead, where the TFT model *EFAS* consistently outperforms the *OBS* model. During high-flow periods (Figure S10b, Supplementary), the initial streamflow persistence provides the *OBS* model comparable skill to the *EFAS* model only on day 1. Beyond that, its skill rapidly declines, dropping below climatology after roughly one week. This contrast emphasizes that while the *OBS* model effectively leverages persistence to

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

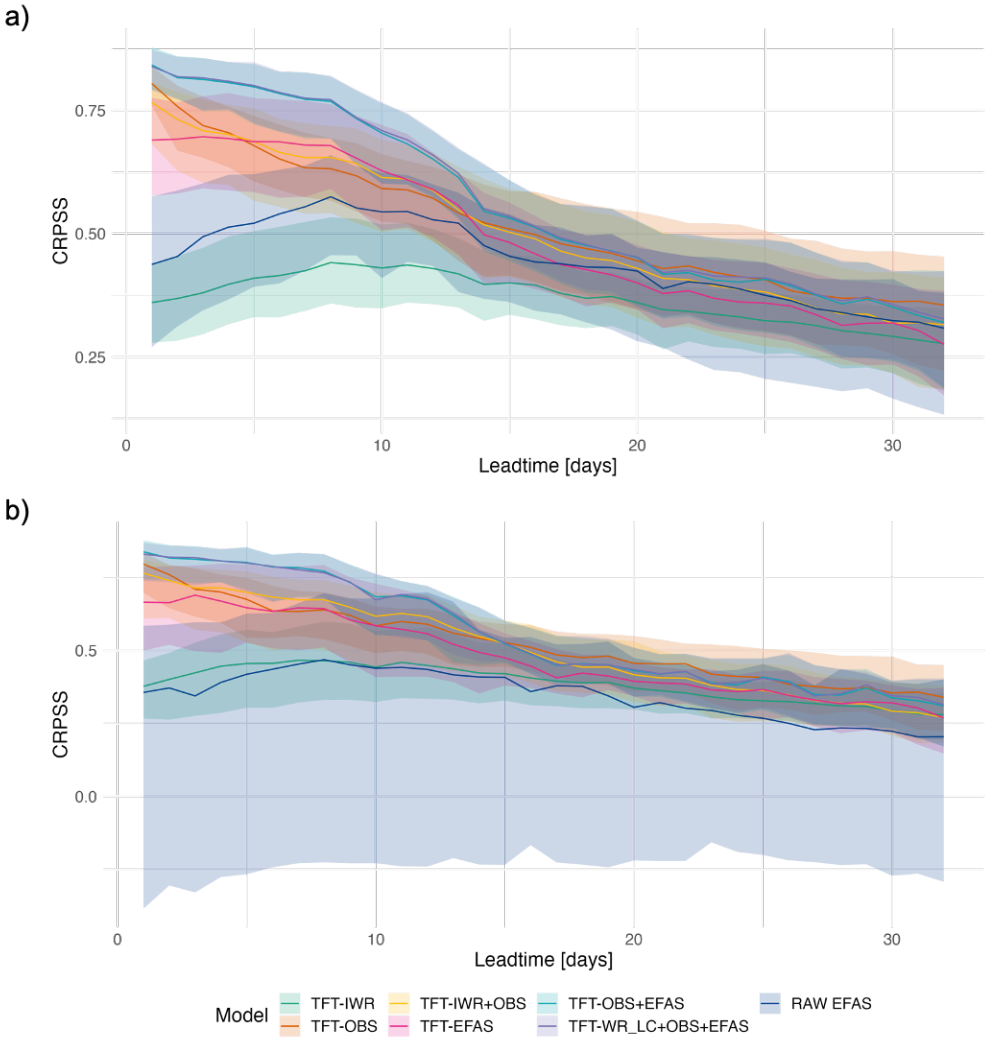

Figure 5. Comparison of CRPSS over a 32-day lead time for selected models with various input features. Information on model setup can be found in Table 3. The raw EFAS output, shown in blue, serves as a baseline. The shaded regions represent the interquartile range of CRPSS values across 101 stations and 11 seeds during the testing period (July 1, 2015 - December 28, 2018). Panel (a) presents results for low flow days across all stations, while panel (b) focuses on non-calibrated EFAS stations on low flow days. Additional year-round and high-flow results are provided in Figure S10 in the Supplementary.

capture slower dynamics typical of low-flow conditions, it struggles to maintain predictive skill under faster, more dynamic high-flow conditions.

The best performing models are the TFT model *OBS+EFAS* and the TFT model *WR_LC+OBS+EFAS*, both achieving similar CRPSS values that surpass all other models during the first weeks. While certain individual stations in specific seasons

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

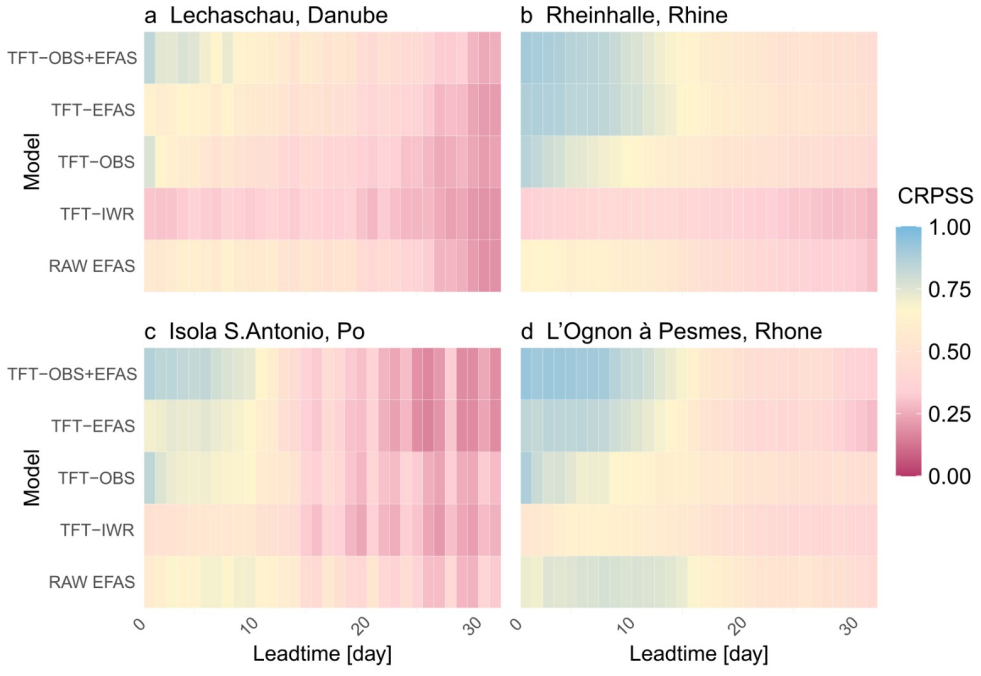

Figure 6. CRPSS comparison at selected stations. The panel name specifies the station and river basin names. Results include low flow periods only.

may benefit from the inclusion of WR data, the overall added value of WR across all stations is not evident. Given that the TFT model *OBS+EFAS* attains comparable skill as the TFT model *WR_LC+OBS+EFAS* with less input data, it is regarded as the most effective model. The skill observed when combining in-situ observations with EFAS inputs likely arises from their complementary nature. In-situ observations provide precise, localized data that enhances the initial conditions, while EFAS offers broad-scale hydrological insights. This combination allows the model to leverage the accuracy of in-situ data while still benefiting from the broader coverage and predictive capability of EFAS, ultimately improving overall forecast skill.

The skill gap between the TFT models and the raw EFAS becomes difficult to distinguish after week two when considering all stations (Figure 5a), suggesting that the models may reach a ceiling of possible improvement with the current setup. When focusing solely on non-calibrated stations, the TFT models converge to a similar CRPSS value while maintaining a clear gap above the raw EFAS. This is likely because these stations present greater discrepancies and bias compared to observation, allowing the models to derive more valuable information from the input features. It's also worth noting that the TFT models notably reduce the IQR of the raw EFAS, particularly at non-calibrated stations. The variability due to seeds is shown to be relatively small (see Figure S9 in the Supplementary), suggesting the variability in Figure 5 is due to the differences among stations.

We investigate the improvement in CRPSS achieved by the TFT model *OBS+EFAS* for the first three days of forecasts relative to the raw EFAS model for the different subgroups (see Figure 7 and Table A1 in the Appendix). Low flow peri-

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

ods experience slightly less CRPSS improvement compared to high flow periods, indicating the model's greater efficacy in correcting high flow events. Seasonally, the largest improvements are observed during the summer season of JJA (June-July-August), while the other seasons, particularly MAM (March-April-May), exhibit more modest improvements. The high level of improvements in JJA could be due to the fact that more high flow events occur during the summer months.

Supporting the results shown in Figure 5, a comparison between calibrated and non-calibrated EFAS stations in Figure 7a shows that non-calibrated stations experience greater improvements with the TFT model *OBS+EFAS*. This indicates the TFT model is particularly effective at compensating for initial biases or deficiencies in the raw EFAS forecasts when calibration is absent. The spatial pattern in Figure 7b further illustrates that these improvements are not uniformly distributed, with the highest gains concentrated at non-calibrated stations. Such variability underscores the importance of targeted implementation of TFT models, suggesting that forecasting systems could achieve optimal effectiveness by prioritising deployment at locations where traditional model calibration is limited or absent, thus enhancing decision-making in areas currently underserved by calibrated forecasts.

Figure 8 compares the number of skilful days for the TFT model *OBS+EFAS* versus the raw *EFAS* model during low-flow periods across different CRPSS thresholds. The majority of stations are positioned above the 1:1 line, indicating a clear extension of the skilful horizon for the TFT model *OBS+EFAS*. Notably, the stations with the greatest gains in skilful days are predominantly non-calibrated stations, demonstrating the model's effectiveness in improving forecasts where the raw *EFAS* model typically underperforms. It is important to note that several stations may be represented by the same point on the plot, especially when they share the same marker shapes, which can make overlaps difficult to see. This overlap highlights the consistency of these skill gains across similar station types.

6 Discussion

6.1 Interpretability

6.1.1 Feature importance

Figure 9 presents the feature importance analysis of the TFT model *'WR_LC+OBS+EFAS* during the testing phase with reanalysis data. The static features generally exhibit varying levels of importance across different stations, reflecting the distinct static characteristics unique to each location. Since static features are constant and time-invariant, their importance remains consistent across different initialisation dates for a given station. Figure 9a presents the percentage importance of each static feature across all stations and random seeds, ordered by their corresponding median values. Glacier coverage stands out with the highest median importance, highlighting its critical role in streamflow generation in the Alpine region. Although the median and mean importance percentages show only slight differences across the static features, there are noticeable differences in their variabilities. Nevertheless, the mean importance values cluster around the baseline, suggesting that no single static feature overwhelmingly dominates the model's predictions.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

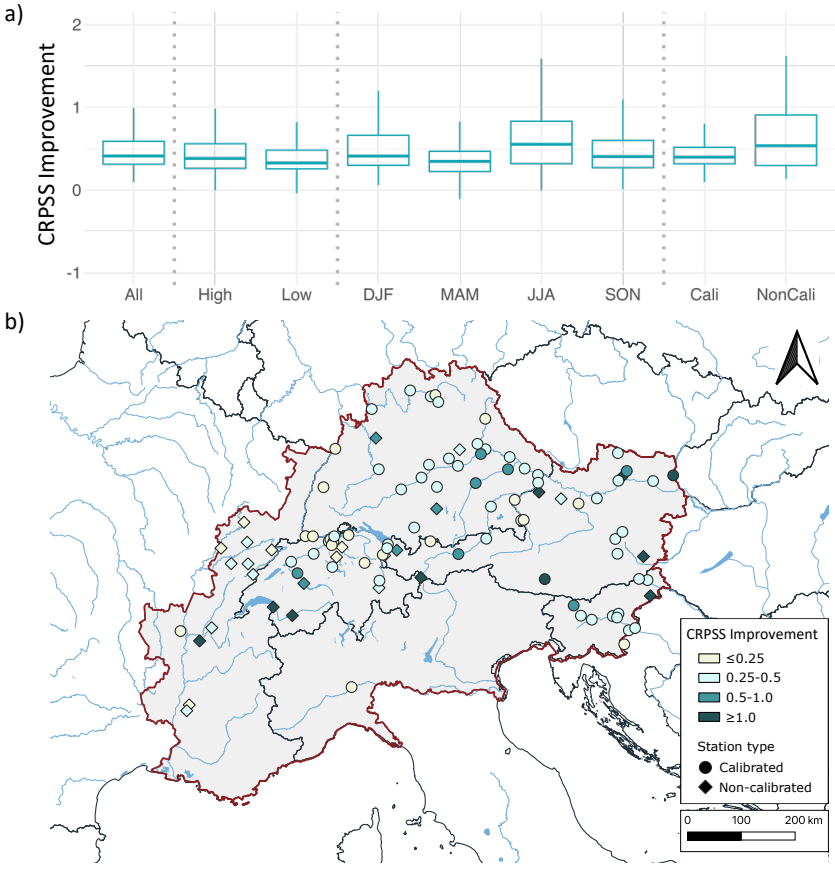

Figure 7. CRPSS improvement achieved by the best-performing TFT model ('OBS+EFAS') compared to the raw EFAS output. Panel (a) displays the results categorized into various subgroups. Detailed descriptions of these subgroups can be found in Table A1. Panel (b) illustrates the low flow results spatially on a map. Calibrated stations are represented by circles, while non-calibrated stations are indicated by diamonds.

This large spread in the distributions indicates that the relevance of static features is highly station-specific. Different stations rely on different static features to varying extents, likely due to their unique geographical or environmental characteristics. The absence of a universally dominant feature across all stations implies that the model does not rely heavily on a fixed set of static features for its predictions. Instead, it demonstrates adaptability, leveraging a diverse set of features depending on the specific context of each station. The balanced yet low overall importance of static features may indicate that these features, while informative, are not the primary drivers of the model's predictive performance. This could suggest that other types of features, such as temporal or dynamic factors, are more critical in predicting streamflow. Additionally, the variability in feature importance highlights the complexity of the underlying factors influencing streamflow and suggests that a more nuanced or localised approach might be necessary for further refinement.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

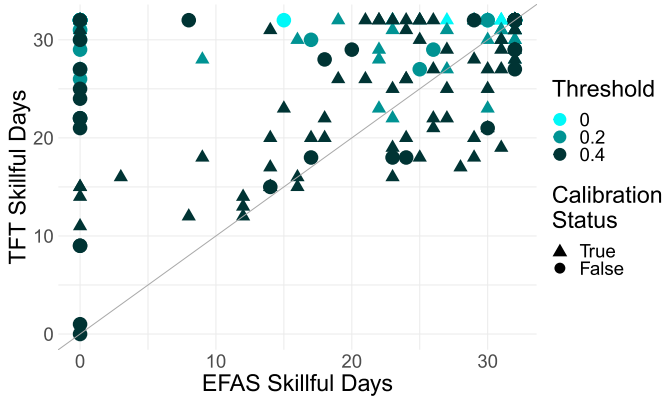

Figure 8. Skillful forecast horizon in days by TFT model 'OBS+EFAS' compared to the raw EFAS output. The comparison includes both calibrated (circles) and non-calibrated (triangles) stations, with marker colors representing the magnitude of the threshold. The diagonal line indicates where the skillful forecast days of the TFT model 'OBS+EFAS' equals that of the raw EFAS output. Points above the line indicate a gain in skillful days with TFT model 'Obs+EFAS', while points below indicate a reduction.

Figure 9b illustrates the frequency with which each feature appears as one of the three most important features for a station. The figure suggests that features of glacier coverage, catchment area, catchment elevation, local river name, and lake coverage are more frequently among the top three important features, consistent with Figure 9a. However, the differences among these features are relatively small (also consistent with Figure 9a).

Overall, the model shows signs of generalisation and flexibility, effectively using the most relevant features based on the specific conditions of each station. This context-dependent feature importance emphasises the necessity of considering local conditions in predictive modelling, which may guide future efforts to improve feature selection and model accuracy.

While the importance of static features does not change from one prediction to another, the importance of encoder and decoder variables does. Figures 9c and d show the importance of the dynamic features in the TFT model. The encoder features (Fig.9c) show that, in the encoder period, EFAS streamflow output is the most important feature, followed by streamflow observations. The high feature importance of EFAS data is somewhat surprising, as one might expect the initial condition from streamflow observations to carry more weight in shaping short-term predictions. This outcome suggests that the model prioritises the broader hydrological context and process-based forecasts provided by EFAS over the localised conditions captured by direct observations. In the decoder period (Fig.9d), EFAS input clearly dominates, as expected, given its role in providing process-based information. WR data is deemed least important in both periods, which matches the results where WR data addition to TFT model *OBS+EFAS* did not further improve skill (cf. Fig. 4).

6.1.2 Attention

The attention and self-attention mechanisms are fundamental components of transformer-based algorithms, allowing the model to attend to the most informative part of the input by assigning varying importance to different time steps (Vaswani et al., 2017).

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

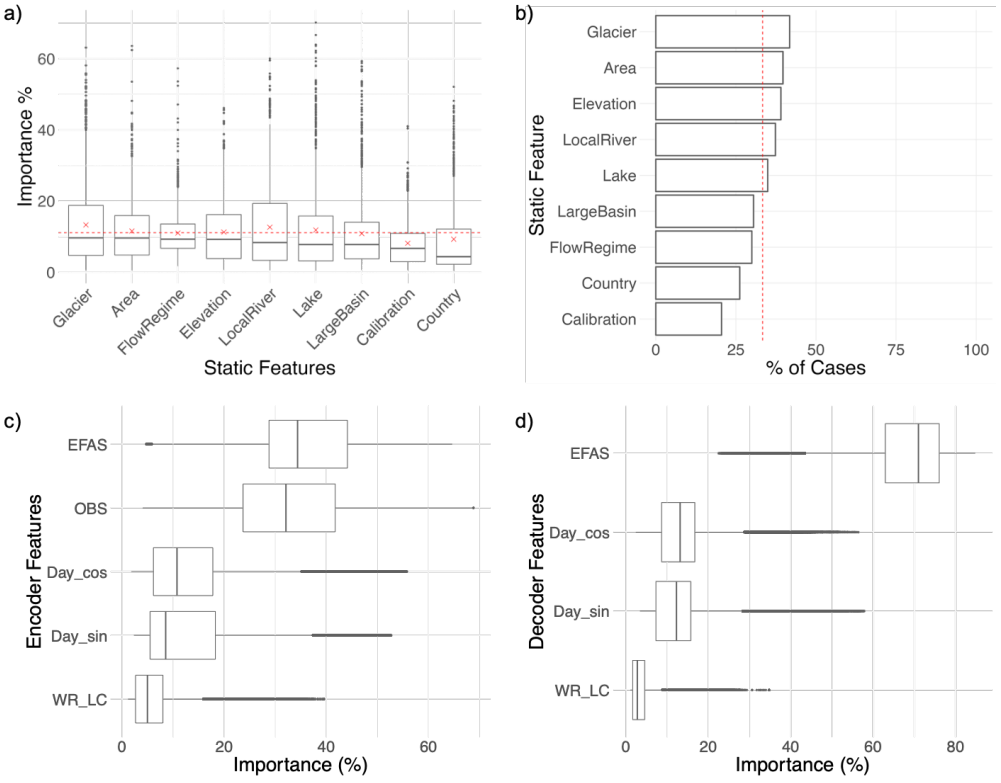

Figure 9. Feature analysis of TFT model *WR_LC+OBS+EFAS'*. Panel (a) displays the importance of each static feature across all stations and seeds as box plots, ordered by their corresponding median values. The red dotted line (11.1%) represents the reference level for equal importance among features, and the red cross markers indicate the mean importance of each feature. Panel (b) shows the frequency with which each static feature appears as a top three important feature, with the red dotted line representing the reference point based on a random order of features for each station, Panels (c) and (d) illustrate the feature importance for the encoder and decoder features, respectively.

Figure 10a shows the analysis of the average attention weights assigned to each time step (in days) during the encoding period, effectively illustrating how the model prioritises information from different days. Days closer to the start of the prediction period (i.e. decoder phase) generally receive higher attention weights. However, this relationship is not strictly linear, as evidenced by the fluctuations seen in Figure 10c and d. These panels show specific instances where certain days within the encoding period exhibit higher importance, likely due to their relevance in capturing significant hydrological events that influence the forecast.

Figure 10b, inspired by Padrón et al. (2025), presents the mean attention weights for each initialisation across all stations for the year 2018. The plot shows a pattern where the attention gradually increases as key events approach the start of the forecasting horizon (seen as sloped straight lines representing a fixed point in time), indicating the model's ability to dynamically adjust its focus based on the temporal proximity of important hydrological occurrences. This behaviour, also observed in Padrón et al. (2025), suggests that the model is not merely relying on recent data but is instead selectively emphasizing time

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

steps that contribute most effectively to the prediction, highlighting the adaptive nature of the attention mechanism in capturing relevant temporal features, that is events with hydrological relevance in this case.

Moreover, when connecting to the WR analysis in Section 3.2, this attention pattern raises an interesting opportunity for further investigation. Specifically, one could analyse the dominant WR types that coincide with days receiving higher attention weights. Such an analysis could provide insights into how specific large-scale circulation patterns influence the model's focus and, consequently, the predictive skill during critical periods. This could also help refine the model's interpretability, enabling a clearer understanding of how the TFT model leverages both temporal and atmospheric information to enhance hydrological forecasts.

6.2 Limitations and outlook

This section outlines the limitations of the study and proposes potential directions for future research to address these gaps and build upon the current findings.

- This study develops a hybrid forecasting framework using diverse input data types, but it does not compare the proposed TFT approach directly with conventional post-processing methods such as quantile mapping or simpler ML models. Simpler ML methods typically require separate models for each station, limiting their ability to capture spatial relationships and complicating model management. While simpler models might initially appear computationally more efficient, advancements in transformer-based deep learning now support efficient parallel processing. Additionally, simpler models often provide less interpretability compared to attention-based deep learning approaches, which help identify key drivers of predictability. Future research could systematically evaluate the trade-offs between computational efficiency, interpretability, and predictive skill by directly comparing TFT with simpler ML methods under controlled conditions.

- The process-based hydrological data used in this study are from EFAS v4.0, which was the latest version providing a complete set of reanalysis data at the time of our framework development. Although EFAS v5.0 (https://confluence.ecmwf.int/display/CEMS/EFAS+versioning+system), a newer version offering higher spatial resolution and other improvements, became available during the study period, we selected EFAS v4.0 to ensure data completeness and consistency. Given the performance of the hybrid framework with EFAS v4.0, we expect it to perform similarly well with EFAS v5.0. Future research should evaluate the framework's application using EFAS v5.0 to assess potential gains from enhanced model inputs, as well as explore its adaptability to other process-based models to support broader generalizability.

- The spatial distribution of the selected stations within the study area is uneven, with certain regions, like the river Po in Italy, being under-represented, which is due to the availability of high-quality observational data. A more balanced station distribution could potentially reduce spatial discrepancies in model performance, thereby providing a more comprehensive assessment of the model's generalizability across the Alpine region.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

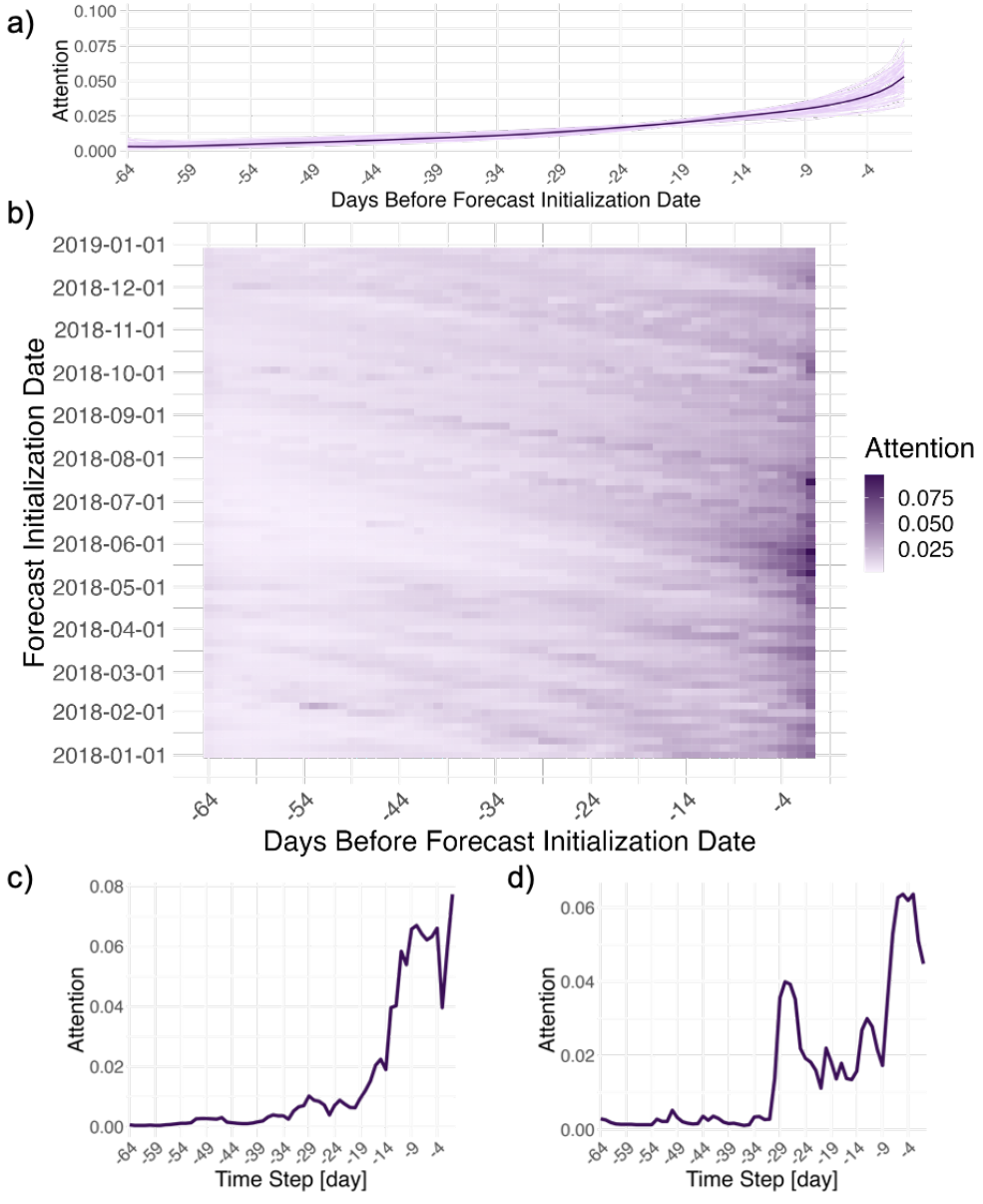

Figure 10. Attention weight analysis of TFT model 'WR_LC+OBS+EFAS'. The visualisation is adapted from the approach by Padrón et al. (2025). Panel (a) shows the attention weight per station averaged over all seeds and initialisation dates, with each thin line representing a station and the thick purple line indicating the median attention weight across all stations. Panel (b) presents a heatmap of attention weights, where each row corresponds to one initialisation date averaged over all stations and seeds, focusing on the year 2018. Panels (c) and (d) provide examples of attention weights at stations Rheinhalle (CH) and Neudorf bei Ilz (AT) initialised on 2018-07-17.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

- Our results show that the greatest improvements in skill are observed at the non-calibrated EFAS stations, hihglighting the potential of the hybrid model in ungauged areas (Figure 7). To assess its generalization ability, future studies should test the model on unseen non-calibrated stations, offering insights into its applicability for enhancing hydrological forecasting in data-scarce regions.

- Our study does not evaluate the hybrid system's ability to assimilate human influence or directly incorporate anthropogenic influences, such as the extensive network of reservoirs used for hydropower generation and flood protection, which significantly alter natural streamflow patterns. Future research should focus on evaluating the ML model's ability to capture patterns associated with human activities and identifying additional predictors or proxies that incorporate human factors to enhance predictive accuracy. Despite challenges in data availability, such work can potentially lead to a deeper understanding of the hydrological processes in regions affected by human activities.

7 Conclusions

In this study, we develop a hybrid framework for low-flow forecasting in the European Alps using the Temporal Fusion Transformer (TFT). We systematically evaluate the contribution of three distinct datasets: large-scale circulation weather regime data, in-situ observation data, and process-based model simulation data from the European Flood Awareness System (EFAS).

Our analysis of the physical connections between hydrological behaviour and weather regimes forms the basis for our selection of input features, ensuring that the model does not merely represent a black box but is instead grounded in an understanding of the physical system. The findings suggest that the TFT model, even when using only the weather regime data, achieves a forecast skill superior to that of climatology. The model performance further improves when combining weather regime data with observation data, underscoring the value of integrating diverse sources of information. The inclusion of process-based simulation data from EFAS leads to another significant enhancement in model performance. However, once the process-based information is incorporated, adding weather regime data does not yield further improvements in forecast skill.

Additionally, we quantify the sources of model performance by analysing feature importance and attention weights within the model. This analysis provides deeper insights into how each data source contributes to the forecasting skill, highlighting the critical role of process-based simulations when combined with observational data.

In conclusion, the hybrid framework developed here demonstrates clear bias correction capabilities, with predictive skill surpassing that of traditional approaches and enabling seamless forecasting up to a one-month horizon. Achieving this level of performance is made possible by leveraging the latest deep learning algorithm TFT, alongside the integration of diverse, large datasets. This approach enables the hybrid system to address catchments with high spatial variability while delivering physically realistic results, paving the way for more accurate and reliable low-flow predictions in complex environments beyond the Alps.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

Appendix A:

Table A1. Description of the different temporal analysis subgroups

Subgroup	Descriptions
All	All stations and all initialization dates
High	All stations and only initialization dates where at least one quarter of the decoder periods experience flow higher than the 85th percentile of climatological flow
Low	All stations and only initialization dates where at least one quarter of the decoder periods experience flow lower than the 15th percentile of climatological flow
DJF	All stations and only initialization dates in the months of December, January, and February
MAM	All stations and only initialization dates in the months of March, April, and May
JJA	All stations and only initialization dates in the months of June, July, and August
SON	All stations and only initialization dates in the months of September, October, and November
Cali	Only calibrated stations (73). All initialization dates included.
NonCali	Only non-calibrated stations (28). All initialization dates included.

Author contributions. AC: Writing – original draft, Conceptualisation, Validation, Methodology, Formal analysis, Visualisation. SH: Writing – review & editing, Methodology, Data curation, Conceptualisation. MHR: Writing – review & editing, Methodology, Conceptualisation. CMG: Writing – review & editing, Conceptualisation, Data curation. MZ: Writing – review & editing, Supervision, Investigation, Funding acquisition, Data curation, Conceptualisation. DD: Writing – review & editing, Supervision, Conceptualisation. KB: Writing – review & editing, Supervision, Investigation, Funding acquisition, Conceptualisation, Visualisation.

Competing interests. The authors declare that they have no known competing interests.

Acknowledgements. The authors AC, MZ and KB were supported by the MaleFix project, which is part of the WSL Program Extremes. 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. CMG acknowledges partial funding by KIT and the Helmholtz Association as part of the Young Investigator Group "Sub-seasonal Predictability: Understanding the Role of Diabatic Outflow" (SPREADOUT, grant VH-NG-1243). The author MHR was supported by the CIPRHES and the DRHYM projects, both funded by the Agence Nationale de la Recherche (ANR) in France. MHR contribution was also part of her WSL-supported fellowship. Support from the Swiss National Science Foundation, Switzerland, through project PP00P2_198896 to DD is gratefully acknowledged. We thank Marisol Osman for providing the updated weather regime forecasts

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

565 data and François Tilmant for the streamflow observation data in France. Finally, we thank Ryan Padrón for sharing his ML technical knowledge in support of this project.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

References

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, Journal of Hydrology, 517, 913–922, https://doi.org/https://doi.org/10.1016/j.jhydrol.2014.06.035, 2014.

Andersson, J., Pechlivanidis, I., Gustafsson, D., Donnelly, C., and Arheimer, B.: Key factors for improving large-scale hydrological model performance, European Water, 49, 77–88, 2015.

Bard, A., Renard, B., Lang, M., Giuntoli, I., Korck, J., Koboltschnig, G., Janža, M., d'Amico, M., and Volken, D.: Trends in the hydrologic regime of Alpine rivers, Journal of Hydrology, 529, 1823–1837, https://doi.org/https://doi.org/10.1016/j.jhydrol.2015.07.052, 2015.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrology and Earth System Sciences, 13, 141–153, https://doi.org/10.5194/hess-13-141-2009, 2009.

Bogner, K. and Pappenberger, F.: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, Water Resources Research, 47, https://doi.org/https://doi.org/10.1029/2010WR009137, 2011.

Brunner, M. I., Götte, J., Schlemper, C., and Van Loon, A. F.: Hydrological Drought Generation Processes and Severity Are Changing in the Alps, Geophysical Research Letters, 50, e2022GL101 776, https://doi.org/https://doi.org/10.1029/2022GL101776, e2022GL101776 2022GL101776, 2023.

Büeler, D., Ferranti, L., Magnusson, L., Quinting, J. F., and Grams, C. M.: Year-Round Sub-Seasonal Forecast Skill for Atlantic-European Weather Regimes, Quarterly Journal of the Royal Meteorological Society, n/a, https://doi.org/10.1002/qj.4178, in press, 2021.

Caldas, F. M. and Soares, C.: A Temporal Fusion Transformer for Long-Term Explainable Prediction of Emergency Department Overcrowding, in: Machine Learning and Principles and Practice of Knowledge Discovery in Databases, edited by Koprinska, I., Mignone, P., Guidotti, R., Jaroszewicz, S., Fröning, H., Gullo, F., Ferreira, P. M., Roqueiro, D., Ceddia, G., Nowaczyk, S., Gama, J., Ribeiro, R., Gavaldà, R., Masciari, E., Ras, Z., Ritacco, E., Naretto, F., Theissler, A., Biecek, P., Verbeke, W., Schiele, G., Pernkopf, F., Blott, M., Bordino, I., Danesi, I. L., Ponti, G., Severini, L., Appice, A., Andresini, G., Medeiros, I., Graça, G., Cooper, L., Ghazaleh, N., Richiardi, J., Saldana, D., Sechidis, K., Canakoglu, A., Pido, S., Pinoli, P., Bifet, A., and Pashami, S., pp. 71–88, Springer Nature Switzerland, Cham, ISBN 978-3-031-23618-1, 2023.

CEMS: EFAS versioning system - EFAS v4.0, https://confluence.ecmwf.int/display/CEMS/EFAS+v4.0, accessed: 05.03.2024, 2023.

Chang, A. Y., Bogner, K., Grams, C. M., Monhart, S., Domeisen, D. I., and Zappa, M.: Exploring the Use of European Weather Regimes for Improving User-Relevant Hydrological Forecasts at the Subseasonal Scale in Switzerland, Journal of Hydrometeorology, 24, 1597–1617, https://doi.org/10.1175/JHM-D-21-0245.1, 2023.

Chang, A. Y.-Y., Ramos, M.-H., Harrigan, S., Prudhomme, C., Tilmant, F., Domeisen, D. I., and Zappa, M.: Exploring hydrological system performance for alpine low flows in local and continental prediction systems, Journal of Hydrology: Regional Studies, 56, 102 056, https://doi.org/https://doi.org/10.1016/j.ejrh.2024.102056, 2024.

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

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., and Poligot-Pitsch, S.: Finalisation de l'opération HYDRO 3 de modernisation du système d'information national des données hydrométriques, LHB, 108, 2099 317, https://doi.org/10.1080/27678490.2022.2099317, 2022.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

Fountain, A. G. and Tangborn, W. V.: The Effect of Glaciers on Streamflow Variations, Water Resources Research, 21, 579–586, https://doi.org/https://doi.org/10.1029/WR021i004p00579, 1985.

Giacomazzi, E., Haag, F., and Hopf, K.: Short-Term Electricity Load Forecasting Using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources, in: Proceedings of the 14th ACM International Conference on Future Energy Systems, e-Energy '23, p. 353–360, Association for Computing Machinery, New York, NY, USA, ISBN 9798400700323, https://doi.org/10.1145/3575813.3597345, 2023.

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

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Monthly Weather Review, 133, 1098 – 1118, https://doi.org/10.1175/MWR2904.1, 2005.

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., and Wernli, H.: Balancing Europe's wind-power output through spatial deployment informed by weather regimes, Nature Climate Change, 7, 557–562, https://doi.org/10.1038/NCLIMATE3338, 2017.

Grams, C. M., Magnusson, L., and Ferranti, L.: How to Make Use of Weather Regimes in Extended-Range Predictions for Europe, ECMWF, 2020.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Gurtz, J., Zappa, M., Jasper, K., Lang, H., Verbunt, M., Badoux, A., and Vitvar, T.: A comparative study in modelling runoff and its components in two mountainous catchments, Hydrological Processes, 17, 297–311, https://doi.org/https://doi.org/10.1002/hyp.1125, 2003.

Hajek, P. and Novotny, J.: Beyond Sentiment in Stock Price Prediction: Integrating News Sentiment and Investor Attention with Temporal Fusion Transformer, in: Artificial Intelligence Applications and Innovations, edited by Maglogiannis, I., Iliadis, L., Macintyre, J., Avlonitis, M., and Papaleonidas, A., pp. 30–43, Springer Nature Switzerland, Cham, ISBN 978-3-031-63219-8, 2024.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979–present, Earth System Science Data, 12, 2043–2060, https://doi.org/10.5194/essd-12-2043-2020, 2020.

Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N.: The suitability of a hybrid framework including data driven approaches for hydrological forecasting, Hydrology and Earth System Sciences Discussions, pp. 1–20, 2022.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559–570, 2000.

Horton, P., Schaefli, B., and Kauzlaric, M.: Why do we have so many different hydrological models? A review based on the case of Switzerland, WIREs Water, 9, e1574, https://doi.org/https://doi.org/10.1002/wat2.1574, 2022.

Jenko, J. and Costa, J. P.: Using Temporal Fusion Transformer Predictions to Maximise Use of Renewable Energy Sources, in: 2024 International Workshop on Artificial Intelligence and Machine Learning for Energy Transformation (AIE), pp. 1–10, https://doi.org/10.1109/AIE61866.2024.10561261, 2024.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264–277, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

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

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train an LSTM on a single basin, Hydrology and Earth System Sciences Discussions, 2024, 1–19, https://doi.org/10.5194/hess-2023-275, 2024.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T.: Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting, 37, 1748–1764, https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.03.012, 2021.

Liu, J., Bian, Y., Lawson, K., and Shen, C.: Probing the limit of hydrologic predictability with the Transformer network, Journal of Hydrology, p. 131389, https://doi.org/https://doi.org/10.1016/j.jhydrol.2024.131389, 2024.

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., and Liniger, M. A.: Skill of Subseasonal Forecasts in Europe: Effect of Bias Correction and Downscaling Using Surface Observations, Journal of Geophysical Research: Atmospheres, 123, 7999–8016, https://doi.org/10.1029/2017JD027923, 2018.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., and Matias, Y.: Flood forecasting with machine learning models in an operational framework, Hydrology and Earth System Sciences, 26, 4013–4032, https://doi.org/10.5194/hess-26-4013-2022, 2022.

Ng, K., Huang, Y., Koo, C., Chong, K., El-Shafie, A., and Najah Ahmed, A.: A review of hybrid deep learning applications for streamflow forecasting, Journal of Hydrology, 625, 130 141, https://doi.org/https://doi.org/10.1016/j.jhydrol.2023.130141, 2023.

Osman, M., Beerli, R., Büeler, D., and Grams, C. M.: Multi-model assessment of sub-seasonal predictive skill for year-round Atlantic–European weather regimes, Quarterly Journal of the Royal Meteorological Society, 149, 2386–2408, https://doi.org/https://doi.org/10.1002/qj.4512, 2023.

Padrón, R. S., Zappa, M., Bernhard, L., and Bogner, K.: Extended-range forecasting of stream water temperature with deep-learning models, Hydrology and Earth System Sciences, 29, 1685–1702, https://doi.org/10.5194/hess-29-1685-2025, 2025.

Papacharalampous, G. and Tyralis, H.: A review of machine learning concepts and methods for addressing challenges in probabilistic hydrological post-processing and forecasting, Frontiers in Water, 4, https://doi.org/10.3389/frwa.2022.961954, 2022.

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

Poncelet, C., Merz, R., Merz, B., Parajka, J., Oudin, L., Andréassian, V., and Perrin, C.: Process-based interpretation of conceptual hydrological model performance using a multinational catchment set, Water Resources Research, 53, 7247–7268, https://doi.org/https://doi.org/10.1002/2016WR019991, 2017.

Quin, A. and Destouni, G.: Large-scale comparison of flow-variability dampening by lakes and wetlands in the landscape, Land Degradation & Development, 29, 3617–3627, https://doi.org/https://doi.org/10.1002/ldr.3101, 2018.

https://doi.org/10.5194/egusphere-2025-3411
Preprint. Discussion started: 1 September 2025

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., and Zappa, M.: Hybrid forecasting: blending climate predictions with AI models, https://doi.org/10.5194/hess-27-1865-2023, 2023.

Stephan, R., Erfurt, M., Terzi, S., Žun, M., Kristan, B., Haslinger, K., and Stahl, K.: An inventory of Alpine drought impact reports to explore past droughts in a mountain region, Natural Hazards and Earth System Sciences, 21, 2485–2501, https://doi.org/10.5194/nhess-21-2485-2021, 2021.

Sutanto, S. J. and Lanen, H. A. J. V.: Catchment memory explains hydrological drought forecast performance, Scientific Reports, 12, 2689, https://doi.org/10.1038/s41598-022-06553-5, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, https://arxiv.org/abs/1706.03762, 2017.

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

Wei, Y., Wang, R., and Feng, P.: Improving Hydrological Modeling with Hybrid Models: A Comparative Study of Different Mechanisms for Coupling Deep Learning Models with Process-based Models, Water Resources Management, https://doi.org/10.1007/s11269-024-03780-5, 2024.

White, C. J., Domeisen, D. I. V., Acharya, N., Adefisan, E. A., Anderson, M. L., Aura, S., Balogun, A. A., Bertram, D., Bluhm, S., Brayshaw, D. J., Browell, J., Büeler, D., Charlton-Perez, A., Chourio, X., Christel, I., Coelho, C. A. S., DeFlorio, M. J., Monache, L. D., Giuseppe, F. D., García-Solórzano, A. M., Gibson, P. B., Goddard, L., Romero, C. G., Graham, R. J., Graham, R. M., Grams, C. M., Halford, A., Huang, W. T. K., Jensen, K., Kilavi, M., Lawal, K. A., Lee, R. W., MacLeod, D., Manrique-Suñén, A., Martins, E. S. P. R., Maxwell, C. J., Merryfield, W. J., Ángel G. Muñoz, Olaniyan, E., Otieno, G., Oyedepo, J. A., Palma, L., Pechlivanidis, I. G., Pons, D., Ralph, F. M., Reis, D. S., Remenyi, T. A., Risbey, J. S., Robertson, D. J. C., Robertson, A. W., Smith, S., Soret, A., Sun, T., Todd, M. C., Tozer, C. R., Vasconcelos, F. C., Vigo, I., Waliser, D. E., Wetterhall, F., and Wilson, R. G.: Advances in the Application and Utility of Subseasonal-to-Seasonal Predictions, Bulletin of the American Meteorological Society, 103, E1448 – E1472, https://doi.org/10.1175/BAMS-D-20-0224.1, 2022.

Wilks, D.: Statistical Methods in the Atmospheric Sciences, ISSN, Elsevier Science, ISBN 9780123850232, https://books.google.de/books?id=fxPiH9Ef9VoC, 2011.

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