Supplement of Hydrol. Earth Syst. Sci., 29, 1685–1702, 2025
https://doi.org/10.5194/hess-29-1685-2025-supplement

Hydrology and Earth System Sciences
Open Access
EGU

Supplement of

Extended-range forecasting of stream water temperature with deep-learning models

Ryan S. Padrón et al.

Correspondence to: Ryan S. Padrón (ryan.padron@wsl.ch)

Table S1: List of 54 stream water temperature stations from the Swiss Federal Office for the Environment monitoring network (https://www.hydrodaten.admin.ch/en/) with their corresponding characteristics used as static features when training the deep learning models. *Indicates the subset of 20 stations not used for training the models when evaluating the predictive skill at new and ungauged stations. For each of these stations a corresponding similar station is indicated in parentheses whose data are used to normalize water temperature when evaluating the predictive skill at ungauged stations.

Station ID	Name	Area [km^2]	Mean elevation [m asl]	Glacierized fraction [%]	Coord X [m]	Coord Y [m]	Station elevation [m asl]	Water temp mean [ºC]	Water temp stdev [ºC]
2009	Rhône - Porte du Scex	5238	2127	11.1	557660	133280	377	8.36	2.42
2016	Aare - Brugg	11681	1000	1.5	657000	259360	332	13.31	5.76
2018* (2016)	Reuss - Mellingen	3386	1259	1.8	662830	252580	345	13.41	6.01
2019	Aare - Brienzwiler	555	2135	15.5	649930	177380	570	7.72	2.21
2029	Aare - Brügg, Aegerten	8249	1142	2.1	588220	219020	428	13.12	5.91
2030	Aare - Thun	2459	1746	6.9	613230	179280	548	12.39	4.90
2034	Broye - Payerne, Caserne d 'aviation	416	715	0	561660	187320	441	13.01	6.87
2044	Thur - Andelfingen	1702	770	0	693510	272500	356	12.98	6.46
2056* (2276)	Reuss - Seedorf	833	2013	6.4	690085	193210	438	8.35	3.15
2068	Ticino - Riazzino	1613	1643	0.1	713670	113500	200	10.95	4.20
2070* (2343)	Emme - Emmenmatt, nur Hauptstation	443	1065	0	623610	200430	638	10.81	5.39
2084	Muota - Ingenbohl	317	1363	0	688262	206170	438	9.32	3.48
2085* (2029)	Aare - Hagneck	5112	1368	3.4	580680	211650	437	13.18	5.21
2091	Rhein - Rheinfelden, Messstation	34524	1068	0.8	627189	267845	265	13.54	5.91
2104	Linth - Weesen, Biäsche	1062	1584	1.6	725160	221380	419	12.08	5.14
2106	Birs - Münchenstein, Hofmatt	887	728	0	613570	263080	268	12.28	4.97
2109* (2019)	Lütschine - Gsteig	381	2050	13.5	633130	168200	585	7.81	2.80
2112	Sitter - Appenzell	74.4	1256	0.1	749040	244220	769	9.72	5.13
2113* (2091)	Aare - Felsenau, K.W. Klingnau (U.W.)	17687	1060	1.4	659150	271790	312	14.14	5.90
2126	Murg - Wängi	80.2	652	0	714105	261720	466	12.50	4.85
2135	Aare - Bern, Schönau	2941	1596	5.8	600710	198000	502	12.62	5.04
2150* (2276)	Landquart - Felsenbach	614	1797	0.7	765364	204910	571	8.55	4.26
2152	Reuss - Luzern, Geissmattbrücke	2254	1504	2.8	665330	211800	432	13.34	5.84
2159* (2126)	Gürbe - Belp, Mülimatt	116	846	0	604810	192680	522	12.82	5.84
2167	Tresa - Ponte Tresa, Rocchetta	609	803	0	709580	92145	268	16.13	7.10

2179	Sense - Thörishaus, Sensematt	351	1071	0	593350	193020	553	11.62	6.39
2210* (2179)	Doubs - Ocourt	1275	952	0	572530	244460	417	11.90	5.39
2243	Limmat - Baden, Limmatpromenade	2384	1131	0.7	665640	258690	351	13.83	6.51
2256* (2269)	Rosegbach - Pontresina	66.5	2704	21.7	788795	151694	1766	5.89	4.23
2269	Lonza - Blatten	77.4	2624	24.7	629130	140910	1520	5.87	2.72
2276	Grosstalbach - Isenthal	43.9	1819	6.7	685500	196050	767	8.51	2.09
2288	Rhein - Neuhausen, Flurlingerbrücke	11930	1239	0.6	689145	281975	383	13.29	6.53
2307* (2414)	Suze - Sonceboz	127	1036	0	579810	227350	642	10.03	2.79
2308	Goldach - Goldach, Bleiche, nur Hauptstation	50.4	832	0	753190	261600	399	11.89	7.01
2343	Langeten - Huttwil, Häberenbad	59.9	760	0	629560	219135	597	10.81	4.06
2351* (2617)	Vispa - Visp	786	2648	23.1	634030	125900	659	7.54	2.86
2369	Mentue - Yvonand, La Mauguettaz	105	675	0	545440	180875	449	11.41	5.98
2374* (2343)	Necker - Mogelsberg, Aachsäge	88.1	956	0	727110	247290	606	10.92	6.25
2392	Rhein (Oberwasser) - Rheinau	11950	1238	0.6	687420	277140	353	13.28	6.49
2414	Rietholzbach - Mosnang, Rietholz	3.19	794	0	718840	248440	682	9.77	4.62
2415	Glatt - Rheinsfelden	417	503	0	678040	269720	336	14.25	5.94
2432	Venoge - Ecublens, Les Bois	228	686	0	532040	154160	383	12.37	5.18
2433* (2493)	Aubonne - Allaman, Le Coulet	105	952	0	520720	147410	390	10.39	4.17
2457* (2369)	Aare - Ringgenberg, Goldswil	1138	1951	12.1	633730	171510	564	11.53	4.58
2462* (2617)	Inn - S-Chanf	616	2463	6.1	795800	165910	1645	7.09	4.26
2467	Saane - Gümmenen	1881	1131	0.1	585100	199240	473	11.61	5.35
2473* (2112)	Rhein - Diepoldsau, Rietbrücke	6299	1771	0.7	766280	250360	410	9.49	3.91
2485* (2432)	Allaine - Boncourt, Frontière	212	562	0	567830	261200	366	12.82	4.22
2493	Promenthouse - Gland, Route Suisse	120	1027	0	510080	140080	394	10.93	3.70
2604* (2112)	Biber - Biberbrugg	31.9	1003	0	697240	223280	825	9.88	6.14
2608* (2369)	Sellenbodenbach - Neuenkirch	10.4	608	0	658530	218290	515	11.75	5.51
2612	Riale di Pincascia - Lavertezzo	44.5	1705	0	708060	123950	536	8.98	5.78
2617	Rom - Müstair	128	2184	0	830800	168700	1236	7.50	3.37
2623	Rhone - Oberwald	93.3	2466	19.3	669870	154080	1368	5.25	2.42

Table S2: Hyperparameters of the deep learning models, the explored hyperparameter space, and their tuned values. Hyperparameter names agree with the PyTorch Forecasting documentation (https://pytorch-forecasting.readthedocs.io/en/stable/models.html).

Model	Hyperparameter	Tunning space	Tuned value
RNNED	hidden_size	32 – 128	116
	rnn_layers	1 – 4	1
	gradient_clip_val	0.1 – 0.3	0.144
	dropout	0 – 0.3	0.276
TFT	hidden_size	32 – 128	60
	hidden_continuous_size	8 – 32	12
	lstm_layers	1 – 4	2
	attention_head_size	1 – 4	1
	gradient_clip_val	0.1 – 0.3	0.204
	learning_rate	0.001 – 0.1	0.0033
	dropout	0 – 0.3	0.222
NHITS	hidden_size	32 – 128	51
	num_blocks	1 – 3	3
	num_stacks	1 – 3	3
	gradient_clip_val	0.1 – 0.3	0.105
	learning_rate	0.001 – 0.1	0.0033
	dropout	0 – 0.3	0.173

Table S3: Hyperparameter influence on the prediction skill of the RF and MLP models (ARX has no hyperparameters). The analysed RF hyperparameters are: num.trees (number of trees), mtry (number of variables to possibly split at in each node), and min.node.size (minimal node size to split at). The analysed MLP hyperparameters are: n.hidden (number of hidden nodes), iter.max (maximum iterations of the optimization algorithm), and n.trials (number of repeated trials used to avoid local minima). The prediction skill is evaluated as the average CRPS over the 32 lead times, 54 stations, and 90 forecasts distributed over the year 2022. Overall, there is little variability of the CRPS for the different hyperparameter settings, so the default option (in bold) is used throughout the manuscript. *Here we use only 25 members instead of the 51 members of the ensemble forecasts to reduce computing time, which is why the RF has a slightly different CRPS value than that reported in Fig. 4.

Model	Hyperparameters	Average CRPS*
RF	**500 trees, 2 mtry, 5 minimum node size**	0.810
	100 trees, 2 mtry, 5 minimum node size	0.811
	1000 trees, 2 mtry, 5 minimum node size	0.809
	500 trees, 1 mtry, 5 minimum node size	0.840
	500 trees, 5 mtry, 5 minimum node size	0.807
	500 trees, 2 mtry, 2 minimum node size	0.809
	500 trees, 2 mtry, 10 minimum node size	0.811
MLP	**2 hidden nodes, 5000 max iterations, 5 trials**	0.808
	5 hidden nodes, 5000 max iterations, 5 trials	0.788
	10 hidden nodes, 5000 max iterations, 5 trials	0.790
	2 hidden nodes, 1000 max iterations, 5 trials	0.804
	2 hidden nodes, 10000 max iterations, 5 trials	0.800
	2 hidden nodes, 5000 max iterations, 1 trials	0.805
	2 hidden nodes, 5000 max iterations, 10 trials	0.805

Table S4: List of start dates of the 90 forecasts during 2022 that are used for model evaluation. Forecasts are generated twice per week, with data missing for 2022-06-30, 2022-07-04, 2022-09-08, 2022-09-12, and 2022-10-03. For each of these 32-day forecasts the modelled water temperature at each station is compared against observations.

Number	Forecast start date	Number	Forecast start date	Number	Forecast start date
1	2022-01-03	31	2022-04-18	61	2022-08-08
2	2022-01-06	32	2022-04-21	62	2022-08-11
3	2022-01-10	33	2022-04-25	63	2022-08-15
4	2022-01-13	34	2022-04-28	64	2022-08-18
5	2022-01-17	35	2022-05-02	65	2022-08-22
6	2022-01-20	36	2022-05-05	66	2022-08-25
7	2022-01-24	37	2022-05-09	67	2022-08-29
8	2022-01-27	38	2022-05-12	68	2022-09-01
9	2022-01-31	39	2022-05-16	69	2022-09-05
10	2022-02-03	40	2022-05-19	70	2022-09-15
11	2022-02-07	41	2022-05-23	71	2022-09-19
12	2022-02-10	42	2022-05-26	72	2022-09-22
13	2022-02-14	43	2022-05-30	73	2022-09-26
14	2022-02-17	44	2022-06-02	74	2022-09-29
15	2022-02-21	45	2022-06-06	75	2022-10-06
16	2022-02-24	46	2022-06-09	76	2022-10-10
17	2022-02-28	47	2022-06-13	77	2022-10-13
18	2022-03-03	48	2022-06-16	78	2022-10-17
19	2022-03-07	49	2022-06-20	79	2022-10-20
20	2022-03-10	50	2022-06-23	80	2022-10-24
21	2022-03-14	51	2022-06-27	81	2022-10-27
22	2022-03-17	52	2022-07-07	82	2022-10-31
23	2022-03-21	53	2022-07-11	83	2022-11-03
24	2022-03-24	54	2022-07-14	84	2022-11-07
25	2022-03-28	55	2022-07-18	85	2022-11-10
26	2022-03-31	56	2022-07-21	86	2022-11-14
27	2022-04-04	57	2022-07-25	87	2022-11-17
28	2022-04-07	58	2022-07-28	88	2022-11-21
29	2022-04-11	59	2022-08-01	89	2022-11-24
30	2022-04-14	60	2022-08-04	90	2022-11-28

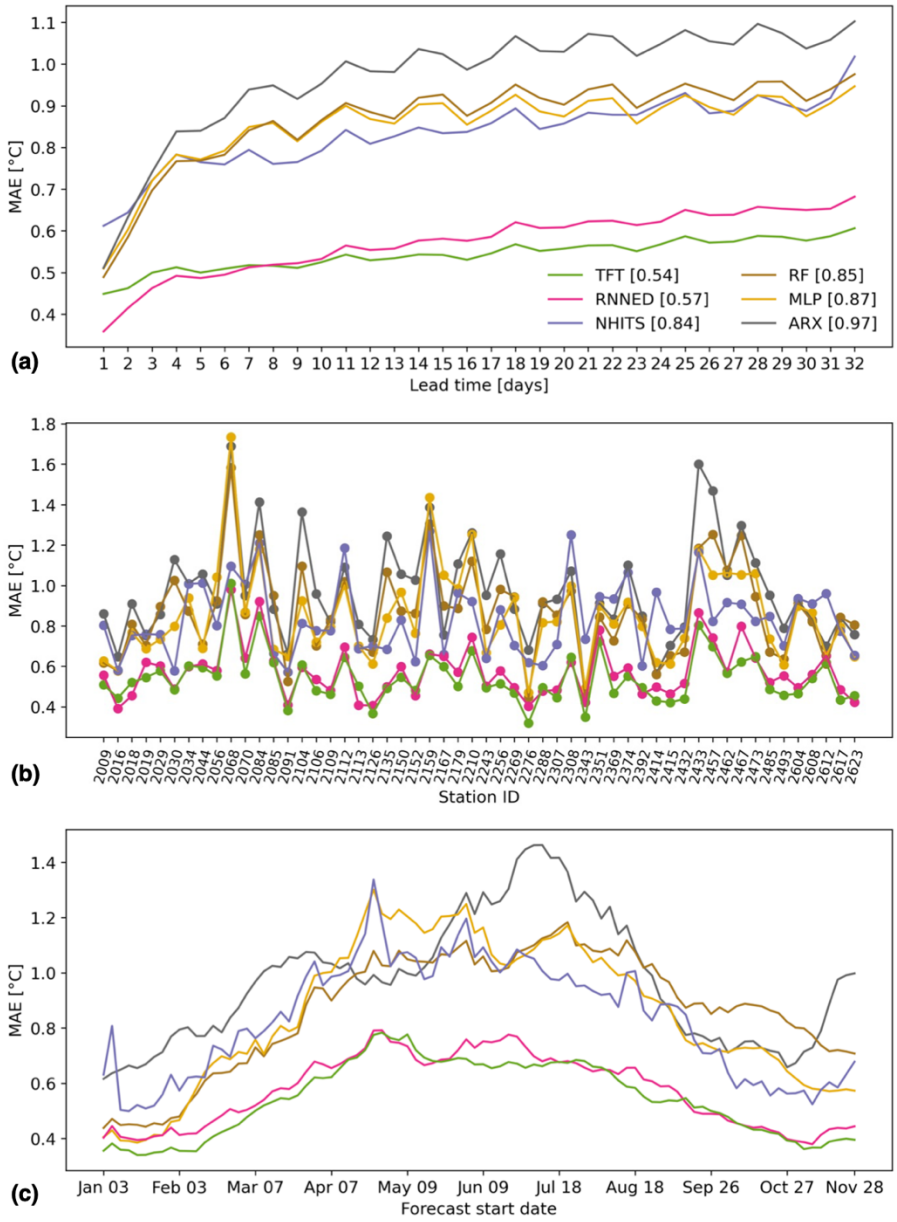


Figure S1: Model comparison of predictive skill when omitting the uncertainty of meteorological forecasts. The mean absolute error (MAE) of each model is shown as a function of **(a)** lead time averaged over all stations and forecasts, **(b)** station averaged over all lead times and forecasts, and **(c)** forecast start date averaged over all lead times and stations. The legend indicates the different models and their average MAE over all 32 lead times, 54 stations, and 90 forecasts distributed over the year 2022.

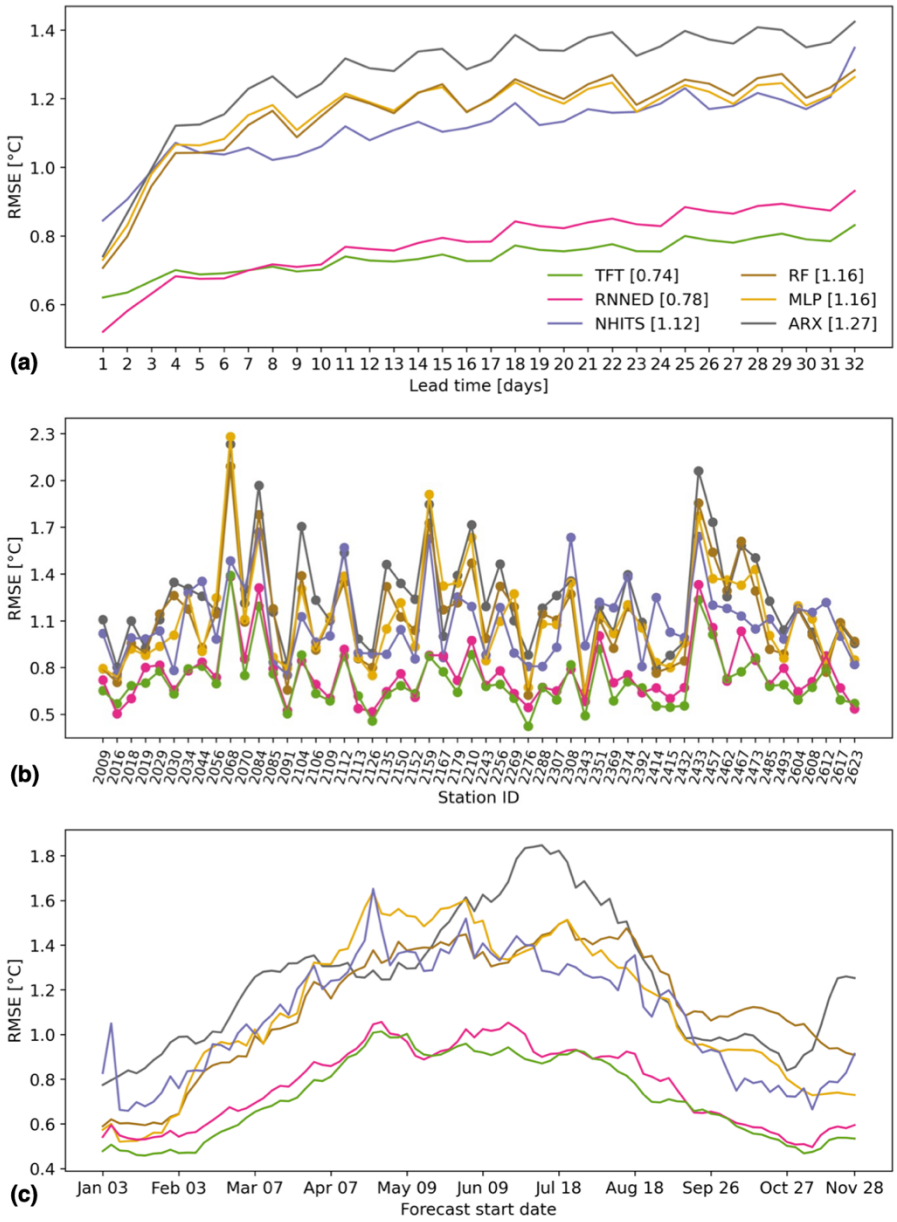


Figure S2: Model comparison of predictive skill when omitting the uncertainty of meteorological forecasts. The root mean squared error (RMSE) of each model is shown as a function of **(a)** lead time averaged over all stations and forecasts, **(b)** station averaged over all lead times and forecasts, and **(c)** forecast start date averaged over all lead times and stations. The legend indicates the different models and their average RMSE over all 32 lead times, 54 stations, and 90 forecasts distributed over the year 2022.

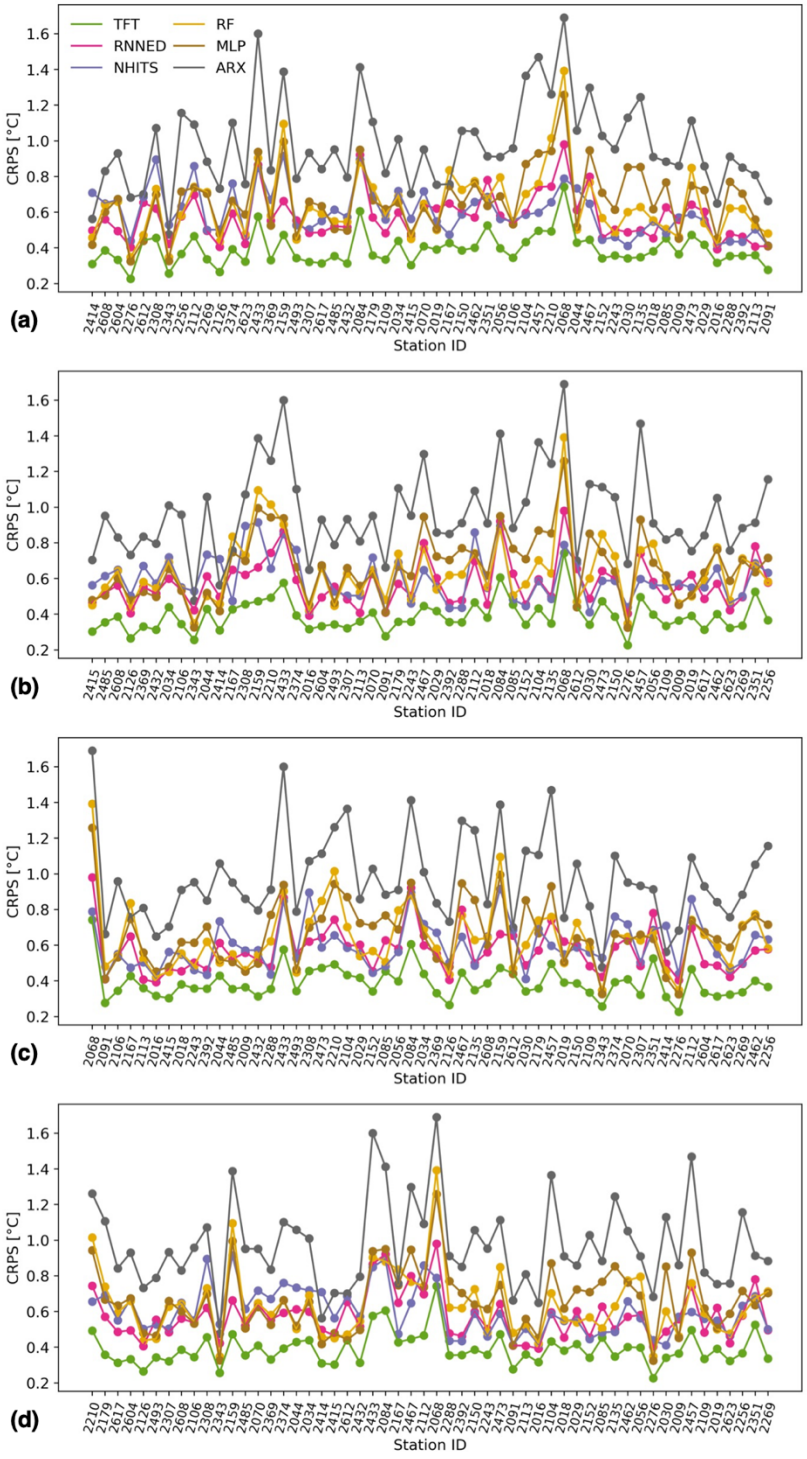


Figure S3: Model predictive skill per station when omitting the uncertainty of meteorological forecasts. Stations are sorted in ascending fashion according to **(a)** catchment area, **(b)** mean catchment elevation, **(c)** station elevation, and **(d)** glacierized fraction when omitting the uncertainty of meteorological forecasts. The CRPS is averaged over all lead times and forecasts.

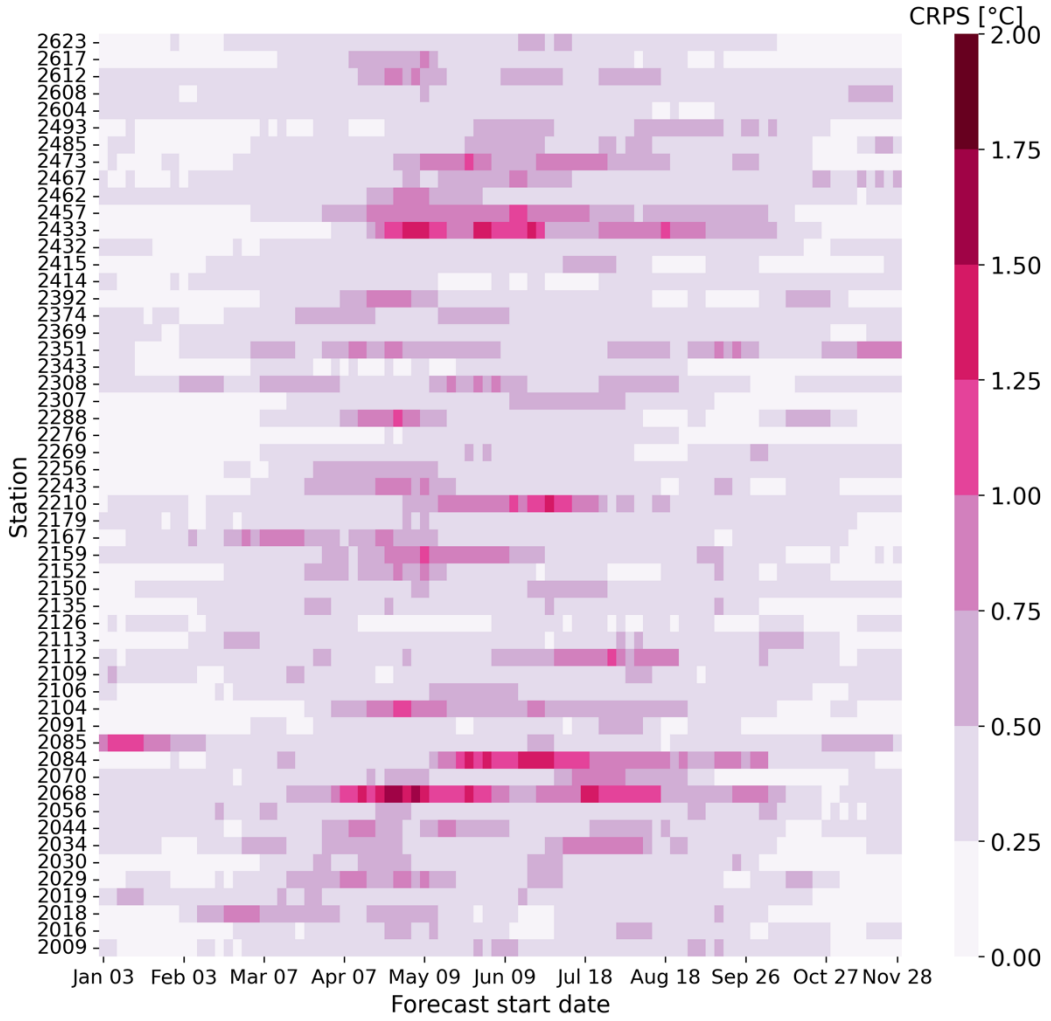


Figure S4: TFT predictive skill per forecast start date and station when omitting the uncertainty of meteorological forecasts. The CRPS is averaged over all lead times.

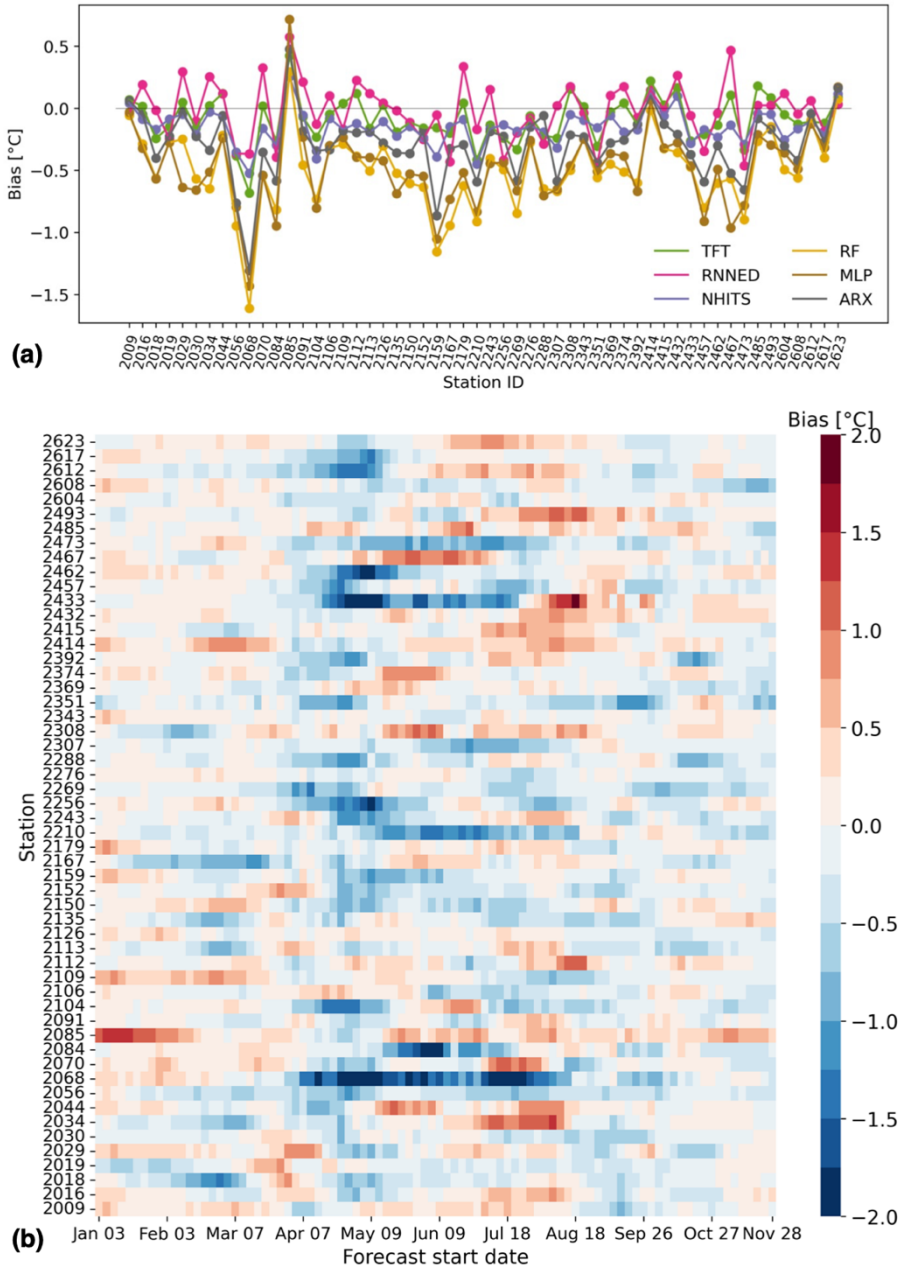


Figure S5: Water temperature forecast bias when omitting the uncertainty of meteorological forecasts. **(a)** Model comparison of forecast bias as a function of station averaged over all lead times and forecast start dates. **(b)** TFT forecast bias per forecast start date and station averaged over all lead times.

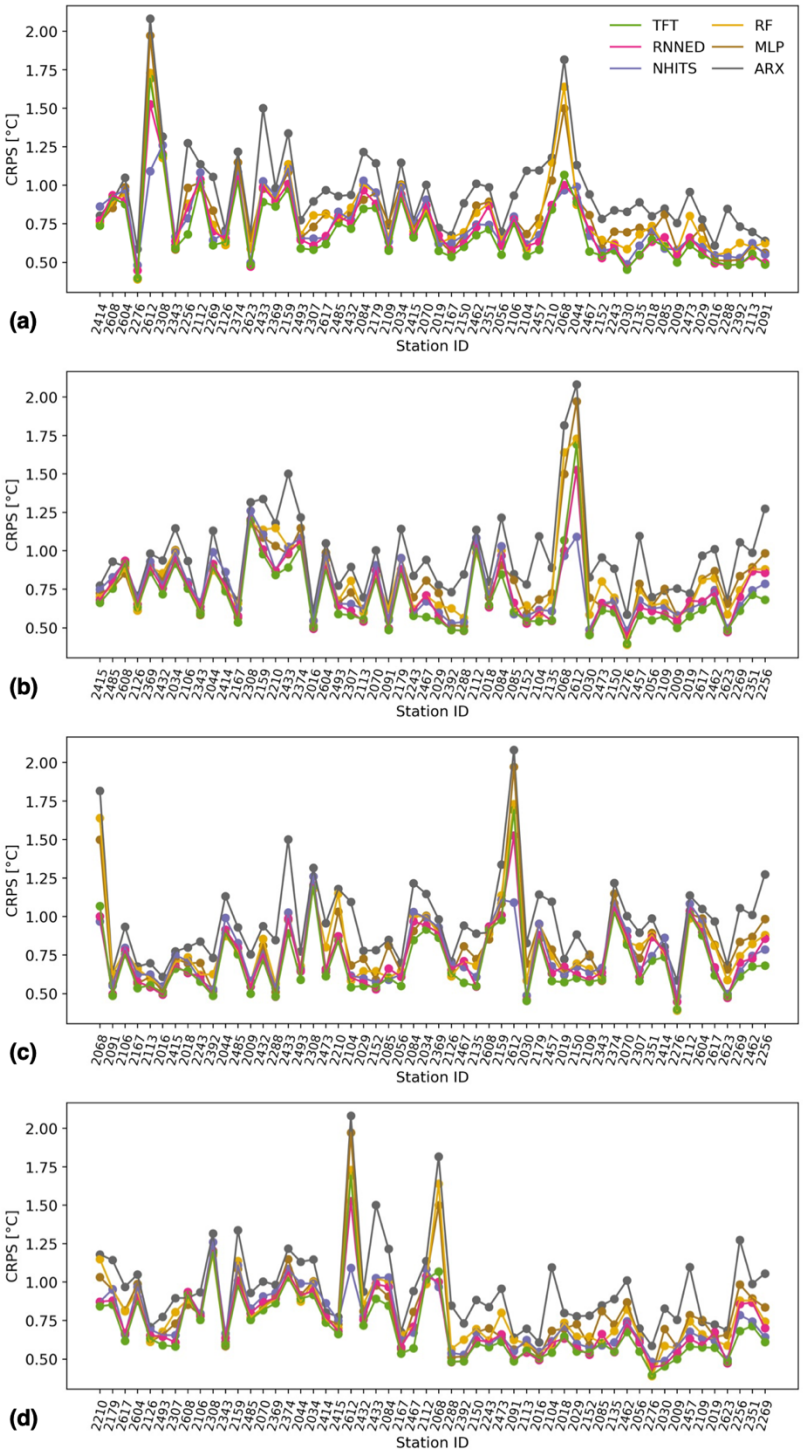


Figure S6: Model predictive skill per station when including the uncertainty of meteorological forecasts. Stations are sorted in ascending fashion according to **(a)** catchment area, **(b)** mean catchment elevation, **(c)** station elevation, and **(d)** glacierized fraction when omitting the uncertainty of meteorological forecasts. The CRPS is averaged over all lead times and forecasts.

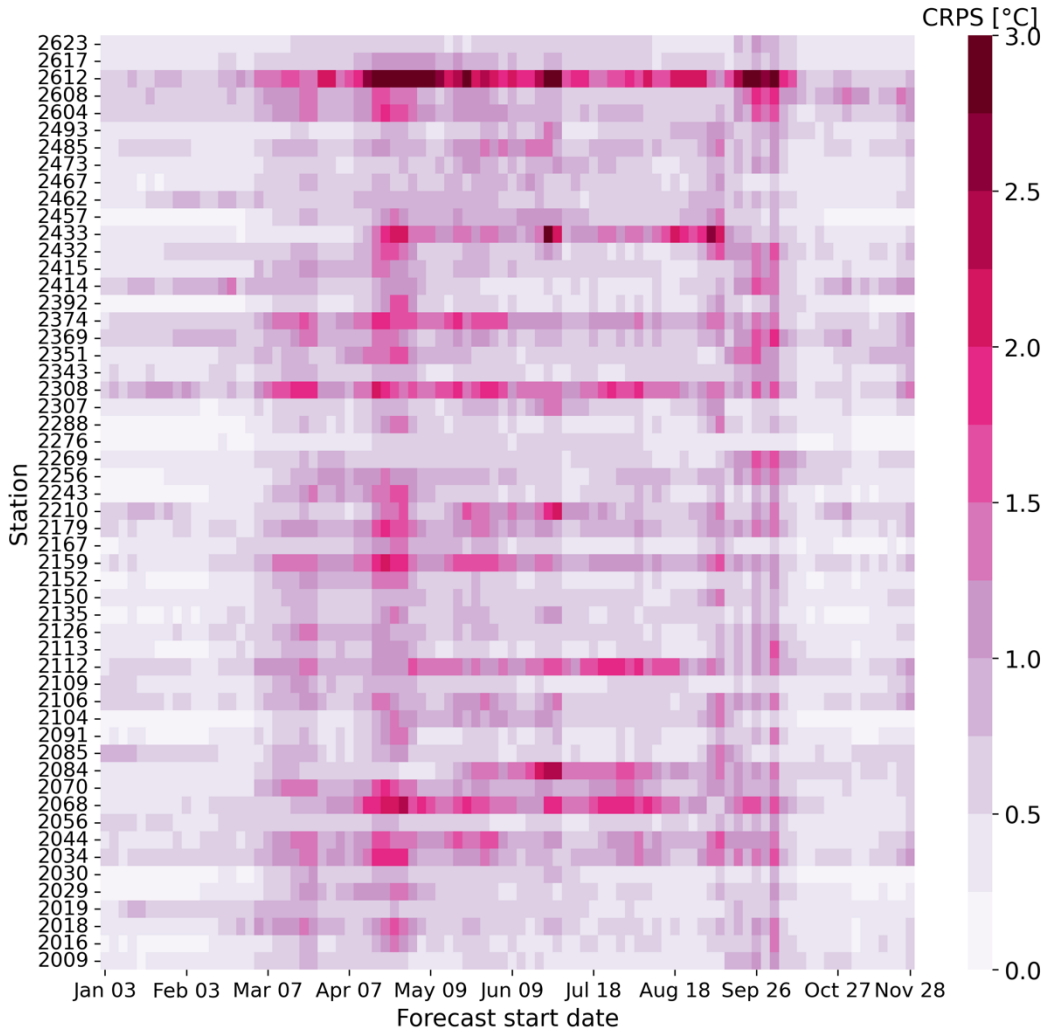


Figure S7: TFT predictive skill per forecast start date and station when including the uncertainty of meteorological forecasts. The CRPS is averaged over all lead times.

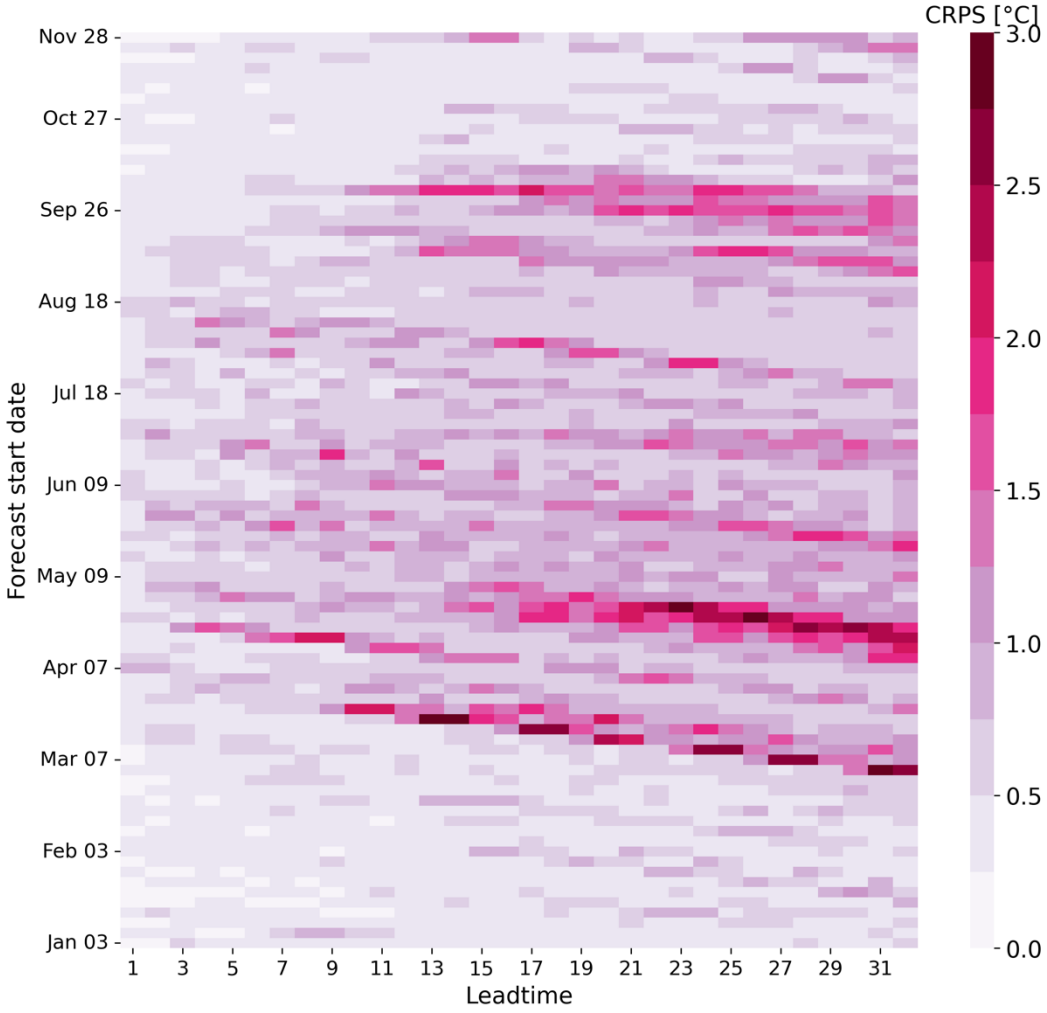


Figure S8: TFT predictive skill per lead time and forecast start date when including the uncertainty of meteorological forecasts. The CRPS is averaged over all stations.

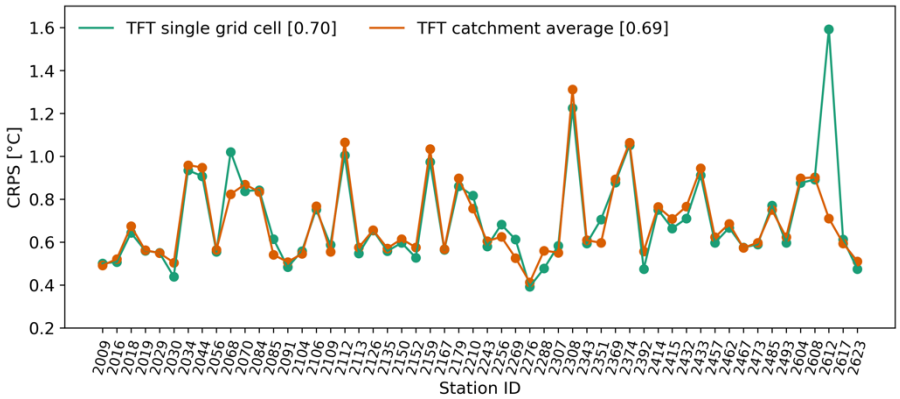


Figure S9: Comparison of TFT predictive skill per station if we use catchment average AT, P, and SD as predictors (orange) instead of their values at the single grid cell where the station is located as in Fig. 4 (green). Here we include the uncertainty of meteorological forecasts when estimating the predictive skill. The cumulative rank probability score (CRPS) is averaged over all lead times and forecasts. The legend indicates the average CRPS over all lead times, forecasts, and stations.

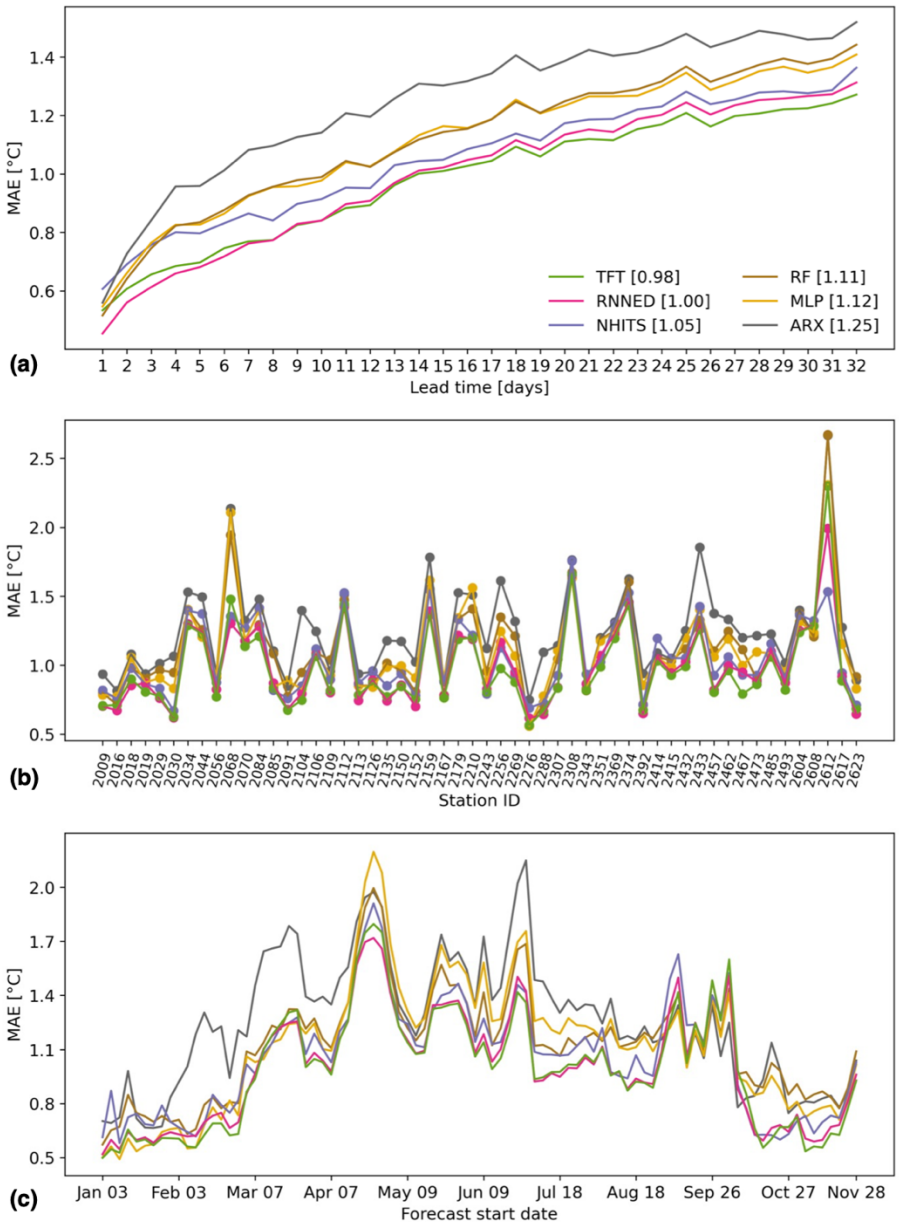


Figure S10: Model comparison of predictive skill when including the uncertainty of meteorological forecasts. The mean absolute error (MAE) of each model is shown as a function of **(a)** lead time averaged over all stations and forecasts, **(b)** station averaged over all lead times and forecasts, and **(c)** forecast start date averaged over all lead times and stations. The legend indicates the different models and their average MAE over all 32 lead times, 54 stations, and 90 forecasts distributed over the year 2022.

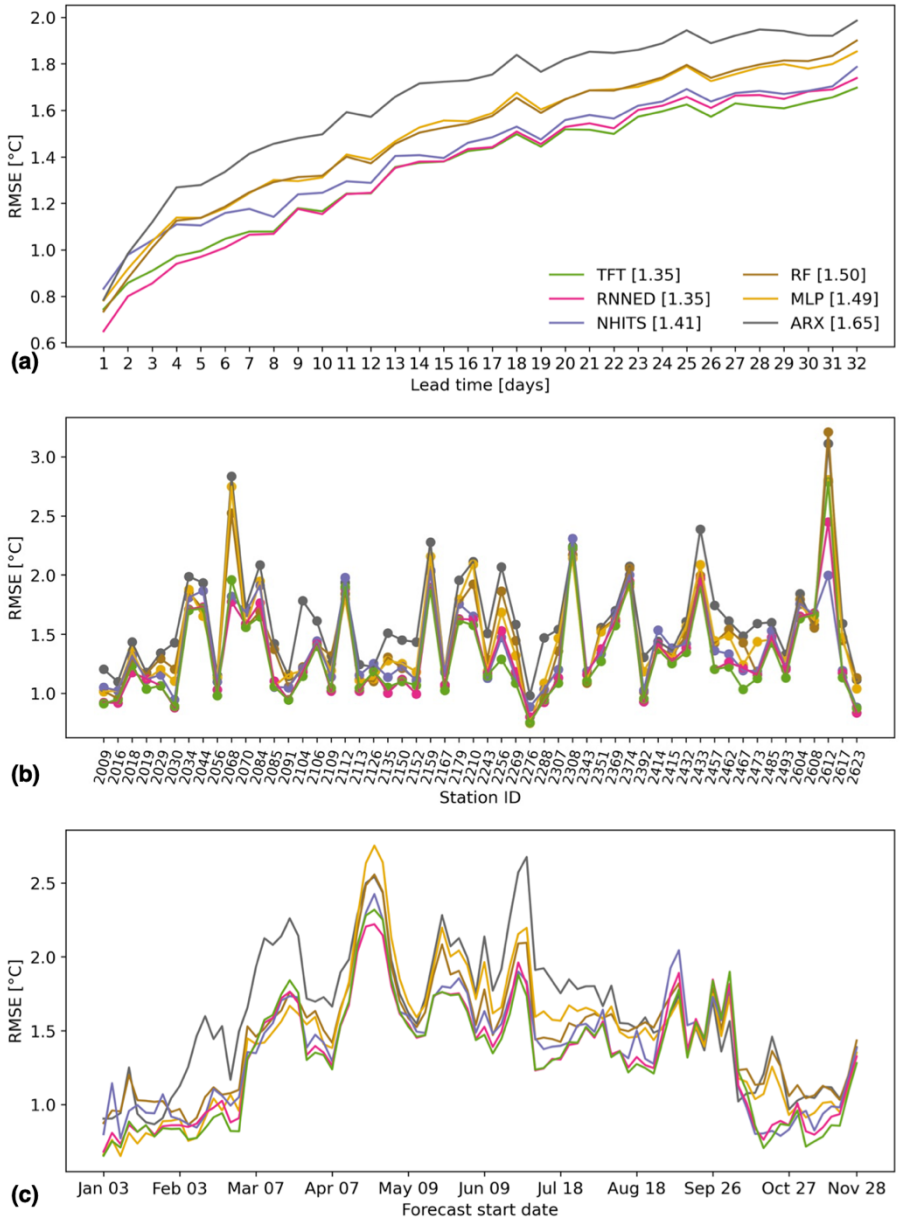
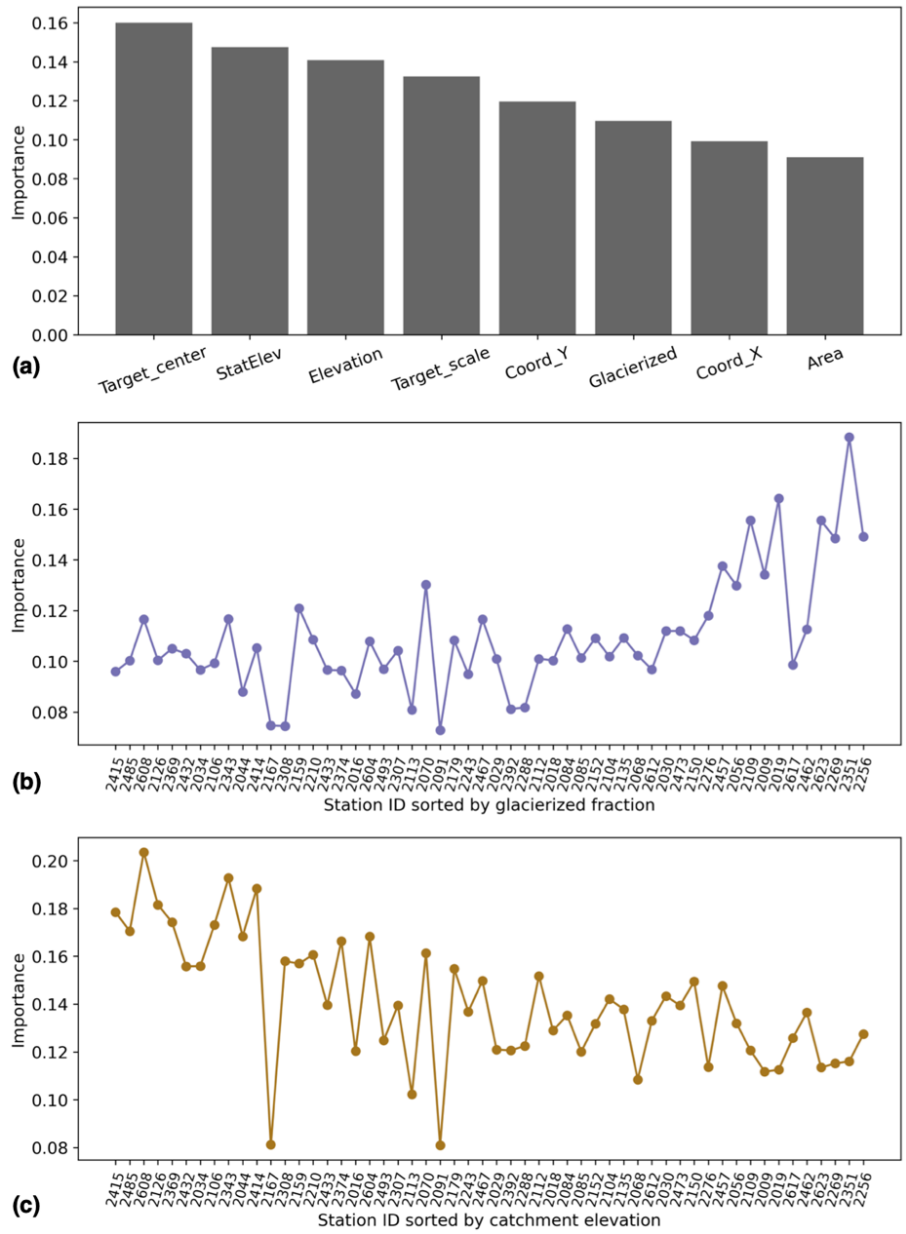


Figure S11: Model comparison of predictive skill when including the uncertainty of meteorological forecasts. The root mean squared error (RMSE) of each model is shown as a function of **(a)** lead time averaged over all stations and forecasts, **(b)** station averaged over all lead times and forecasts, and **(c)** forecast start date averaged over all lead times and stations. The legend indicates the different models and their average RMSE over all 32 lead times, 54 stations, and 90 forecasts distributed over the year 2022.



145 **Figure S12:** TFT static feature importance. **(a)** Feature importance averaged over 10 random seeds and 54 stations. **(b)** Glacierized fraction importance per station, with stations sorted according to the glacierized fraction of their catchments. **(c)** Catchment elevation importance per station, with stations sorted according to their mean catchment elevation.

150

155

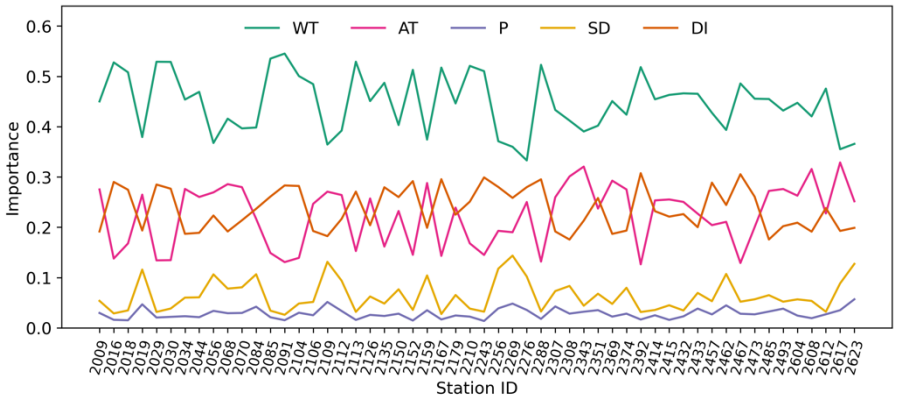


Figure S13: RF feature importance. The importance is deduced from the decrease in accuracy when randomly permuting the values of a feature. The larger the accuracy decrease, the higher the importance of the feature. Here the feature importance weights are obtained with the data from 2012 to 2021used for training the RF model.

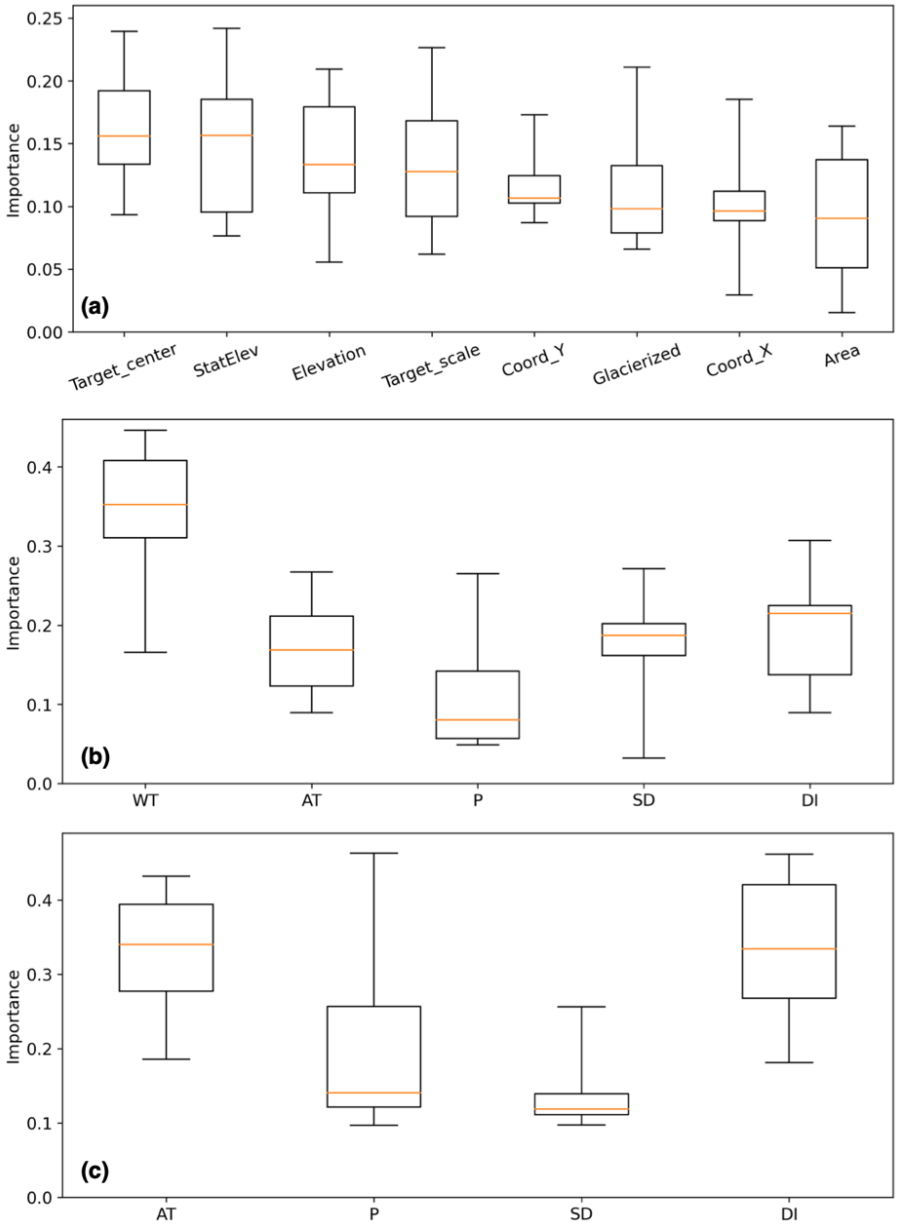


Figure S14: TFT average feature *importance* variability across 10 models trained with different random seeds. Results are shown for **(a)** static features, **(b)** encoder features, and **(c)** decoder features. Boxplots indicate the minimum, median, maximum, and interquartile range of the average *importance* across time steps, stations and forecast start dates.

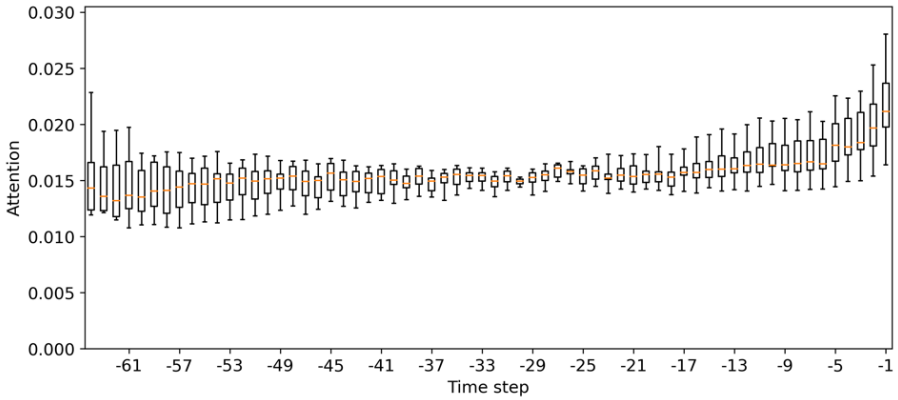


Figure S15: TFT *attention* variability across 10 models trained with different random seeds as a function of encoder time step. Boxplots indicate the minimum, median, maximum, and interquartile range of the average *attention* across stations and forecast start dates.