Supplementary information

Static attribute statistics

Attributes from Arsenault et al. (2020) are included directly in the dataset, whereas attributes from Addor et al. (2017) and MacKinnon (2010) are calculated from the hydrometeorological timeseries data in HYSETS. Seasonality attributes are calculated for flow and temperature using an additive seasonal decomposition for a frequency of one year. The seasonality is then given as:

$$\text{Seasonality} = 1 - \frac{\sum(w_t - \overline{w_t})^2}{\sum((w_t + s_t) - \overline{(w_t + s_t)})^2} \tag{1}$$

where s_t and w_t are the seasonal and residual (i.e. white noise) terms of additive seasonal decomposition. Negative seasonality values are truncated to 0, resulting in seasonality values ranging between 0 and 1, corresponding to no and high seasonality.

Table S1. Basin static attribute statistics (mean, standard deviation, 10th, 50th, and 90th percentile) calculated for 2363 Canadian catchments.

	mean	**std**	**10%**	**50%**	**90%**	**source**
elevation (m)	703	740	93.5	380	1.89e+03	(Arsenault et al., 2020)
slope (deg)	6.12	6.52	0.596	3.02	16	(Arsenault et al., 2020)
gravelius (-)	1.69	0.359	1.32	1.61	2.17	(Arsenault et al., 2020)
aspect (deg)	172	89	59.2	162	302	(Arsenault et al., 2020)
land use forest (frac)	0.435	0.294	0.0269	0.45	0.826	(Arsenault et al., 2020)
land use grass (frac)	0.0843	0.143	0.0018	0.027	0.24	(Arsenault et al., 2020)
land use wetland (frac)	0.0448	0.0892	0	0.008	0.145	(Arsenault et al., 2020)
land use water (frac)	0.0142	0.0315	0	0.0036	0.0349	(Arsenault et al., 2020)
land use urban (frac)	0.0945	0.167	0.0011	0.0464	0.219	(Arsenault et al., 2020)
land use shrubs (frac)	0.0987	0.153	0.0001	0.0336	0.293	(Arsenault et al., 2020)
land use crops (frac)	0.223	0.284	0	0.0754	0.745	(Arsenault et al., 2020)
land use snow ice (frac)	0.00485	0.0325	0	0	0.0001	(Arsenault et al., 2020)
permeability (log; m^2)	-13.9	1.18	-15.4	-14	-12.3	(Arsenault et al., 2020)
porosity (frac)	0.123	0.0641	0.0286	0.124	0.205	(Arsenault et al., 2020)
precip mean (mm/d)	2.74	1.19	1.29	2.71	3.93	(Addor et al., 2017)
precip high (mm/d)	13.7	5.94	6.44	13.6	19.6	(Addor et al., 2017)
precip low (mm/d)	1	0	1	1	1	(Addor et al., 2017)
precip mean (mm/mo)	83.2	36.2	38.6	82.5	120	(Addor et al., 2017)
precip mean (mm/y)	998	435	464	990	1.43e+03	(Addor et al., 2017)
precip high freq (d/yr)	3.93	0.528	3.21	4.06	4.46	(Addor et al., 2017)
precip high dur (d)	1.24	0.131	1.12	1.2	1.42	(Addor et al., 2017)
precip low freq (d/yr)	48.6	5.82	41.2	49	55.7	(Addor et al., 2017)
precip low dur (d)	5.5	2.78	3.53	4.65	8.26	(Addor et al., 2017)
baseflow index (-)	0.646	0.168	0.416	0.665	0.846	(Addor et al., 2017)
q mean (mm/d)	1.3	1.38	0.151	0.98	2.56	(Addor et al., 2017)
q high (mm/d)	11.7	12.5	1.36	8.82	23	(Addor et al., 2017)
q low (mm/d)	0.261	0.277	0.0302	0.196	0.512	(Addor et al., 2017)
q high freq (d/yr)	95.8	122	1	49	268	(Addor et al., 2017)
q high dur (d)	2.63	4.13	1	1.74	5.12	(Addor et al., 2017)
q low dur (d)	26.8	44.9	3.56	15.4	56.8	(Addor et al., 2017)
q zero freq (d/yr)	351	1.43e+03	0	0	698	(Addor et al., 2017)
q 95 (mm/d)	4.43	4.52	0.429	3.41	8.81	(Addor et al., 2017)
q 5 (mm/d)	0.162	0.316	0	0.0722	0.389	(Addor et al., 2017)
runoff ratio (-)	0.45	0.374	0.101	0.371	0.842	(Addor et al., 2017)
q adf (-)	-12.4	6.99	-19.4	-11.6	-5.16	(MacKinnon, 2010)
q seasonality (-)	0.272	0.231	0.0543	0.192	0.65	Eqn. 1
tmean seasonality (-)	0.782	0.0633	0.724	0.795	0.831	Eqn. 1
tmean (c)	9.95	5.32	3.77	9.81	17.5	-
tmax annual mean (c)	34.6	3.05	30.8	35	38.1	-
tmin annual mean (c)	-20.8	9.73	-33	-21.3	-7.76	-
q var $(\mathrm{mm/d})^2$	9.5	27.3	0.129	3.5	19	-

Sensitivity of silhouette scores to minimum cluster size

The clustering outcome and silhouette scores are highly dependent on the minimum cluster size. Unconstrained k-means clustering results in an optimum K value of 8, however, some of these clusters do not contain enough basins to use for validation. A comparison of silhouette scores across the numbers of clusters for minimum cluster sizes of 2, 8, 16, 32, 64, is provided in Fig. 1 below. Although greater numbers of clusters produce more hydrologically diverse sets of basins, there are simply not enough basins in each cluster to form a training dataset with balanced hydrological conditions (i.e., an equal number of basins from each cluster).

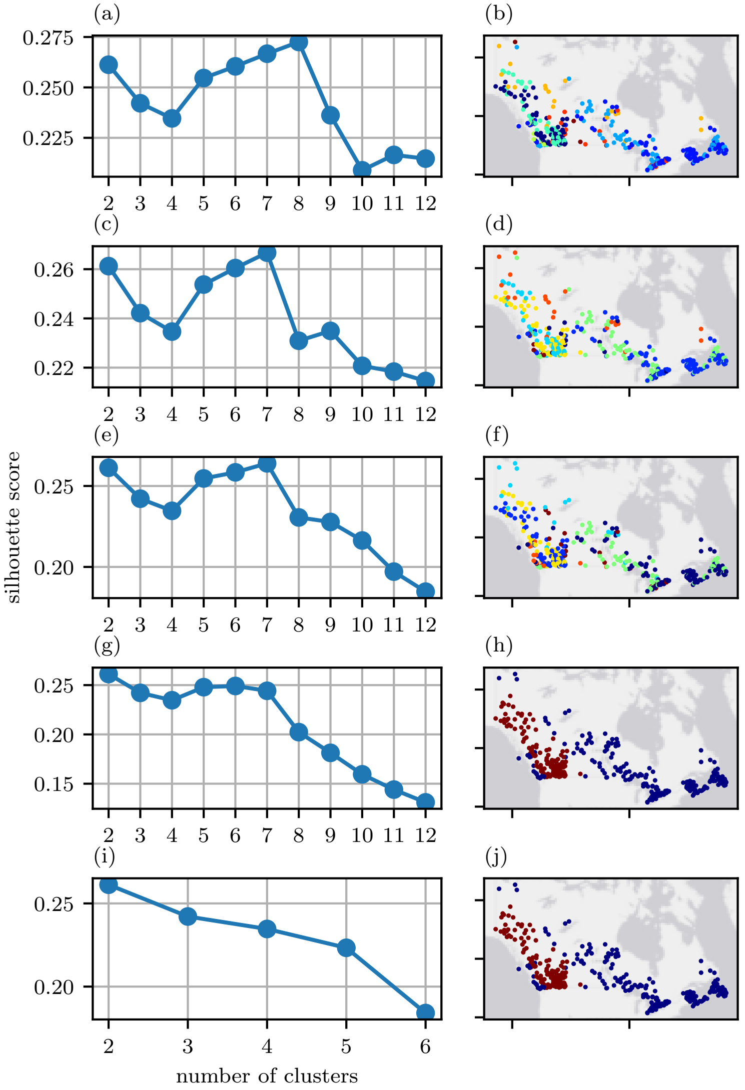

Figure S1. Silhouette scores (left) for minimum cluster sizes of 2 (a), 8 (c), 16 (e), 32 (g), and 64 (i). Respective clustering results (right) for K values (corresponding to maximum silhouette score) of 8 (b), 7 (d, f), and 2 (h, j). World Gray Canvas basemap provided by ESRI.

Two-cluster attributes

Table S2. Mean and standard deviation of basin attributes for 2 cluster configuration.

	C0 mean	C0 stdev	C1 mean	C1 stdev
elevation (m)	307.05	179.09	1432.84	327.84
slope (deg)	3.11	3.82	16.15	6.84
gravelius (-)	2.11	0.5	1.86	0.37
aspect (deg)	155.91	88.78	169.09	100.34
land use forest (frac)	0.58	0.27	0.6	0.2
land use grass (frac)	0.04	0.08	0.2	0.13
land use wetland (frac)	0.04	0.07	0.0	0.0
land use water (frac)	0.06	0.07	0.02	0.02
land use urban (frac)	0.04	0.1	0.01	0.01
land use shrubs (frac)	0.09	0.11	0.09	0.05
land use crops (frac)	0.14	0.26	0.02	0.11
land use snow ice (frac)	0.0	0.01	0.05	0.1
permeability (log; m^2)	-14.19	1.17	-14.07	0.9
porosity (frac)	0.1	0.06	0.11	0.06
precip mean (mm/d)	2.68	1.23	2.14	1.43
precip high (mm/d)	13.39	6.15	10.69	7.14
precip low (mm/d)	1.0	0.0	1.0	0.0
precip mean (mm/mo)	81.2	37.74	64.45	43.81
precip mean (mm/y)	974.38	452.9	773.45	525.76
precip high freq (d/yr)	3.57	0.5	3.33	0.66
precip high dur (d)	1.16	0.11	1.23	0.07
precip low freq (d/yr)	45.0	5.12	45.65	6.79
precip low dur (d)	3.99	1.63	4.64	0.98
baseflow index (-)	0.71	0.13	0.78	0.1
q mean (mm/d)	1.8	1.54	1.86	1.79
q high (mm/d)	16.18	13.88	16.77	16.13
q low (mm/d)	0.36	0.31	0.37	0.36
q high freq (d/yr)	69.38	91.85	47.28	83.23
q high dur (d)	2.67	2.69	2.35	2.54
q low dur (d)	29.66	34.52	42.33	35.39
q zero freq (d/yr)	183.47	970.46	135.96	1064.88
q 95 (mm/d)	6.13	5.27	6.42	5.57
q 5 (mm/d)	0.21	0.27	0.24	0.29
runoff ratio (-)	0.6	0.38	0.87	0.57
q adf (-)	-14.24	5.33	-11.23	4.58
q seasonality (-)	0.3	0.2	0.51	0.27
tmean seasonality (-)	0.8	0.1	0.76	0.11
tmean (c)	3.69	3.81	3.31	3.94
tmax annual mean (c)	31.03	2.27	31.17	2.83
tmin annual mean (c)	-30.24	9.02	-31.26	10.03
q var $(\mathrm{mm/d})^2$	11.02	26.07	10.2	21.08

Additional two-cluster comparison performance metrics

The NSE α and β decompositions are taken from (Gupta et al., 2009):

$$\alpha - \text{NSE} = \sigma_{\hat{q}} - \sigma_q \tag{2}$$

where $\sigma_{\mathrm{\hat{q}}}$ and σ_{q} are the standard deviations of the simulated and observed streamflow.

$$\beta - \text{NSE} = \frac{\mu_{\hat{q} - \mu_q}}{\sigma_q} \tag{3}$$

where $\mu_{\mathrm{\hat{q}}}$ and μ_{q} are the means of the simulated and observed streamflows.

The overfitting ratio, which quantifies the relative difference in train and test error, is calculated based on the MSE:

$$\text{MSE} = \frac{1}{n}\sum(q_t - \hat{q}_t)^2 \tag{4}$$

The overfitting ratio is given as:

$$\text{OVERFITTING RATIO} = 1 - \frac{\text{MSE}_{\text{test}}}{\text{MSE}_{\text{train}}} \tag{5}$$

where MSE_{test} and $\text{MSE}_{\text{train}}$ are the MSE scores on the test, and train partitions, respectively. The overfitting ratio ranges $-\infty$ to an ideal value of 1, which corresponds to 0 error on the test dataset. Scores of 0 indicate the same error on train and test data.

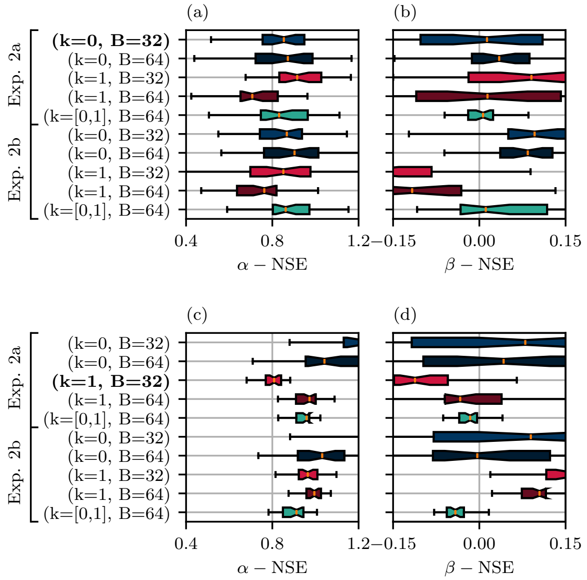

Figure S2. Boxplots showing the $\alpha - \text{NSE}$ (left) and $\beta - \text{NSE}$ of models evaluated on 32 C0 basins (a, c) and 32 C1 basins (b, d) for a lead time of 3 days. The baseline model, which is trained uniquely using the evaluation basins, is indicated in bold. Blue, red, and green colours indicate models trained on C0, C1, and both types of basins.

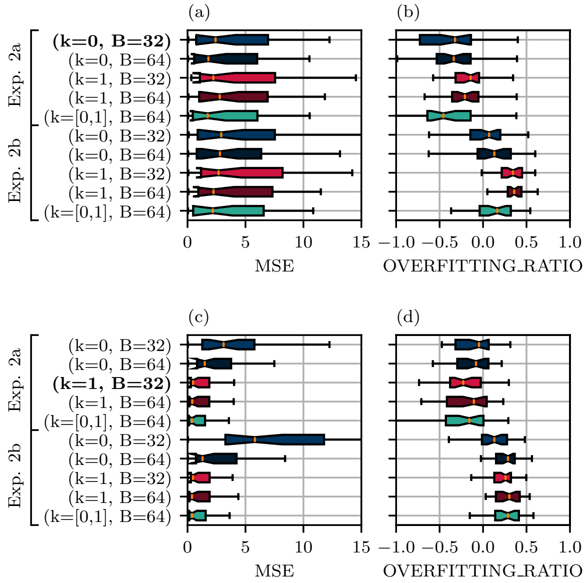

Figure S3. Boxplots showing the MSE (left) and the overfitting ratio (right) of models evaluated on 32 C0 basins (a, c) and 32 C1 basins (b, d) for a lead time of 3 days. The baseline model, which is trained uniquely using the evaluation basins, is indicated in bold. Blue, red, and green colours indicate models trained on C0, C1, and both types of basins.