

This paper explores a hybrid forecasting framework to predict low flows in the European Alps at a sub-seasonal scale, testing combinations of different input variables and evaluating performance in both reanalysis and forecast modes. The study provides valuable insights into the implementation of hybrid forecasting systems, and the use of EFAS adds practical relevance for future operational applications. This is a promising work which I would recommend acceptance after revisions addressing the following aspects.

Thank you for your positive evaluation. The feedback has helped improve the quality of the manuscript. Detailed responses to the individual comments are provided below in purple.

1. One point regards to the focus on low-flow prediction specifically. Given that the models appear to be trained across all seasons (as I understand), it would be helpful for the authors to explain why the evaluation is restricted to low-flow conditions only.

It is correct that the models are trained using data from all seasons. Low flow conditions in this study are defined by the 15th percentile threshold based on a time-varying climatology, following the same approach as in Chang et al. (2023). Under this definition, low flow conditions are not restricted to a specific season and can occur throughout the year. Training the models on all seasons is therefore necessary to ensure that low flow conditions are represented across different seasons. To help the models capture the seasonality, day-of-year is included as a predictor. We acknowledge that the definition of low flow conditions was not clearly described in the main text of the original manuscript. We will therefore revise the manuscript accordingly.

While we evaluated models in all flow conditions (both high and low), the analyses included in the main text focus on low flow periods because improving low flow prediction is the main objective of this study. Results for overall performance across all flows and for high flow conditions are provided in the Supplementary Material.

2. The methodology is generally well explained, some additional clarifications would be needed. In particular, the encoder-decoder architecture of the TFT model could be described more clearly. Are there new inputs introduced during the decoder phase, or does the decoder rely only on the internal representation generated by the encoder? Moreover, are the precipitation/temperature/weather regime for the 32-day forecast horizon used as inputs too, or the model use only information from the previous 64 days?

We agree that we did not describe the encoder–decoder configuration of the TFT model sufficiently in the original manuscript. Section 4 will be revised to clarify this aspect.

In the TFT implementation used in this study, no new input types are introduced during the decoder phase. The encoder processes past information over a 64-day input window, including different combinations of past streamflow, weather regime information, and EFAS data, depending on the model configuration. If weather

regime and EFAS data are included during the encoder period, they are also provided during the 32-day decoder period. In what we refer to as the forecast mode, only reforecast data (e.g. for weather regime and EFAS data) are used during the decoder period (see Section 4.3.2). Observational streamflow data are never included during the decoder phase, as these values constitute the prediction target. The only meteorological information used in the model is the weather regime classification; precipitation and temperature are not used as direct inputs in this study.

3. Line 102, the use of the 31-day moving window centered on each day is clear, but the 7-day smoothed data needs more clarification. How is the 7-day smoothing interact with the 31-day window?

The 7-day smoothing is applied first to the daily streamflow time series to reduce random day-to-day variability. The 31-day moving window is then applied to the smoothed time series to compute the climatology. We will clarify this sequence in the revised manuscript.

4. Line 110, here all stations are assigned to one of the two categories (nival/pluvial). If I understand correctly, these categories are not used during model training but only for comparison, could the authors confirm this?

The hydrological flow regime classification (nival/pluvial) is included as a static feature to the model (see Section 4.1 and Table 2). We acknowledge that this was not stated explicitly in Section 2.1, and we will therefore revise the section to indicate the use of this classification.

5. Line 123, could the authors confirm that the streamflow is used in unit of mm?

Yes, streamflow is expressed in units of mm (specific runoff). Streamflow is converted from volumetric discharge ( $\text{m}^3/\text{s}$ ) to area-normalized units to account for the large variability in catchment sizes across the study domain. This is described in Section 3.1, and we will clarify the unit explicitly in the revised manuscript.

6. Line 137, Do the weather-regime come with ensemble members and a 6-hour lead time, as implied? How are these used as inputs to the TFT model, do you aggregate them and assign one weather regime per daily time step?

The section referred to in Line 137 introduces the general concept of weather-regime (WR) and their computation. We agree that the description of how WR data are used in the TFT model was not clear. We will update the manuscript to improve this clarity.

For model training, validation, and testing in non-forecast mode, we use WR reanalysis data. These data consist of a single member available at 3-hourly resolution, which is first aggregated to a daily scale. For the WR indices (IWR), one daily value is provided for each WR type (yields 7 in total), whereas for the WR life-cycle (WR\_LC) representation, a single dominant WR type is assigned per day.

In forecast mode, WR reforecast data are used in the decoder period instead of reanalysis data. These reforecast data are provided to the study at daily resolution and include 11 ensemble members, so no temporal aggregation is performed. Rather than using an ensemble mean, we run each model configuration separately for each ensemble member.

We hope this clarification is helpful to better understand the model setup.

7. Line 169, the format of "" needs adjusted.

Thank you for noting this formatting issue. The quotation marks will be corrected and applied consistently throughout the manuscript.

8. Line 250, could the authors explain the reason of splitting the dataset starting from different seasons?

This is a valid and relevant question. The dataset is split starting from different seasons to reduce seasonal bias and to ensure that the validation and testing periods include a sufficient number of dry and low flow conditions, which are the focus of this study. We will revise the section to make this rationale clear to the readers.

9. Line 262, Including country as an input variable seems somewhat unconvincing. While it may provide coarse locational information, national borders are not hydrologically meaningful. Could the authors elaborate on why this variable was chosen and whether more physically relevant spatial descriptors (e.g. hydrological regions, climatic zones) were considered?

We agree that national borders are not hydrologically meaningful and that including country as an input feature may appear unusual. In this study, country is used as a categorical feature to provide coarse contextual information rather than to represent hydrological processes. This choice was motivated by the imbalance in the number of stations across countries. The model already incorporates several physically relevant spatial descriptors as static features (see Section 4.1 and Table 2), including catchment attributes and hydrological context provided through local river and large basin identifiers. We will clarify this rationale in the revised manuscript.

10. Line 286, This is an interesting point. There is debate in the community regarding the inclusion of day of the year as an input feature, some consider it a form of "cheating" because it implicitly encodes seasonal information so the model doesn't need to learn this from the dataset, while others view it as a legitimate predictor. Given that this feature appears to have relatively high importance (Figure 9, higher than WR-LC) in the results, it would be valuable to hear the authors' perspective on this.

We agree that the inclusion of day of the year (DOY) as an input feature is debated in the community, and we appreciate the opportunity to clarify our perspective. In this study, DOY is included as a deterministic temporal feature to explicitly represent

recurring temporal patterns in streamflow. In the submitted manuscript, DOY was briefly described as a feature that helps the model capture seasonal and weekly patterns in the data. We will expand this section to further explain the seasonal signal associated with natural hydrological regimes, as well as potential weekly patterns related to hydropower regulation and operation.

Including DOY provides the model with explicit information on recurring temporal structure, indicating whether a given forecast period typically corresponds to wetter or drier conditions within the annual cycle. The relatively high importance of DOY in Figure 9 reflects the strong influence of such recurring temporal patterns on streamflow. The relatively lower importance of WR-LC suggests that, within the limited set of time-varying predictors considered here, weather regimes provide a more modest incremental contribution.

11. Line 303, could the authors clarify how lead time is handled when transitioning from reanalysis to forecast data during the decoder phase?

Thank you for pointing out that this point requires clarification. In forecast mode, the transition from reanalysis to forecast data occurs at the start of the decoder phase. The encoder always uses reanalysis and/or observational data over the past input window (64 days). During the decoder phase, all time-varying inputs provided to the model (e.g. weather regime and EFAS data, when applicable) are taken from their corresponding reforecast products for the full 32-day forecast horizon. Lead time is therefore handled by the decoder, with no mixing of reanalysis and forecast data at a given lead time. We will update the manuscript accordingly.

12. Figure 5, the lines in this figure are kind of difficult to distinguish. Please consider improving color contrast or line styles.

We agree that Figure 5 is not easy to unpack, particularly due to the overlap of the interquartile range (IQR) of the different experiments. To improve readability, we will remove the shaded IQR area and revise the line styles by introducing distinct line types, so the different experiments are more distinguishable. As the IQR remains important for assessing model performance, we will provide the corresponding version with shaded uncertainty bands in the Supplementary Material.