Stochastic Environmental Research and Risk Assessment (2026) 40:12
https://doi.org/10.1007/s00477-025-03138-2

ORIGINAL PAPER

Improving sub-seasonal hydrological forecasts utilizing the randomness in Deep Learning models

Konrad Bogner[1] · **Ryan S. Padrón**[1,2]

Received: 23 September 2025 / Accepted: 17 November 2025

Abstract
Recent advances in Deep Learning have demonstrated significant potential for improving hydrological forecasts, yet the inherent stochasticity of these models–particularly the impact of random seeds–remains underexplored. Whereas reproducibility often dictates fixing random seeds, the resulting variability of multiple random seeds in model performance across stations and lead times presents an opportunity to enhance probabilistic forecasting through combination techniques. We use the Temporal Fusion Transformer (TFT) model to forecast water temperature and streamflow across 22 gauging stations in Switzerland. The TFT is trained globally, optimizing quantile losses for all stations and lead times jointly. A key aspect of our study is that here we use multiple random seeds when training the TFT model, resulting in an ensemble of seed-specific models to generate predictions. To optimally combine the forecasts of these different models we use two methods: Nonhomogeneous Gaussian Regression (NGR) and Beta-transformed Linear Pool (BLP). These combination methods improve the predictive skill compared to forecasts from the individual seed-specific models. The average Continuous Ranked Probability Score across stations and lead times for water temperature (streamflow) predictions improved from 0.85 °C (0.95 mm/d) for the average of the seed-specific models to 0.73 °C (0.81 mm/d) for the NGR and 0.73 °C (0.79 mm/d) for the BLP. Whereas both methods perform similarly well for predicting water temperature (near-Gaussian), the BLP performs better for streamflow as it is less dependent than the NGR on the underlying distribution of the data. Nevertheless, the skill of the NGR streamflow predictions can match those of the BLP by assuming a heavier-tailed distribution such as the lognormal. Overall, we demonstrate that Deep Learning model ensembles built from random seeds, coupled with principled combination methods, can improve forecast skill across hydrological variables.

Keywords Forecast combination · Quantile regression · Machine learning

1 Introduction

In recent years, Machine Learning–and especially Deep Learning (DL)–methods have been commonly applied with great success in the geophysical domain (Yu and Ma 2021). This is also the case for hydrometeorological forecasting, with DL methods performing as well or mostly better than process-based models at one step ahead predictions and short range forecasts up 10 days (Kratzert et al. 2018, 2019; Xie et al. 2021). The Temporal Fusion Transformer (TFT) is a novel DL model (Lim et al. 2021) that has been used for streamflow predictions (Rasiya Koya and Roy 2024) and sub-seasonal forecasts of water temperature (Padrón et al. 2025). In both cases, the TFT demonstrated its potential to predict hydrological variables by performing at least as well as Long Short-Term Memory models (LSTM).

A factor that is rarely analyzed when using Machine Learning methods is that they are random in nature. One exception is the study of Altarabichi et al. (2024), who provide an overview of randomization techniques in DL and their effects. Randomness is highly important in Machine Learning methods for learning, optimization, generalization

Ryan S. Padrón have contributed equally to this work.

✉ Konrad Bogner
konrad.bogner@wsl.ch

Ryan S. Padrón
ryan.padron@wsl.ch

[1] Hydrological Forecasting, Swiss Federal Research Institute WSL, Zuercherstrasse 111, 8903 Birmensdorf, Switzerland

[2] Institute for Atmospheric and Climate Science, ETH Zürich, Zürich, Switzerland

Published online: 08 January 2026

and the initialization of weights. The inherent unpredictability in random weight initialization helps break symmetries within the network. This property is crucial because it allows the network to explore different feature representations of the data during training, making it less likely to converge to suboptimal solutions (Narkhede et al. 2022). To prevent overfitting in Deep Neural Networks, regularization techniques like dropout come into play. Dropout randomly deactivates a fraction of neurons during training, serving to diversify the network's internal representations. This controlled randomness prevents overreliance on specific features, thereby enhancing model generalization (Srivastava et al. 2014). Determining optimal hyperparameters for Machine Learning models can be an arduous endeavor. Techniques such as randomized search and Bayesian optimization employ random numbers to sample hyperparameters from predefined distributions (Bergstra and Bengio 2012; Wu et al. 2019). This randomization efficiently explores the hyperparameter space, leading to improved model configurations. Stochastic optimization techniques employ random numbers to introduce variability during the gradient descent process (Bottou 2012). Instead of a fixed step size for each update, methods like mini-batch gradient descent randomly sample subsets of the training data. This stochasticity helps escape local minima and can expedite convergence (Goodfellow et al. 2016).

The main objective of this study is to analyze how the randomness of DL models influences forecast skill, and to assess the potential of different forecast combination methods to improve the predictive skill. To do so, we train a TFT model with n different random seeds, following the post-processing methods developed to calibrate hydrometeorological probabilistic forecasts (Li et al. 2017) and to derive predictive uncertainties (Tyralis and Papacharalampous 2024). These n seed-specific models generate slightly different forecasts that can then be optimally combined by applying appropriate weights derived from a previous set of forecasts. Here we estimate these weights using a Non-homogeneous Gaussian Regression (NGR, Gneiting et al. 2005) and a Beta-transformed Linear Pool (BLP, Ranjan and Gneiting 2010), as it was done by Bogner et al. (2017). We do not use ensemble learning methods based on Machine Learning (Mienye and Sun 2022) to determine the optimal combination of the n seed-specific forecasts because this would not solve the problem of randomization, and we would end up running in circles.

2 Methods

2.1 Model setup

We train a TFT model with data from 22 gauging stations across Switzerland (Fig. 1) to predict either daily maximum water temperature or daily average streamflow for the next 32 days. This is a global model that best fits the data from all stations and lead times. It is split into an encoder and a decoder. The encoder comprises past known information, including observations of the target variable to benefit from auto-regressive properties. On the other hand, the decoder comprises data known in the future, such as meteorological forecasts. One of the main advantages of the TFT is the combination of LSTM models for temporal processing of past and future input with multi-head attention layers for capturing long-term temporal dependencies. For details about the TFT refer to Lim et al. (2021).

The input and output features of the TFT model are past continuous and categorical predictors until forecast initialization time t ($x^{(p)}_{[:t]}$), known future predictors until the

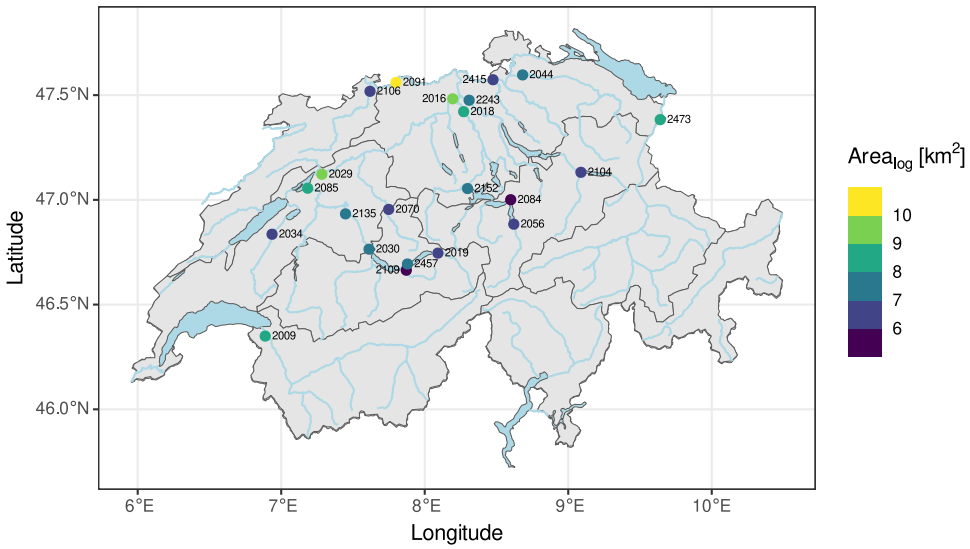


Fig. 1 Locations of the Swiss gauging stations. The color of the circles of the 22 stations selected for this study indicate the logarithm of the upstream area

forecast horizon H ($x^{(f)}_{[:t+H]}$), static input features $x^{(s)}$, and past observations of the target variable $y_{[:t]}$. These features are put together to constitute a quantile regression model with the conditional probability:

$$\mathbb{P}(y_{[t+1:t+H]}|y_{[:t]}, x^{(p)}_{[:t]}, x^{(f)}_{[:t+H]}, x^{(s)}) \quad (1)$$

The model parameters are estimated by minimizing the average quantile loss over 99 quantiles (0.01, 0.02, ...,0.99), all stations, and all lead times. This is equivalent to having the Continuous Ranked Probability Score (CRPS, Gneiting and Raftery (2007)) as the loss function, given that the overall average quantile loss corresponds to the CRPS divided by 2 (Bröcker 2012; van der Meer et al. 2024). The tuning of the TFT hyperparameters and the setup to train the model is done as described in Padrón et al. (2025). We train the TFT models using daily data from 2012 to 2022, with the year 2022 used for validation during the training process. We test the predictive skill of the models in 2023 and 2024, and assess the potential of different forecast combination methods to improve it.

Table 1 shows the variables we use as features in the TFT models to predict daily maximum water temperature and daily average streamflow with a forecast horizon H of 32 days. Here we define an encoder length of 64 days to incorporate information prior to the start of the forecast at time t. In addition, note that we also use wavelet-transformed air temperature data as an input feature. By doing this we incorporate valuable information about long-term averages and trends in air temperature. As in Bogner and Pappenberger (2011), we apply a non-decimated wavelet transform given by the Haar à trous algorithm (Dutilleux 1987) to decompose air temperature observations and forecasts. This wavelet transformation simply pairs up input values, stores the difference, and passes the sum recursively to provide the next scale (for a review see Stankovic and Falkowski 2003). The à trous wavelet transform is based on successive convolutions with the discrete low-pass filter $h(l)$ given by Eq. 2, where the finest scale is the original signal $s_0(t) = y(t)$ at time t.

$$s_{i+1}(t) = \sum_{l=-\infty}^{\infty} h(l) s_i(t + 2^i l) \quad (2)$$

From the sequence of smoothings of the signal, the differences are used to calculate the wavelet coefficients d_i (Eq. 3), which capture the details of the signal.

$$d_{i+1}(t) = s_i(t) - s_{i+1}(t) \quad (3)$$

The Haar à trous wavelet transform is a simplified version, where the low-pass filter h is given by (1/2, 1/2) and Eq. 2 reduces to

Table 1 Input features used by the TFT models to predict water temperature and streamflow

Input feature	Data type/Unit	Source
Known timeseries (past and future)		
Air temperature (AirTemp)	Extracted from gridded data [2 km]/°C	MeteoSwiss
Precipitation	Extracted from gridded data [2 km]/*mm/d*	MeteoSwiss
Sunshine duration	Extracted from gridded data [2 km]/%	MeteoSwiss
AirTemp wavelet transformed details (Level 6)	°C	calculated
AirTemp wavelet transformed smoothed (Level 6)	°C	calculated
Day index	Periodic function of the day of the year	calculated
Week index	Periodic function of the week of the year	calculated
Static features		
Station number	Group ID	FOEN[1]
Catchment area	km^2	FOEN[1]
Mean catchment elevation	*m*	FOEN[1]
Glacierized area	%	FOEN[1]
Coord_X	*m*	FOEN[1]
Coord_Y	*m*	FOEN[1]
Station elevation	*m*	FOEN[1]
Target center	Long-term average of the target variable °C, *mm/d*	calculated
Target scale	Long-term standard deviation of the target variable °C, *mm/d*	calculated
Past observations (Target)		
Water temperature / Streamflow	Point measurements [°C, *mm/d*]	FOEN[1]

[1]FOEN: Federal office for the environment (Switzerland)

$$s_{i+1}(t) = \frac{1}{2}(s_i(t) + s_i(t - 2^i)). \quad (4)$$

More details about the Haar à trous wavelet transform and its application for forecast purposes can be found in Benaouda et al. (2006).

In correspondence with our encoder length of 2^6, we chose a decomposition level of 6. To avoid multicollinearity problems only d_6 and s_6 are included as input features, which can be interpreted as the long-term average and the trend in the air temperature over the past 64 days. In Fig. 10 an example of a wavelet decomposition is shown, with the d_6 and s_6 levels shown as triangles. Figure 11 shows how the CRPS improves when predicting water temperature with a model that includes wavelet-transformed air temperature data compared to a model without these input features.

2.2 Combination of ensemble forecasts

Operationally, there are 51 ensemble members available from sub-seasonal meteorological forecasts to use as input to our TFT models for predicting water temperature and streamflow. Therefore, each seed-specific TFT model can generate 51 different probabilistic forecasts. Here we first obtain a single probabilistic forecast for each TFT model by combining their 51 estimates, so we can then optimally combine the resulting individual forecasts of TFT models trained with different random seeds to improve the predictive skill. The ensemble members of the meteorological forecasts have no physical correspondence from one forecast to the next (Molteni et al. 1996; Bröcker and Kantz 2011), which is why the combination of their corresponding 51 hydrological forecasts cannot be optimized. The ensemble meteorological forecasts that we use are provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and pre-processed by the Swiss Federal Office for Meteorology and Climatology (Monhart et al. 2018; Bogner et al. 2022; Chang et al. 2023). For each of the 51 ensemble members the TFT model generates a probabilistic forecast described by estimates of 99 quantiles (Fig. 2).

First, we fit a normal distribution to each set of quantile predictions by minimizing the sum of squared errors between the predicted quantile values and the theoretical quantiles of the distribution. For a given quantile level p (where $p \in (0, 1)$), the theoretical quantile of the normal distribution is given by the inverse cumulative distribution function (CDF) $\mu + \sigma \cdot \Phi^{-1}(p)$. This optimization was initialized using the empirical mean and standard deviation of the quantile values. In addition, given that streamflow data are not typically normally distributed, we also fit a lognormal distribution. In this case, the theoretical quantiles are given by $\exp(\mu + \sigma \cdot \Phi^{-1}(p))$, with μ initialized to the log-median and σ approximated from log-transformed data. Optimization is performed via L-BFGS-B (Zhu et al. 1997), with bounds ensuring positivity of scale parameters. Figure 2c illustrates that it is appropriate to fit a normal distribution to the quantile predictions of water temperature, whereas it is better to fit a lognormal distribution to the quantile predictions of streamflow as shown in Fig. 2d

Second, we combine the resulting 51 probability distributions (one for each ensemble member of the meteorological forecasts) into one overall distribution. For normal distributions with parameters (μ_i, σ_i^2), the parameters of the overall distribution are:

$$\mu_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N} \mu_i \quad (5)$$

$$\sigma^2_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N} \left(\sigma_i^2 + \mu_i^2\right) - \mu^2_{\text{total}} \quad (6)$$

To combine an ensemble of lognormal distributions, their underlying normal parameters μ_i and σ_i^2 can be used, given that a lognormal distribution is defined as:

$$X_i \sim \mathcal{LN}(\mu_i, \sigma_i^2) \iff \ln(X_i) \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad (7)$$

Once the normal parameters of $ln(X_i)$ for each of the 51 distributions are estimated, they are then combined into a single overall distribution using Eqs. 5 and 6. The overall lognormal distribution is given then by:

$$X_{\text{total}} \sim \mathcal{LN}(\mu_{\text{total}}, \sigma^2_{\text{total}}) \quad (8)$$

The expected value and the variance of a lognormal distribution are defined as:

$$\mathbb{E}[X_i] = e^{\mu_i + \sigma_i^2/2} \quad (9)$$

$$\text{Var}(X_i) = \left(e^{\sigma_i^2} - 1\right) e^{2\mu_i + \sigma_i^2} \quad (10)$$

In Fig. 3 we show the probability density function of the single overall distribution derived from the 51 probability distributions for the example forecast of Fig. 2 at one station and one lead time. We use normal distributions for water temperature predictions and lognormal distributions for streamflow predictions. For comparison, we use the kernel density estimation (KDE, Silverman (1986)) method to draw 51×500 samples from the estimated normal and lognormal distributions (one parameter set for each of the 51 ensemble members of the meteorological forecasts). Both options show similar results for water temperature. For streamflow,

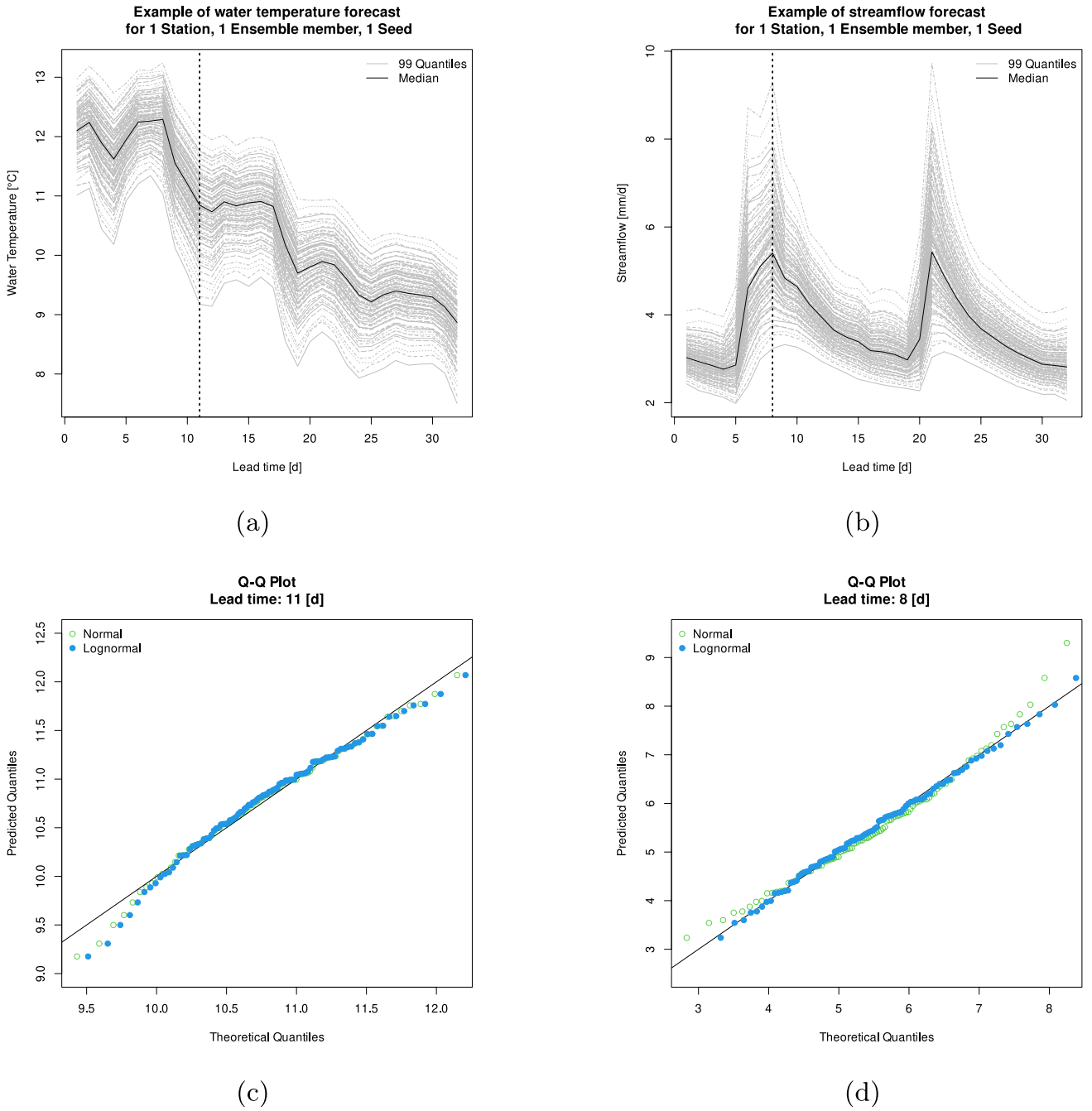


Fig. 2 Example probabilistic forecasts of water temperature (**a**) and streamflow (**b**) at one station, and for one ensemble member. Grey lines indicate the predictions for quantiles 0.01, 0.02,…, 0.99, whereas the median (q50) is shown in black. The dashed vertical lines indicate the lead times at which Q-Q plots are drawn for water temperature (**c**) and streamflow (**d**). The Q-Q plots compare the predicted quantiles with the theoretical quantiles from a fitted normal and lognormal distribution. The diagonal 1:1 line is shown in black

the KDE estimate of the probability density function shows a sharper shape, which could cause difficulties in estimating probabilities for extreme events. It is important to note that merging the forecasts derived from the meteorological ensemble into a single unimodal normal or lognormal distribution could smooth out existing multi-modal or heavily skewed structures. While this merging step is necessary for the purpose of our study, it is still advisable to also provide end-users with the predictions based on individual meteorological ensemble members.

Finally, we have one probabilistic forecast per seed-specific TFT model after combining the TFT forecasts generated when using each of the 51 ensemble members of the meteorological forecasts as input (Fig. 4). We can then

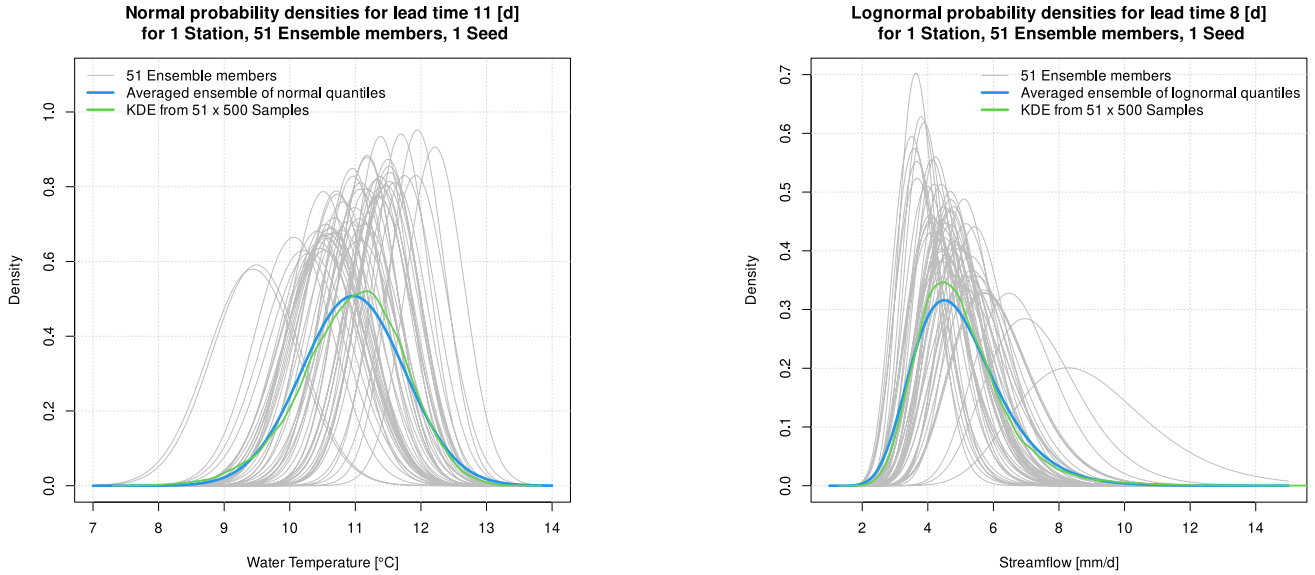


Fig. 3 Combination of probability densities from 51 ensemble members to an average overall distribution at one station and one lead time corresponding to the example forecast of Fig. 2. Probability densities of water temperature predictions are shown in (**a**) and of streamflow predictions are shown in (**b**). For comparison, the result of the Kernel Density Estimation (KDE) from 51 × 500 samples drawn from the estimated normal and lognormal distributions are shown in green

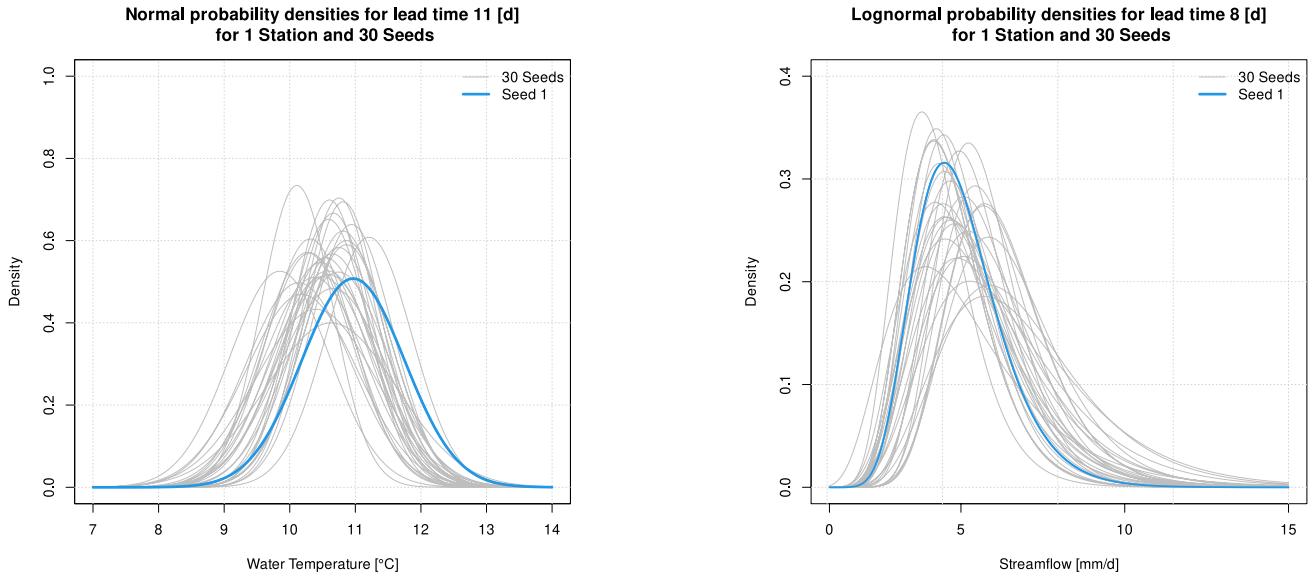


Fig. 4 Example probabilistic forecasts of water temperature (**a**) and streamflow (**b**) from 30 different seed-specific TFT models. These forecasts correspond to the one station and lead time shown in Figs. 2 and 3. The blue line corresponds to the probability densities shown in Fig. 3 for one seed-specific TFT model after combining the probability densities of the 51 ensemble members of the meteorological forecasts

determine how to combine these resulting forecasts to maximize the predictive skill.

2.3 Optimal combination of forecasts from seed-specific models

Here we use Nonhomogeneous Gaussian Regression (NGR, Gneiting et al. (2005)) and Beta-transformed Linear Pool (BLP, Ranjan and Gneiting (2010)) to optimally combine the forecasts from seed-specific TFT models. At each forecast initialization date, we train the NGR and BLP by using the TFT predictions of the previous 35 forecast initialization dates for each station and lead time separately. These 35 forecast dates span 122 days into the past, given that the meteorological forecasts are available twice per week. Typically, shorter training periods between 20 and 60 days are used for estimating the combination parameters when forecasts are initialized every day (e.g. Gneiting et al. (2005); Baran and Lerch (2015)). A shorter training period could adapt better to seasonal and environmental changes, whereas longer periods with more forecast initialization dates help stabilize the variability of the parameters. With our choice of 35 forecast initialization dates (122 days) we aim to adapt to seasonal changes, while still counting with enough data for stable parameter estimates. In Fig. 12 we evaluate the sensitivity of our results to the chosen number of past forecast initialization dates used when training the combination methods.

The NGR and BLP are trained to minimize the CRPS. For a normal distribution (μ, σ^2), as we use for predicting water temperature, the CRPS is given by:

$$CRPS_{normal} = \sigma \begin{bmatrix} \frac{y_{obs}-\mu}{\sigma}\left(2\Phi\left(\frac{y_{obs}-\mu}{\sigma}\right)-1\right) \\ +2\phi\left(\frac{y_{obs}-\mu}{\sigma}\right)-\frac{1}{\sqrt{\pi}} \end{bmatrix} \tag{11}$$

where ϕ is the probability density function of the standard normal distribution. On the other hand, for the lognormal streamflow predictions the CRPS is given by:

$$\begin{aligned} \mathrm{CRPS}_{\mathcal{LN}}(F, y_{\mathrm{obs}}) =& y_{\mathrm{obs}}\left[2\Phi\left(\frac{\ln y_{\mathrm{obs}}-\mu}{\sigma}\right)-1\right] \\ &- 2\exp\left(\mu+\frac{\sigma^2}{2}\right)\Phi\left(\frac{\ln y_{\mathrm{obs}}-\mu}{\sigma}-\sigma\right) \end{aligned} \tag{12}$$

where F is the cumulative distribution function of the lognormal distribution with parameters μ (mean of log-transformed variable) and σ (standard deviation of log-transformed variable), y_{obs} is the observed value, Φ is the cumulative distribution function of the standard normal distribution.

2.3.1 Nonhomogeneous Gaussian Regression (NGR)

The NGR method, also known as Ensemble Model Output Statistics (EMOS) is based on multiple linear regression for linear variables. Let y denote again the variable of interest (e.g. water temperature) and let $k_1, \ldots, k_M$ be the corresponding forecast of the M ensemble members or model seeds in our case. If $\mathcal{N}(\mu, \sigma^2)$ denotes a normal density with mean μ and variance σ^2, the NGR predictive distribution is given by

$$\begin{aligned} y|k_1, \ldots, k_M &\sim \mathcal{N}(a_0 + a_1 k_1 + \cdots + a_M k_M, b_0 + b_1 s^2), \\ \text{where} \quad s^2 &= \frac{1}{M}\sum_{m=1}^{M}\left(k_m - \frac{1}{M}\sum_{m=1}^{M} k_m\right)^2 \end{aligned} \tag{13}$$

with regression coefficients $a_0, a_1 + \cdots + a_M$ for estimating the predictive mean μ and the nonnegative coefficients b_0 and b_1 for estimating the predictive variance σ^2 by minimizing $\mathrm{CRPS}_{\mathrm{normal}}$ (Eq. 11). For a lognormal distribution, NGR optimizes the log-space mean (μ) and log-space standard deviation (σ) of the ensemble forecasts of the model seeds to minimize the $\mathrm{CRPS}_{\mathcal{LN}}$ (Eq. 12).

The probability density function of the lognormal distribution is given by:

$$\mathcal{LN}(y \mid \mu, \sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right)$$

As with the case of normal distribution this model is parameterized such that the mean value of the lognormal distribution is a linear function:

$$\mu = a_0 + a_1 k_1 + \cdots + a_M k_M,$$

and the bias-corrected log-variance is given by:

$$\begin{aligned} \sigma^2 &= b_0 + b_1 \cdot \mathrm{Var}(\ln k_{\mathrm{ens}}), \\ \mathrm{Var}(\ln k_{\mathrm{ens}}) &= \frac{1}{M}\sum_{i=1}^{M}(\ln k_i - \frac{1}{M}\sum_{i=1}^{M}(\ln k_i))^2 \end{aligned}$$

2.3.2 Beta-transformed Linear Pool (BLP)

Ranjan and Gneiting (2010) state that any non-trivially weighted average of distinct probability forecasts will be uncalibrated and lack sharpness, even when the individual forecasts have been calibrated. Hence, they suggest a composite of the traditional linear pool with a beta transform. The forecast combination method introduced by Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) considers

the Beta-transformed Linear Pool (BLP) for a set of predictive cumulative density functions $F_1, \ldots, F_M$ as

$$F(y) = B_{\alpha,\beta}\left(\sum_{m=1}^{M} \omega_m F_m(y)\right) \tag{14}$$

for $y \in \mathbb{R}$, where $B_{\alpha,\beta}$ denotes the cumulative density function of the standard Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ and $\omega_1, \ldots, \omega_M$ being nonnegative weights that sum to 1. The BLP density forecast for the component densities $f_i, \ldots, f_M$ then is

$$f(y) = \left(\sum_{m=1}^{M} \omega_m f_m(y)\right) b_{\alpha,\beta}\left(\sum_{m=1}^{M} \omega_m F_m(y)\right) \tag{15}$$

with parameters $\alpha > 0$ and $\beta > 0$ of the Beta density function $b_{\alpha,\beta}$. The advantage of the Beta distribution is its flexibility. By adjusting α and β, the BLP can effectively stretch, compress, and shift the cumulative density function of the linear pool to correct for biases in its location, spread, and skew. For $\alpha = \beta = 1$ the Beta distribution is uniform, so no calibration is applied and the BLP corresponds to the traditional linear opinion pool. For $\alpha < 1, \beta > 1$ the transformation will stretch the lower tail and compress the upper tail of the linear pool, correcting for overconfidence in the lower quantiles. If $\alpha > 1, \beta < 1$ the opposite effect happens, namely a correction for overconfidence in the upper quantiles. If $\alpha > 1, \beta > 1$ the transformation compresses both tails, making the predictive distribution sharper to correct for underconfidence.

Thus $B_{\alpha,\beta}$ can be interpreted as a parametric calibration function for combining $F_1, \ldots, F_M$ with mixture weights $\omega \in \Delta_M$, which assign relative importance to the individual predictive distributions. The parameters $\alpha > 0$ and $\beta > 0$ and the weights $\omega_1, \ldots, \omega_M$ are estimated by minimizing the CRPS (Eq. 11, 12). Berrisch and Ziel (2023, 2024) provide further details about the BLP forecast combination method.

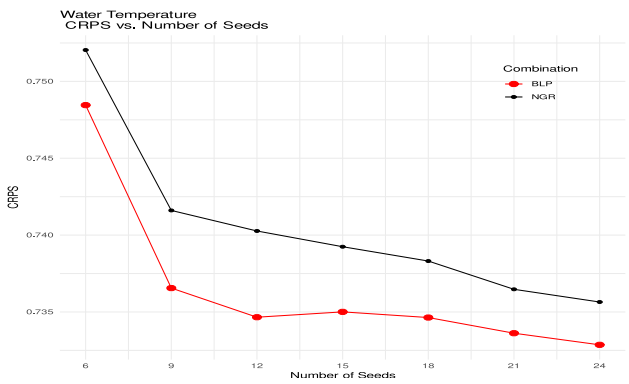


Fig. 5 CRPS as a function of the number of seed-specific TFT models whose water temperature predictions are combined. Results are shown for the NGR and BLP forecast combination methods. The CRPS is averaged across all 22 stations, 32 lead times, and 169 forecasts available during the testing period 2023–2024

3 Results

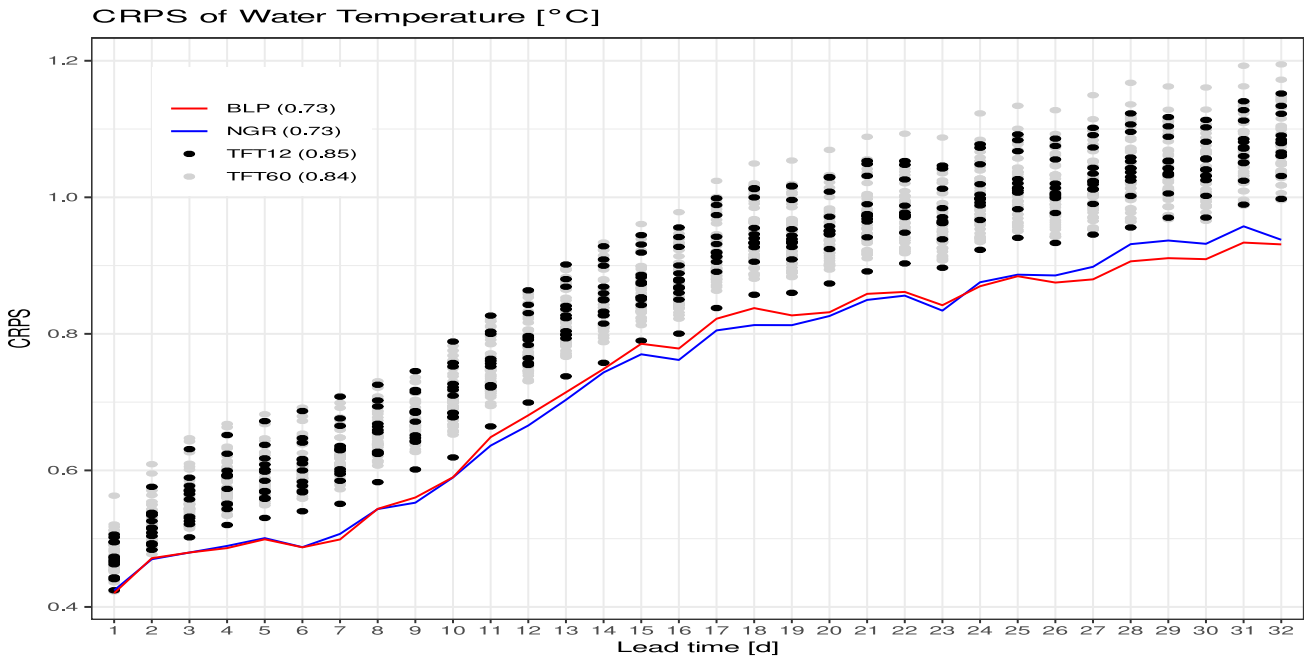
3.1 Optimal number of seed-specific models

We first assess how the average CRPS changes as we increase the number of TFT models with different random seeds whose predictions are optimally combined with the NGR and BLP methods (Fig. 5). Here we train models with 60 different random seeds, and generate predictions for approximately 200 forecast initialization dates over the testing period 2023–2024. The average CRPS across all stations, lead times and forecast initialization dates clearly improves as the water temperature predictions from more seed-specific models are combined. Note, however, that the additional CRPS reduction is small when combining more than 12 seed-specific models. Therefore, we use 12 seed-specific models from here on to balance the gain in predictive skill with the computational resources needed to generate the optimally-combined forecasts. In addition, we test how the resulting average CRPS values depend on which 12 seed-specific models are chosen. To do this we take 30 random samples with replacement of subsets of 12 seed-specific models out of 60 models available, and proceed to compute the average CRPS. The average coefficient of variation is less than 1% for both the NGR and BLP.

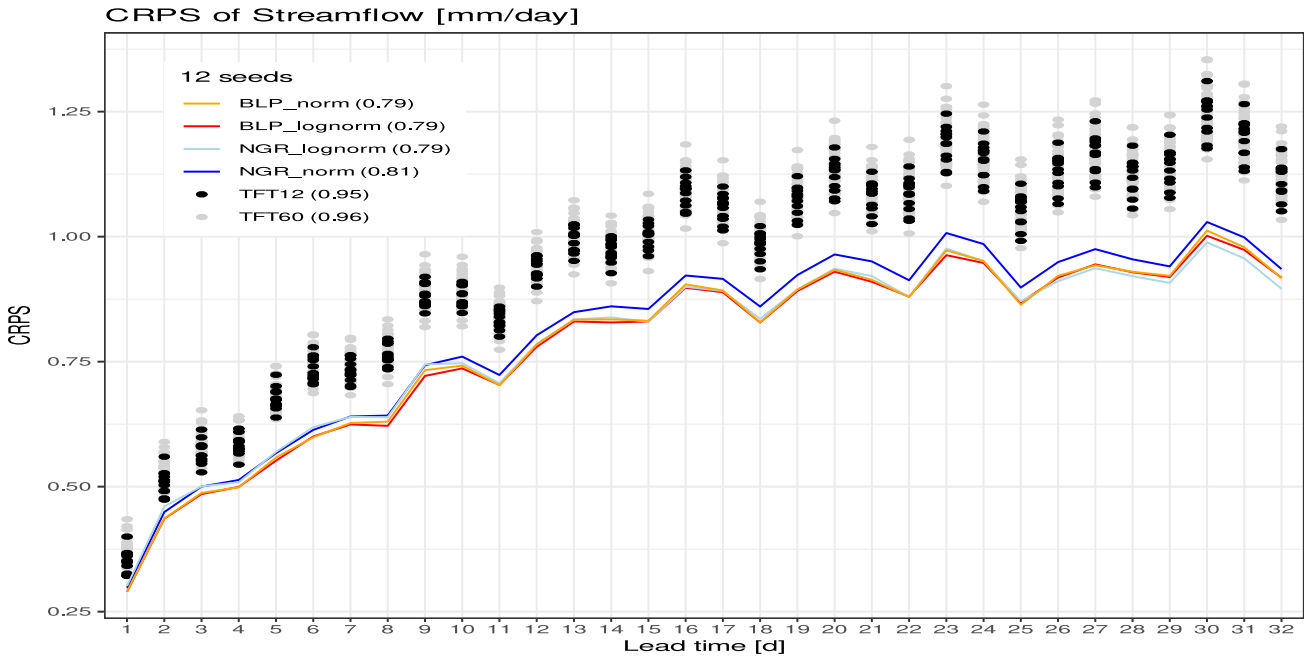
3.2 Improvement of predictive skill with forecast combinations

We find that combining forecasts from seed-specific TFT models with the NGR and BLP methods improves the predictive skill compared to the forecasts from individual TFT models (Fig. 6). This is the case across all lead times for predictions of both water temperature and streamflow. For water temperature, the overall average CRPS of 0.85 °C from 12 seed-specific TFT models is reduced to 0.73 °C after applying either the NGR or BLP optimal combination. Correspondingly, the overall average CRPS improves from 0.95 mm/d to 0.79 mm/d for streamflow. Note that if we had assumed an underlying normal distribution for the streamflow probabilistic forecasts, as opposed to a lognormal distribution, the CRPS of the NGR would have been slightly worse (0.81 mm/d). Besides this higher sensitivity of the NGR to the underlying distribution of the probabilistic forecasts, the BLP and NGR methods lead to similar results.

In Fig. 7 we quantify how much the CRPS improves per lead time and per station when applying the BLP



(a)



(b)

Fig. 6 CRPS per lead time averaged over all 22 stations and 169 forecasts available during the testing period 2023–2024. CRPS results are shown for all 60 seed-specific TFT models, for the subset of 12 seed-specific TFT models used to combine the forecasts, and for the NGR and BLP forecast combination methods with either a normal or lognormal assumed distribution for the probabilistic forecasts. Results for water temperature are shown in (**a**) and for streamflow in (**b**). The numbers in parentheses correspond to the overall average CRPS also across all lead times

combination method, as opposed to taking the average of the 12 seed-specific TFT models. As the CRPS increases with lead time due to the uncertainty of the meteorological forecasts, so does the improvement in predictive skill from using the BLP combination method. At the final lead time of 32 days, the BLP improves the CRPS by 0.18 °C and 0.2 mm/d. We also note that the improvement in predictive skill varies across stations. For water temperature, station 2415 has the lowest average CRPS reduction of 0.03 °C, whereas station 2084 has the highest average CRPS reduction of 0.31 °C. For streamflow, the stations 2034, 2044, 2070, 2106, 2415 show very little to no improvements, whereas the highest improvement of 0.77 mm/d occurs at station 2019. In general, the CRPS improvement would tend to be higher at stations with higher observed day-to-day variability in water temperature and streamflow.

3.3 Forecast examples

The forecast combination quality is mainly determined by the diversity between forecasts, given that there is close to no benefit in combining very similar forecasts (Berrisch and Ziel 2024). In Fig. 8, we show an example forecast with variability across the seed-specific TFT models for 2 stations: (i) station 2016 located at the Aare river in Brugg, with an upstream area of ∼11,680 km^2, and (ii) station 2091 located at the Rhine in Rheinfelden, with a catchment area of ∼34,500 km^2 (see Fig. 1). During the time of this forecast initialized on 27 July 2023 there was high water temperature and low streamflow. The BLP predictions match the observations better over most lead times, and particularly in the second half of the forecast horizon. We additionally show the probability density forecasts of the seed-specific TFT models and of the BLP at the end of the forecast horizon (lead time of 32 days). In this case, the BLP clearly predicts water temperature better than the individual TFT models, but it does not for streamflow. The BLP streamflow predictions remain relatively low for the second half of the forecast horizon, thus capturing the observed low flow conditions better, despite missing the observed peak in the final days.

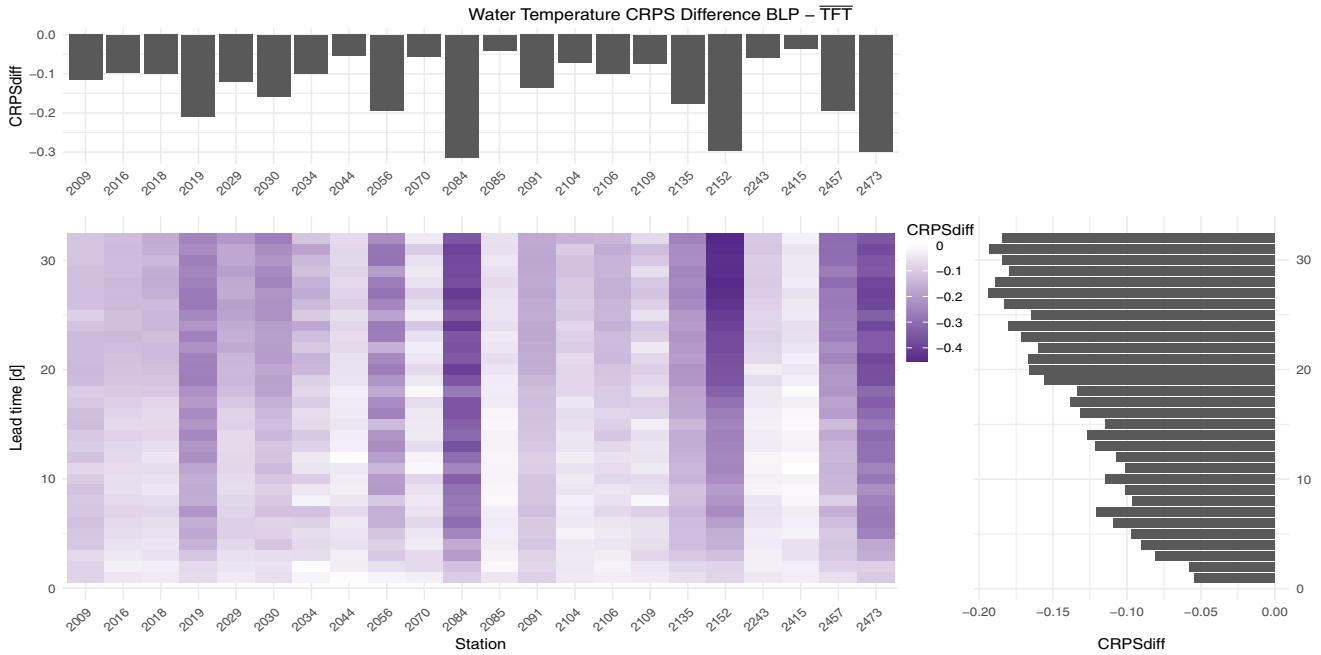
Finally, we highlight the predictive skill of our BLP estimates by comparing it to streamflow predictions from a multi model ensemble of process-based hydrological models (Schirmer et al. 2025). Figure 9 shows an example forecast from 20 November 2023 with observed high streamflow conditions in the weeks that followed at station 2070 (Emme-Emmenmatt). There are not many other stations with available predictions for comparison. We find that for this chosen forecast, the TFT and process-based hydrological models have a similar average CRPS of 4.77 mm/d and 4.64 mm/d, respectively. Meanwhile, the BLP estimate improves the average CRPS down to a value of 3.78 mm/d, which is even better than the CRPS of 4.32 mm/d from the best performing process-based hydrological model.

4 Discussion

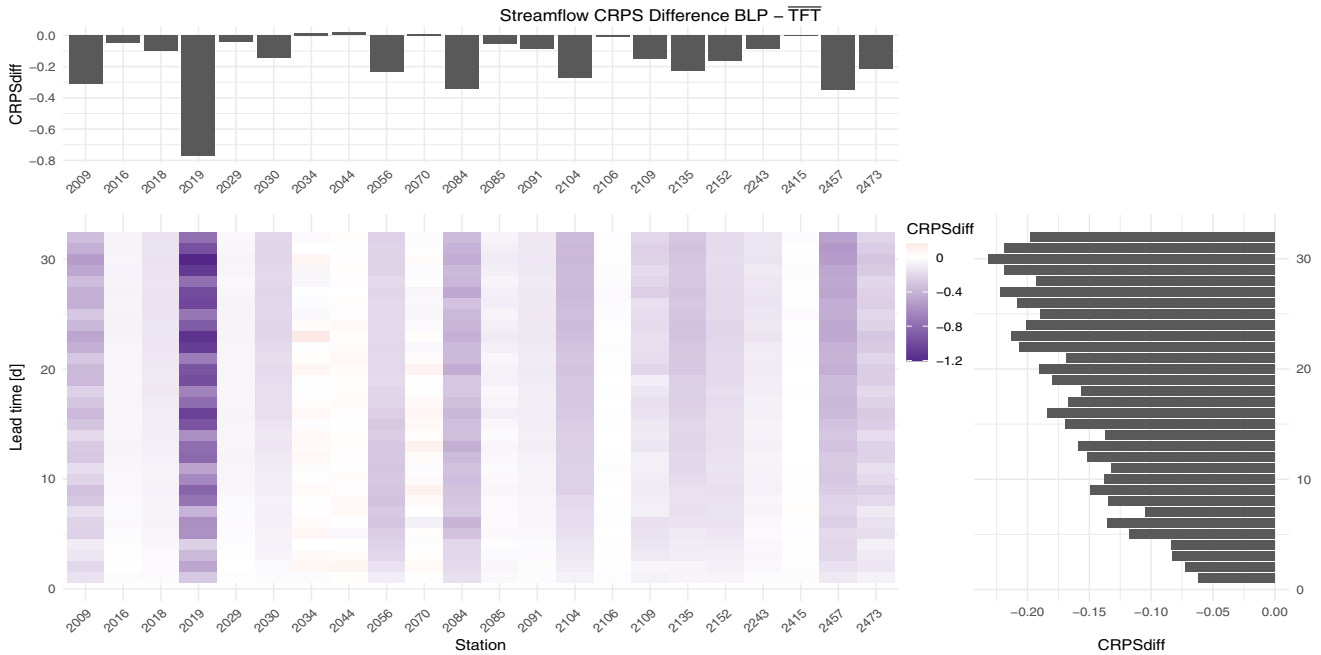
The variability induced by random seeds results in TFT models having performance differences across lead times and stations. For lead times beyond 10 days that are increasingly dominated by meteorological uncertainty, the diversity among seeds provides an opportunity to improve probabilistic forecasts through combination methods (NGR and BLP). Notably, the gains in predictive skill plateaued after combining ∼ 12 seed-specific models (Fig. 5), indicating diminishing returns with larger ensembles. This suggests that operational frameworks can achieve robust performance with a tractable number of model realizations, reducing computational overhead.

For water temperature, a variable with a near normal distribution, BLP and NGR performed similarly well, as both methods effectively calibrated the ensemble spread (Fig. 6). However, for runoff, which follows a heavier-tailed distribution, BLP demonstrated superior robustness when assuming an underlying normal distribution for the probabilistic forecasts of the seed-specific TFT models. While both BLP and NGR are parametric methods, NGR requires a strong assumption about the family of the predictive distribution (e.g., normal or lognormal). In contrast, BLP does not assume a specific distributional form for the combined forecast. Instead, it uses the parameters of the Beta distribution as a flexible calibration function to correct the spread and skew of the linear pool of forecasts. This distribution-agnostic nature makes BLP more robust for variables like runoff, where the true error distribution may not conform perfectly to standard parametric families. These findings echo earlier work advocating for BLP's flexibility in combining non-normal forecasts (Ranjan and Gneiting 2010; Gneiting and Ranjan 2013).

The success of the combination methods hinges on ensemble diversity as illustrated in Fig. 8. It is important to note that the degree of variability in the ensemble can change substantially between different forecast initialization dates, and that consequently this would influence the level of improvement provided by the combination methods. Note also that while the BLP improves the CRPS in our forecast example, substantial differences remain between observed and predicted values. Both rivers shown in this example, the Aare and the Rhein, are highly regulated and influenced by natural lakes and hydro power reservoirs. In addition, hydrological predictions depend strongly on the quality of the meteorological forecasts, which substantially



(a)



(b)

Fig. 7 CRPS differences between the BLP estimate and the average of the 12 seed-specific TFT models. The differences for each station and lead time are shown after averaging over all 169 forecasts available during the testing period 2023–2024. Results correspond to **a** water temperature predictions with an underlying normal distribution, and **b** streamflow predictions with an underlying lognormal distribution

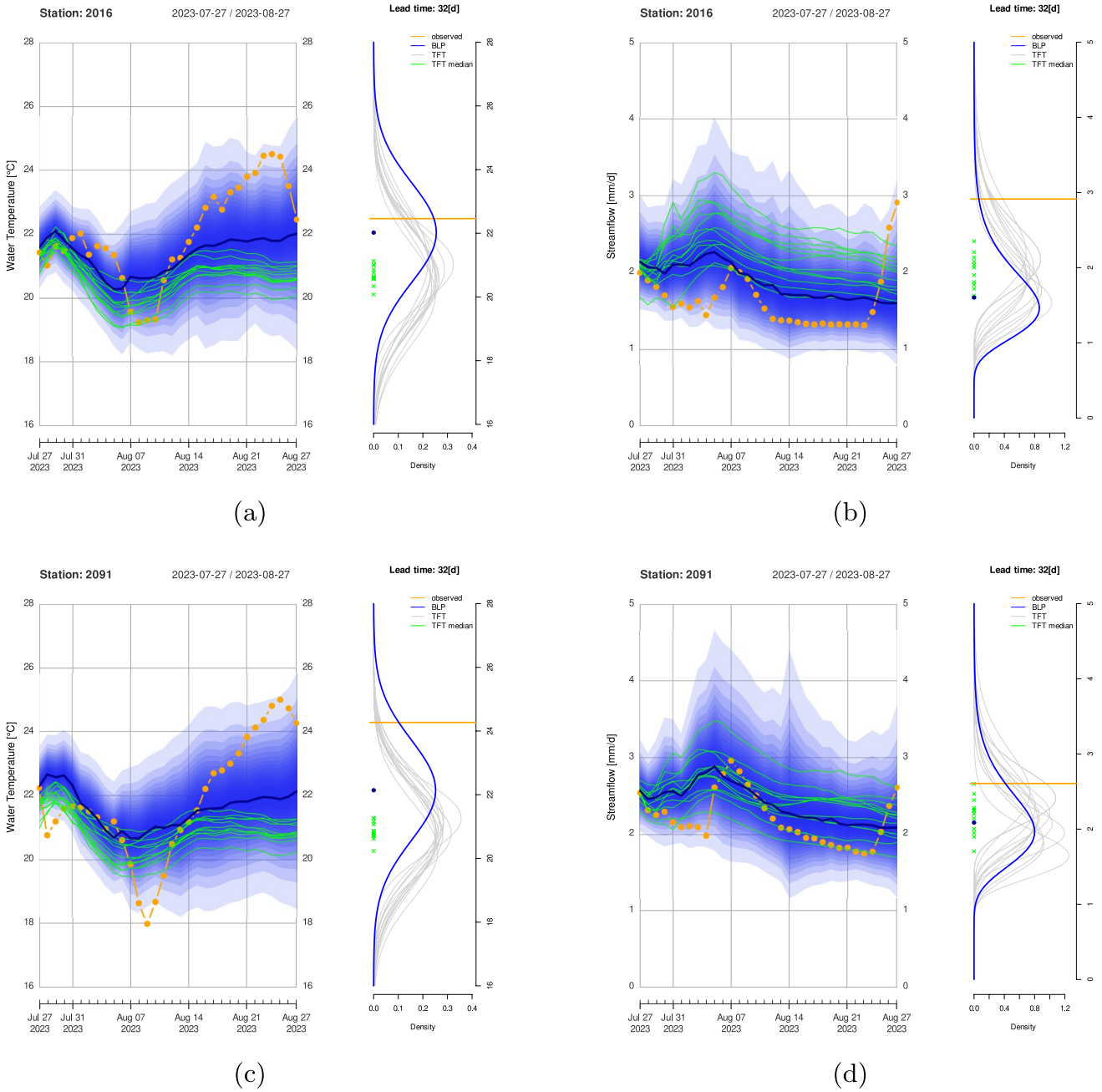
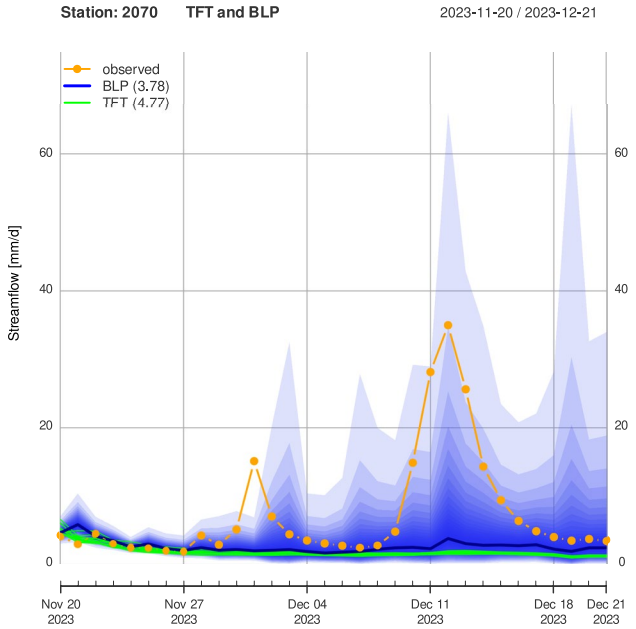


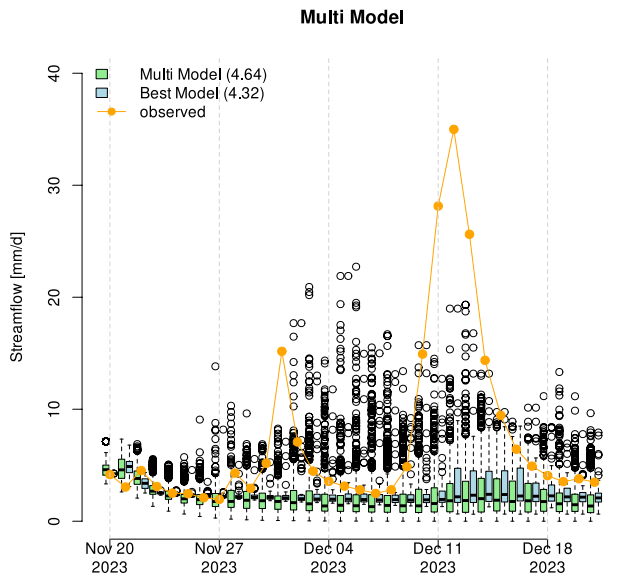
Fig. 8 Example forecast from 27 July 2023 for water temperature (**a**, **c**) and streamflow (**b**, **d**). Predictions are shown for station 2016 (Aare-Brugg) (**a**, **b**) and station 2091 (Rhein-Rheinfelden) (**c**, **d**). Observed values are shown in orange, BLP estimates in blue (with shading indicating different probability densities), best estimates of the 12 seed-specific TFT models in green, and probability densities of these TFT models in grey

decreases for lead times longer than 10 days. Therefore, it would be illusory to expect that the combination of seed-specific TFT models could predict flood peaks one month ahead of time. However, for medium to low flow conditions, the forecast combination improves the predictive skill providing good estimates of the sub seasonal trend. This is an important contribution to be better prepared for the hazard of simultaneous hot and dry conditions in rivers.

We achieve a good trade-off between seasonal adaptability and parameter estimation robustness with our choice to use information from the previous 35 forecast initialization dates (122 days) for the forecast combination methods (Fig. B1). While shorter training periods (e.g., 20–60 days) are common (Gneiting et al. 2005; Baran and Lerch 2015), our extended window ensured sufficient data for reliable weight estimation without sacrificing responsiveness to



(a)



(b)

Fig. 9 Comparison of streamflow forecasts from (**a**) our BLP combination of seed-specific TFT models against (**b**) a multi model ensemble of process-based hydrological models (Schirmer et al. 2025). Shown is an example forecast from 20 November 2023 at station 2070 (Emme-Emmenmatt) with a catchment area of $\sim 440 km^2$. The numbers in parentheses correspond to the average CRPS across all lead times. In panel (**b**) the spread in the predictions arises from 11 different process-based hydrological models times 51 ensemble members of the meteorological forecasts used as input to the models

environmental changes. This is an important aspect in favor of the suitability of our forecast combination framework to be applied operationally.

This study highlights the dual role of randomness in DL-based hydrological forecasting: as a source of uncertainty and as a resource for improving probabilistic predictions. By leveraging the diversity induced by random seeds through principled combination methods, our framework moves beyond ad hoc model selection, offering a systematic approach to harness stochasticity. In addition, the integration of wavelet features demonstrates how domain-specific preprocessing of the input data can enhance DL models, bridging gaps between data-driven and process-based insights.

In principle, the TFT model could also be set up for multivariate applications. For example, to forecast water temperature and streamflow simultaneously, as we plan to do next. However, the difference in running one model for water temperature and streamflow simultaneously or two separate models, one for each variable, does not lead to any changes in the methods presented here for forecast combination, which have to be applied variable by variable. Our results from this study form the basis to evaluate potential improvements when forecasting water temperature and streamflow together.

5 Conclusions

Deep Learning models depend on random number generators (seeds), which can cause substantial differences in sub-seasonal operational forecasts. This is the case for the Temporal Fusion Transformer (TFT) model, which is a global model that performs parameter fitting at all stations and lead times simultaneously. To address this randomness, we run the TFT model trained with different random seeds, and combine the resulting forecasts optimally with two methods: the Non-homogeneous Gaussian Regression (NGR) and the Beta-transformed Linear Pool (BLP). We find that the combination methods improve the forecast quality across almost all 22 stations and 32 days of lead time analyzed. This approach relieves the forecaster from choosing in advance the best performing TFT model among many different realizations. For normally distributed variables, like water temperature, there is almost no difference in the performance between BLP and NGR. However, if the distribution of the variable is heavier tailed, like streamflow, NGR is only comparable with BLP after transferring and fitting the parameters in lognormal space. It is noteworthy that the level of improvement from the combination methods increases with forecast lead time. In summary, we improve hydrological predictions by optimally combining an ensemble of TFT models with different random seeds, based on their performance over the latest forecasts.

Appendix A Wavelet decomposition

We include wavelet-decomposed air temperature features (levels d_6 and s_6) as input to the TFT model. These features capture long-term trends and averages, providing the model with additional contextual information that enhances its ability to predict hydrological variables. This aligns with previous studies demonstrating the utility of wavelet transforms in hydrology for isolating multi-scale temporal patterns (Bogner and Pappenberger 2011). A wavelet decomposition example is shown in Fig. 10. Including wavelet-transformed air temperature data significantly improved the accuracy of water temperature forecasts, particularly for stations with large catchment areas (Fig. 11). It should be noted

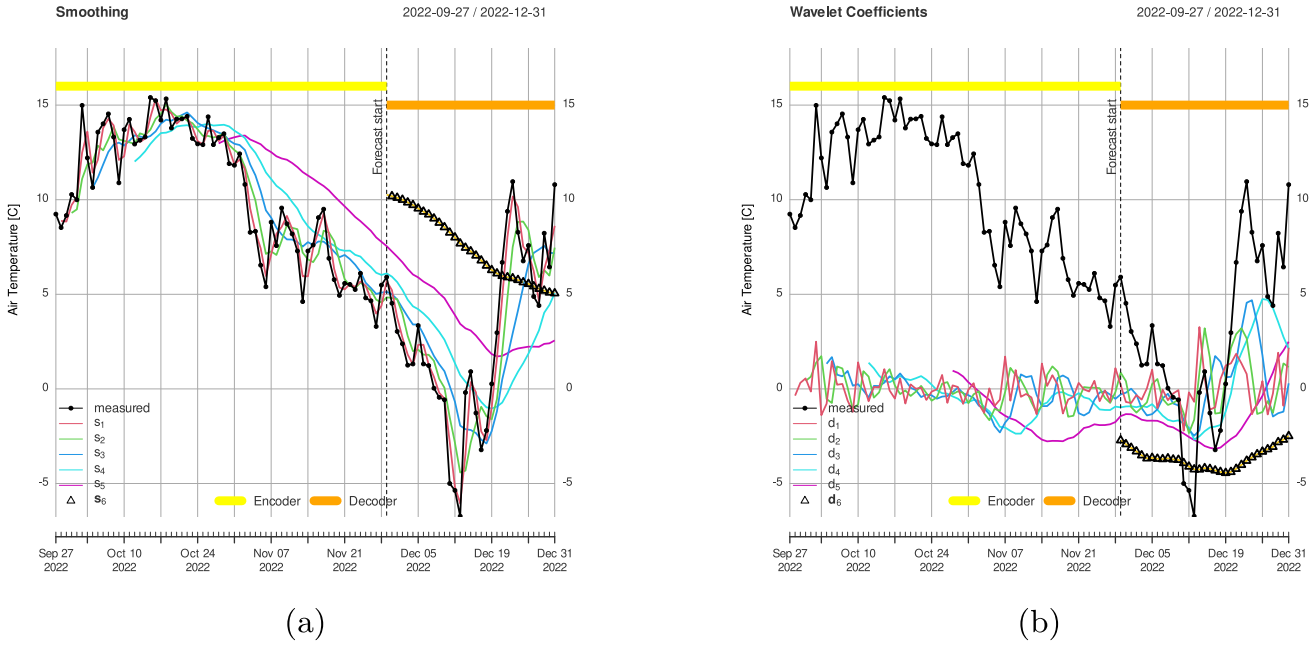


Fig. 10 Example of the wavelet decomposition of air temperature data into 6 levels of smoothing s (**a**) and detail coefficients d (**b**). In our case we use d_6 and s_6 as input features for the TFT model. In an operational setting, the measured air temperatures in the decoder are replaced with meteorological forecasts

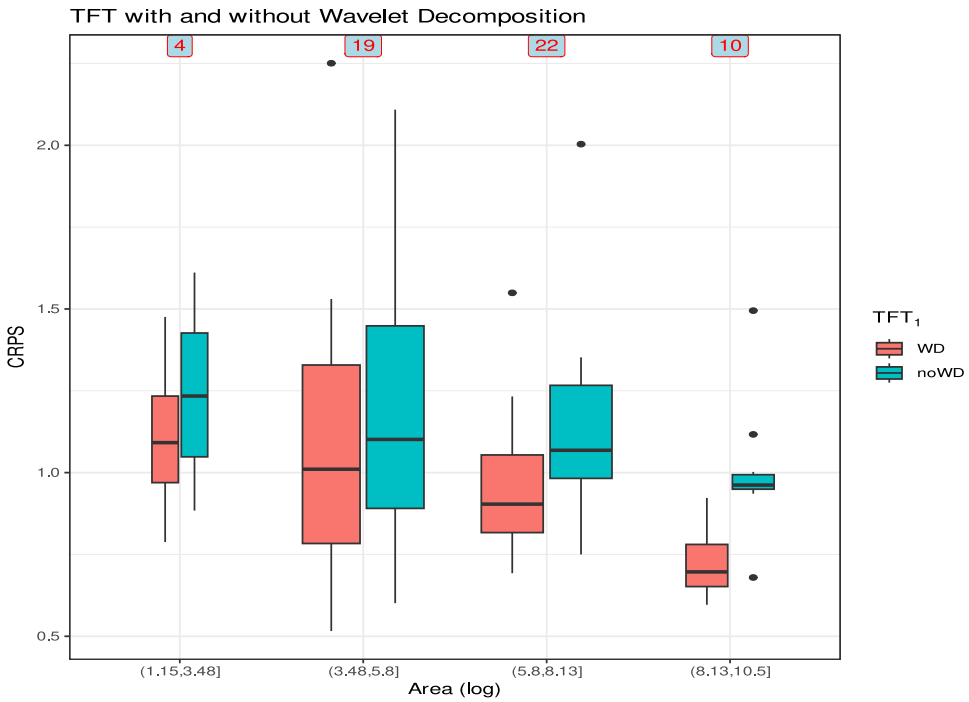


Fig. 11 Comparison of the CRPS of the TFT models (average of 12 seeds) including wavelet decompositions (WD) and without (noWD) wavelet decompositions of air temperature data (d_6 and s_6). Results are shown separately for different categories of catchment area. The number of stations included in each category is reported at the top

that this analysis is based on 55 stations with measurements of water temperature that have been used in the work of Padrón et al. (2025).

Appendix B BLP sensitivity to window size of past forecast initialization dates

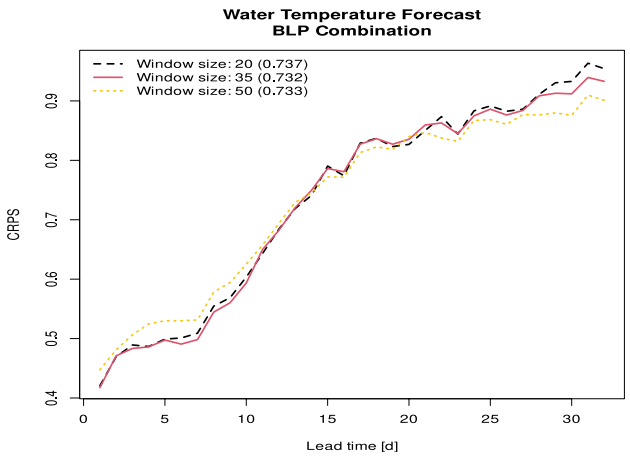


Fig. 12 Comparison of the water temperature CRPS obtained with the BLP combination method when using different window sizes of past forecast initialization dates. The numbers in parentheses correspond to the overall average CRPS also across all lead times

Acknowledgements We acknowledge the Swiss Federal Office for Meteorology and Climatology and the Swiss Federal Office for the Environment for providing the meteorological and hydrological data, respectively. We would especially like to thank Massimiliano Zappa for his expert contributions during fruitful discussions.

Author contributions Both authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by both authors. Both authors drafted and commented on the text. Both authors read and approved the final manuscript.

Funding This research has been supported by the Swiss Federal Institute for Forest, Snow and Landscape Research (EXTREMES program).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Altarabichi MG, Nowaczyk S, Pashami S et al (2024) Rolling the dice for better deep learning performance: a study of randomness techniques in deep neural networks. Inf Sci 667:120500. https://doi.org/10.1016/j.ins.2024.120500

Baran S, Lerch S (2015) Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. Q J R Meteorol Soc 141(691):2289–2299. https://doi.org/10.1002/qj.2521

Benaouda D, Murtagh F, Starck JL et al (2006) Wavelet-based nonlinear multiscale decomposition model for electricity load forecasting. Neurocomputing 70(1–3):139–154

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(1):281–305

Berrisch J, Ziel F (2023) CRPS learning. J Econom 237(2):105221. https://doi.org/10.1016/j.jeconom.2021.11.008

Berrisch J, Ziel F (2024) Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices. Int J Forecast. https://doi.org/10.1016/j.ijforecast.2024.01.005

Bogner K, Pappenberger F (2011) Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. Water Resour Res 47(7):W07524. https://doi.org/10.1029/2010WR009137

Bogner K, Liechti K, Zappa M (2017) Technical note: combining quantile forecasts and predictive distributions of stream-flows. Hydrol Earth Syst Sci 21(11):5493–5502

Bogner K, Chang AYY, Bernhard L et al (2022) Tercile forecasts for extending the horizon of skillful hydrological predictions. J Hydrometeorol 23(4):521–539. https://doi.org/10.1175/JHM-D-21-0020.1

Bottou L (2012) Stochastic gradient descent tricks. In: Montavon G, Orr GB, Müller KR (eds) Neural networks: tricks of the trade, 2nd edn. Springer, Berlin Heidelberg, Berlin, Heidelberg

Bröcker J, Kantz H (2011) The concept of exchangeability in ensemble forecasting. Nonlinear Proc Geophys 18(1):1–5. https://doi.org/10.5194/npg-18-1-2011

Bröcker J (2012) Evaluating raw ensembles with the continuous ranked probability score. Q J R Meteorol Soc 138(667):1611–1617. https://doi.org/10.1002/qj.1891

Chang AYY, Bogner K, Grams CM et al (2023) Exploring the use of European weather regimes for improving user-relevant hydrological forecasts at the subseasonal scale in Switzerland. J Hydrometeorol 24(10):1597–1617. https://doi.org/10.1175/JHM-D-21-0245.1

Dutilleux P (1987) An Implementation of the "algorithme a trous" to Compute the Wavelet Transform. In: Springer-Verlag C, J. M. and Grossman, A., Tchamitchian, Ph. (eds) Wavelets: Time-Frequency Methods and Phase Space. New York

Gneiting T, Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378

Gneiting T, Ranjan R (2013) Combining predictive distributions. Electron J Stat 7:1747–1782. https://doi.org/10.1214/13-EJS823

Gneiting T, Raftery A, Westveld A III et al (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133(5):1098–1118

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press http://www.deeplearningbook.org

Kratzert F, Klotz D, Brenner C et al (2018) Rainfall-runoff modelling using long short-term memory (LSTM) networks. Hydrol Earth Syst Sci 22(11):6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert F, Klotz D, Herrnegger M et al (2019) Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour Res 55(12):11344–11354. https://doi.org/10.1029/2019WR026065

Li W, Duan Q, Miao C et al (2017) A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. WIREs Water 4(6):e1246. https://doi.org/10.1002/wat2.1246

Lim B, Arik S, Loeff N et al (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast 37(4):1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Mienye ID, Sun Y (2022) A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access 10:99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Molteni F, Buizza R, Palmer TN et al (1996) The ECMWF ensemble prediction system: methodology and validation. Q J R Meteorol Soc 122(529):73–119

Monhart S, Spirig C, Bhend J et al (2018) Skill of subseasonal forecasts in Europe: effect of bias correction and downscaling using surface observations. J Geophys Res Atmospheres. https://doi.org/10.1029/2017JD027923

Narkhede MV, Bartakke PP, Sutaone MS (2022) A review on weight initialization strategies for neural networks. Artif Intell Rev 55(1):291–322. https://doi.org/10.1007/s10462-021-10033-z

Padrón RS, Zappa M, Bernhard L et al (2025) Extended-range forecasting of stream water temperature with deep-learning models. Hydrol Earth Syst Sci 29(6):1685–1702. https://doi.org/10.5194/hess-29-1685-2025

Ranjan R, Gneiting T (2010) Combining probability forecasts. J R Stat Soc Ser B (Statistical Methodology) 72(1):71–91. https://doi.org/10.1111/j.1467-9868.2009.00726.x

Rasiya Koya S, Roy T (2024) Temporal fusion transformers for streamflow prediction: value of combining attention with recurrence. J Hydrol 637:131301. https://doi.org/10.1016/j.jhydrol.2024.131301

Schirmer M, Hug Peter D, Lustenberger F (2025) openQUARREL: A meta-package integrating several hydrological model packages in R. http://mw-schirmer.github.io/openQUARREL/

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall/CRC monographs on statistics and applied probability, Chapman and Hall, London

Srivastava N, Hinton G, Krizhevsky A et al (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(56):1929–1958

Stankovic R, Falkowski B (2003) The Haar wavelet transform: its status and achievements. Comput Electr Eng 29(1):25–44

Tyralis H, Papacharalampous G (2024) A review of predictive uncertainty estimation with machine learning. Artif Intell Rev 57(4):94. https://doi.org/10.1007/s10462-023-10698-8

van der Meer D, Pinson P, Camal S et al (2024) CRPS-based online learning for nonlinear probabilistic forecast combination. Int J Forecast. https://doi.org/10.1016/j.ijforecast.2023.12.005

Wu J, Chen XY, Zhang H et al (2019) Hyperparameter optimization for machine learning models based on bayesian optimizationb. J Electr Sci Tech 17(1):26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Xie K, Liu P, Zhang J et al (2021) Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. J Hydrol 603:127043. https://doi.org/10.1016/j.jhydrol.2021.127043

Yu S, Ma J (2021) Deep learning for geophysics: current and future trends. Rev Geophys 59(3):e2021RG000742. https://doi.org/10.1029/2021RG000742

Zhu C, Byrd RH, Lu P et al (1997) Algorithm 778: L-bfgs-b: fortran subroutines for large-scale bound-constrained optimization. ACM Trans Math Softw 23(4):550–560. https://doi.org/10.1145/279232.279236