

1 **Daily streamflow forecasting by machine**
2 **learning methods with weather and climate**
3 **inputs**

4 **Kabir Rasouli**

Department of Earth and Ocean Sciences, University of British Columbia
Vancouver, BC, Canada

William W. Hsieh*

Department of Earth and Ocean Sciences, University of British Columbia
Vancouver, BC Canada

Alex J. Cannon

Meteorological Service of Canada, Vancouver, BC, Canada

Submitted to the Journal of Hydrology (revised)

5 August 16, 2011

6 **Abstract**

7 Weather forecast data generated by the NOAA Global Forecasting System (GFS)
8 model, climate indices, and local meteo-hydrologic observations were used to forecast
9 daily streamflows for a small watershed in British Columbia, Canada, at lead times of
10 1–7 days. Three machine learning methods – Bayesian neural network (BNN), sup-
11 port vector regression (SVR) and Gaussian process (GP) – were used and compared
12 with multiple linear regression (MLR). The nonlinear models generally outperformed
13 MLR, and BNN tended to slightly outperform the other nonlinear models. Among

*Corresponding author: William Hsieh, Department of Earth and Ocean Sciences, University of British Columbia, 6339 Stores Road, Vancouver, BC, Canada V6T 1Z4; email: whsieh@eos.ubc.ca, telephone: 1-604-822-2821

14 various combinations of predictors, local observations plus the GFS output were gen-
15 erally best at shorter lead times, while local observations plus climate indices were
16 best at longer lead times. The climate indices selected include the sea surface tem-
17 perature in the Niño 3.4 region, the Pacific-North American teleconnection (PNA),
18 the Arctic Oscillation (AO) and the North Atlantic Oscillation (NAO). In the binary
19 forecasts for extreme (high) streamflow events, the best predictors to use were the
20 local observations plus GFS output. Interestingly, climate indices contribute to daily
21 streamflow forecast scores during longer lead times of 5–7 days, but not to forecast
22 scores for extreme streamflow events for all lead times studied (1–7 days).

23 Keywords: Streamflow, forecast, machine learning, artificial neural network, sup-
24 port vector regression, Gaussian process, British Columbia

25 **1 Introduction**

26 Short lead time streamflow forecasts are used to estimate the inflow of a given reservoir in
27 a watershed for the next few hours or days. These forecasts are used to plan hydroelectric
28 power scheduling and flood mitigation which requires active regulation of reservoir storage
29 for optimum use of available water resources. Factors affecting the rainfall-runoff process
30 include local micrometeorological conditions, such as soil moisture, soil temperature, and
31 snow budget. Snow budget is sensitive to temperature changes and variability in the
32 atmospheric circulation (Brown and Goodison, 1996).

33 Among large-scale climate variability patterns, the Pacific-North American pattern
34 (PNA) (manifested most clearly in the 500 hPa geopotential height anomalies) shows the
35 strongest relation to snow cover variability over North America (Brown and Goodison,
36 1996), with a higher PNA index relating to lower snow cover. The El Niño-Southern Os-
37 cillation (ENSO), the prominent interannual variability characterized by the equatorial
38 Pacific sea surface temperatures (SST) anomalies, also exerts influence on accumulated
39 snow in western Canada (Hsieh and Tang, 2001), with more snow during cool episodes

40 (La Niña) and less snow during warm episodes (El Niño). The North Atlantic Oscillation
41 (NAO), characterized by the difference of sea level pressure between the Icelandic low and
42 the Azores high (Hurrell, 1995), has been found to influence snow variability across the
43 eastern US and southern Ontario, Canada (Coulibaly and Burn, 2005). NAO is commonly
44 considered to be the main part of the Arctic Oscillation (AO), characterized by zonally
45 symmetric seesaw between sea level pressures in the polar latitudes and the temperate
46 latitudes in the N. Hemisphere (Thompson and Wallace, 1998).

47 In coastal regions the main part of inflow results from seasonal storms and spring
48 snowmelt. To simulate and forecast the streamflows accurately, study of climate variability
49 is needed. Analyzing climate influences on streamflow timing, Burn (2008) found that in
50 headwater basins the spring freshet displayed a trend indicating earlier occurrence. ENSO is
51 associated with lower and higher winter precipitation for El Niño and La Niña episodes, re-
52 spectively, in the Pacific Northwest and oppositely in the desert Southwest, USA (Kennedy
53 *et al.*, 2009, Fleming *et al.*, 2007). Pacific Decadal Oscillation (PDO), a decadal-scale oscil-
54 lation characterized by the leading principal component of the North Pacific SST anomalies
55 (Mantua *et al.*, 1997), and ENSO have influence on runoff timing in the Mackenzie River
56 basin in Canada and in the Upper Klamath Lake, Oregon, USA (Burn, 2008, Kennedy
57 *et al.*, 2009) and on American and European climate variations (Makkeasorn *et al.*, 2008).

58 Climate variability as well as other changes in the land use, the topographic and morpho-
59 logic features of the study area, have direct or indirect effects on the streamflow fluctuations.
60 Fleming *et al.* (2007) found that the ENSO effects on annual hydrometeorological cycles
61 can vary from one stream to the next and they suggested the consideration of the local hy-
62 drological characteristics and dynamics in the analysis of the ocean-atmosphere circulation
63 impacts on streamflow. For instance, high-elevation melt-driven nival watersheds exhibit
64 ENSO-driven streamflow anomalies predominantly during the spring-summer melt freshet.
65 Fleming *et al.* (2006) investigated teleconnections between AO and monthly streamflow for
66 the glacier and snowmelt-fed rivers in British Columbia and Yukon, in northern Canada.

67 Clark and Hay (2004) applied medium-range numerical weather prediction model out-
68 put to produce forecasts of streamflow. They used the forecasted atmospheric variables
69 (e.g. total column precipitable water, 2-m air temperature) as predictors in a forward
70 screening multiple linear regression model to improve local forecasts of precipitation and
71 temperature in stations. Competent predictor selection turns out to be an integral part
72 of skillful forecast models. Makkeasorn *et al.* (2008) studied the global climate variability
73 effects on short-term streamflow forecasting using genetic programming and neural net-
74 works. These and similar recent studies (Moradkhani *et al.*, 2004, Fleming *et al.*, 2007)
75 show an increasing interest in incorporating atmospheric circulation variability and outputs
76 of numerical weather prediction models in hydro-meteorologic forecasting, while in the con-
77 ventional methods the emphasis generally was on using local recorded data as predictors.
78 Also, inclusion of the numerical weather forecasts by the NOAA Global Forecasting System
79 (GFS) model (Hamill *et al.*, 2006) for accumulated precipitation, temperature, zonal and
80 meridional winds, relative humidity, sea level pressure, and precipitable water in models
81 (Makkeasorn *et al.*, 2008) can be useful in terms of capturing the governing short and
82 medium range signals in the study area.

83 Previous studies have not combined climate variability with weather forecasts and local
84 observed information as predictors in short-term streamflow forecasting. In this paper,
85 we aim to fully incorporate all interannual, seasonal and short-term signals to give the
86 most comprehensive and accurate daily streamflow forecasts, and to clarify the role of
87 climate variability information on short-term streamflow forecasting. The atmospheric
88 variables reforecasted by a numerical weather prediction model as well as indices of climate
89 variability (e.g. ENSO, PNA, AO, NAO and PDO) are added to the local observations
90 as extra predictors. Sophisticated nonlinear machine learning techniques, such as support
91 vector regression (SVR), Bayesian neural network (BNN) and Gaussian process (GP), are
92 compared against multiple linear regression (MLR).

93 In Section 2, the study area is introduced and the data used for modelling are pre-

94 sented. The machine learning models are described in Section 3 and the methodology for
95 forecast evaluation is explained in Section 4. The results are given in Section 5, followed
96 by conclusions.

97 **2 Study area and data sources**

98 To explore the applicability of machine learning methods to streamflow forecasting, we
99 chose Stave River above Stave Lake in southern British Columbia, Canada (Figure 1),
100 with a catchment drainage area of about 282 km², for a case study. The measurement
101 record used in our study is 1983–2001 — with 1983–1997 used for model development,
102 and 1998–2001 reserved for model forecast verification (i.e. testing). The basin is a mixed
103 pluvial-nival system, exhibiting an annual snowmelt-driven streamflow maximum in early
104 summer and rainfall-driven peaks in winter. The major source of precipitation in the Stave
105 basin comes during November to February from southwesterly frontal flows of warm moist
106 air aloft, due to the abrupt rise of the Coast Mountains above the flat or rolling topography
107 of the Lower Fraser Valley, causing on average 53% of the annual precipitation to fall during
108 this 4-month period. Precipitation events in the summer are generally from weak fronts
109 or convective storms. The mean annual discharge is 33.78 m³s⁻¹ and the mean annual
110 precipitation is 3041 mm. The annual means of the daily maximum temperature and daily
111 average temperatures are 10.3 and 6.8°C, respectively.

112 The local meteo-hydrological observations, reforecast dataset generated by the GFS
113 model (formerly known as Medium-Range Forecast, MRF) (Hamill *et al.*, 2006) initialized
114 twice a day at 00:00 UTC and 12:00 UTC, and several large-scale climate indices are used to
115 forecast streamflows. Reforecast data have been used for prediction with observed analogs
116 (Clark and Hay, 2004), studying atmospheric predictability, integrating into the numerical
117 weather prediction models, and diagnosing model bias where bias is the mean forecast minus
118 the mean verification (Hamill *et al.*, 2006). The pool of potential predictors important for

119 streamflow forecasting is listed in Table 1. The information for the phase of the seasonal
120 cycle is input as a pair of cyclical variables, namely $\sin(\text{phase}_s)$ and $\cos(\text{phase}_s)$.

121 The Niño 3.4 index is used to represent ENSO activity as it measures the central equato-
122 rial Pacific sea surface temperature (SST) anomalies (Burn, 2008). Monthly data were ob-
123 tained from the website <http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices>
124 and weekly data after 1990 from <http://www.cpc.ncep.noaa.gov/data/indices/wksst>.
125 for. The PDO time series was obtained from the website [ftp://ftp.atmos.washington.](ftp://ftp.atmos.washington.edu/mantua/pnw_impacts/INDICES/PDO.latest)
126 [edu/mantua/pnw_impacts/INDICES/PDO.latest](ftp://ftp.atmos.washington.edu/mantua/pnw_impacts/INDICES/PDO.latest). Linear interpolation was applied to the
127 monthly or weekly data to produce daily values for use in our daily streamflow forecasts.
128 Daily index values for PNA, AO, NAO and AAO (Antarctic Oscillation, characterized
129 by the first principal component of the 700 mb height anomalies poleward of 20°S) were
130 obtained from Climate Prediction Center (CPC), NOAA/National Weather Service via
131 <ftp://ftp.cpc.ncep.noaa.gov/cwlinks/>.

132 **3 Machine learning methods**

133 The history of learning machines can be classified into four periods (Vapnik, 1995): con-
134 structing (i) the first learning machines, (ii) the fundamentals of the theory, (iii) neural
135 networks, and (iv) the alternatives to neural networks. In this study, we will focus in
136 the latter two periods and introduce two alternatives to neural networks. In conventional
137 neural network (i.e. multi-layer perceptron) modelling (Hsieh, 2009), a network is trained
138 using a data set (\mathbf{x}, y) , where \mathbf{x} are the predictors and y the predictand (i.e. the response
139 variable), by adjusting network parameters or weights \mathbf{w} so as to minimize an objective

140 function $E(\mathbf{w})$ (also called a cost/loss/error function),

$$E(\mathbf{w}) = CE_D + E_w, \quad (1)$$

$$E_D = \frac{1}{2} \sum_{n=1}^N [y(\mathbf{x}^n; \mathbf{w}) - y_d^n]^2, \quad (2)$$

$$E_w = \frac{1}{2} \|\mathbf{w}\|^2, \quad (3)$$

141 where E_D is an error function measuring how close the network output $y(\mathbf{x}, \mathbf{w})$ is to the
142 target data y_d , E_w is the weight penalty term to prevent overfitting (i.e. fitting to the noise
143 in the data), N is the training sample size, with n the index for the training samples, and
144 C is the inverse weight penalty parameter (often called a hyperparameter). A small C will
145 strongly suppress the magnitude of \mathbf{w} found by the optimization process, thereby yielding
146 a less complex (i.e. less nonlinear) model. The best value for C is commonly chosen by
147 validating the model performance over independent data not used in training the model.
148 With the optimal C , the model should be neither overfitting nor underfitting the data.

149 An alternative to using validation to find the best value for C is to use a Bayesian
150 neural network (BNN) (MacKay, 1992, Bishop, 2006). The idea of BNN is to treat the
151 network weights as random variables, obeying an assumed prior distribution. Once ob-
152 served data are available, the prior distribution is updated to a posterior distribution using
153 Bayes' theorem. BNN automatically determines the optimal value of C without the need
154 of validation data (Bishop, 2006, Hsieh, 2009). BNN has some advantages: (a) error bars
155 (i.e. prediction intervals) can be automatically estimated for the model predictions, and
156 (b) since an automatic procedure is used for finding C , all available data can be used for
157 training, hence better models may result (van Hinsbergen *et al.*, 2009). The BNN code
158 `trainbr.m` provided by MATLAB was used.

159 Since neural network models solve for the model weights by using nonlinear optimization
160 algorithms to minimize the objective function, they are vulnerable to finding local minima
161 of the objective function. To alleviate the problem of multiple minima, 30 BNN models

162 were built with optimization starting from random initial weights, and the predictions
 163 of the 30 models are ensemble averaged to give the final prediction. Improvement in
 164 performance tends to level out as the ensemble size reaches around 25 (Breiman, 1996,
 165 Cannon and Whitfield, 2002), hence our choice of averaging 30 BNN models. The BNN
 166 models were built using data from 1983–1997 and their forecast performance was tested
 167 over the independent verification period of 1998–2001.

168 The foundations of support vector machines (SVM) have been developed and are gaining
 169 popularity due to many attractive features and promising empirical performance. SVMs
 170 were first developed to solve the classification problem, and then extended to regression
 171 problems (Vapnik *et al.*, 1996). In the support vector regression (SVR) algorithm, the
 172 input data \mathbf{x} are nonlinearly mapped into a higher dimensional feature space, in which the
 173 training data may exhibit linearity, so a linear regression problem is solved in this feature
 174 space, thereby circumventing the multiple minima problem of neural networks resulting
 175 from the nonlinear optimization (Bishop, 2006).

176 Performance of the SVR model depends on the choice of the kernel function K and the
 177 hyperparameters. In this study, we use the Gaussian or radial basis function (RBF) kernel
 178 with the hyperparameter σ controlling the width of the Gaussian function, with the kernel
 179 function

$$K(\mathbf{x}, \mathbf{x}_i) = \exp \left[-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2} \right]. \quad (4)$$

180 SVR with the RBF kernel is a nonlinear regression model but with the robust ϵ -insensitive
 181 error norm instead of the non-robust mean squared error norm as used in MLR. The ϵ -
 182 insensitive error norm is defined by

$$|y(\mathbf{x}; \mathbf{w}) - y_d|_\epsilon = \begin{cases} 0, & \text{if } |y(\mathbf{x}; \mathbf{w}) - y_d| < \epsilon \\ |y(\mathbf{x}; \mathbf{w}) - y_d| - \epsilon, & \text{otherwise,} \end{cases} \quad (5)$$

183 i.e. when the difference between y and y_d is smaller than ϵ , the error is ignored, whereas

184 when the difference between y and y_d is large, the error approximates the mean absolute
185 error, which is more robust to outliers in the data than the mean squared error.

186 We also tested three other types of kernel functions – the linear, the polynomial and
187 the sigmoidal kernel functions. We used the SVR codes from Chang and Lin (2001), down-
188 loadable from the LIBSVM website (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>). The
189 hyperparameters were initially estimated according to Cherkassky and Ma (2004), and then
190 optimized by genetic algorithms (GA) (Holland, 1992, Goldberg, 1989). For this purpose,
191 the GA optimization search is used for the SVR model (SVRGA) to find the optimal hy-
192 perparameters and the type of kernel function used (Wu *et al.*, 2009). Genetic algorithms
193 use techniques inspired by evolutionary biology, e.g. inheritance, mutation, selection, and
194 crossover. GA is well suited to manipulate the models with varying structures since it
195 can search non-linear solution spaces without requiring gradient information or a prior
196 knowledge of model characteristics. Cross-validation (i.e. separating data into training and
197 validation subsets) (Bishop, 2006) was performed over the data period used for model de-
198 velopment (i.e. 1983–1997) to determine the optimal hyperparameters and kernels. The
199 model forecast performance was then tested (Chiang *et al.*, 2004) over the verification pe-
200 riod of 1998–2001. In this study, we also tried the standard grid search method, but in
201 comparison to GA, the results were poorer and in some cases (e.g. small grid size) longer
202 computational times were needed.

203 Another kernel method which has become popular in recent years is the Gaussian process
204 (GP) model for nonlinear regression (Rasmussen and Williams, 2006, Bishop, 2006). Gaus-
205 sian process is a collection of random variables, any finite number of which have consistent
206 joint Gaussian distributions. A GP is fully specified by its mean function and covariance
207 function. This is a natural generalization of the Gaussian distribution, for which the mean
208 and the covariance are a vector and a matrix, respectively – i.e. the Gaussian distribution
209 is over vectors, whereas the Gaussian process is over functions (Bousquet *et al.*, 2004).
210 The Bayesian inference in BNN is only approximate, but is exact in the GP model. The

211 hyperparameters of GP were found by maximizing a likelihood function (Bishop, 2006),
212 in contrast to those of SVR which were found from cross-validation. The only problem in
213 GP application is its large memory and computational needs when the dataset size is large
214 (more than a few thousands points). The GP code of Rasmussen and Williams (2006) was
215 used in our study. The GP models were built using data from 1983–1997 and their forecast
216 performance was tested over the verification period of 1998–2001.

217 4 Streamflow forecast evaluations

218 Since the predictors belong to three categories, i.e. local observations, climate indices and
219 GFS model output (Table 1), the categories can be combined, and we tested four sets of
220 predictors, namely (i) local observations only, (ii) local observations and climate indices, (iii)
221 local observations and GFS output, and (iv) local observations, GFS output and climate
222 indices. Because the distribution of the streamflow predictand is highly skewed, data
223 are transformed by the natural logarithm prior to modelling. To discard irrelevant and
224 redundant predictors, the forward and backward stepwise regression method based on the
225 F test was used to select the relevant predictors for our models.

226 As data for soil moisture, soil temperature and snow depth in higher elevations are
227 usually not available, antecedent flow functions as a suitable memory of the watershed
228 system to capture the local and physiographic features. To indirectly incorporate the snow
229 budget influences on the short-term forecasts, antecedent and present streamflows, as well
230 as the maximum daily temperature, daily temperature range, and precipitation were used
231 in the models. The temperature range over a day is important as it controls the snowmelt
232 processes that affects the freezing and thawing of the snow pack.

233 The machine learning models BNN, SVRGA and GP, together with MLR, are used
234 to determine which of the 4 sets of predictors is most effective in streamflow forecasting.
235 Verification of the forecasts by the linear and nonlinear models is carried out using the

236 Pearson correlation coefficient (Corr), mean absolute error (MAE), root mean squared error
 237 (RMSE), and the Nash-Sutcliffe model efficiency coefficient (NSE) (McCuen *et al.*, 2006)
 238 for all the forecasts. For the binary forecast of extreme (high) events in the streamflow,
 239 the Peirce skill score (PSS) and the extreme dependency score (EDS) are used (Stephenson
 240 *et al.*, 2008). The flowchart in Figure 2 summarizes our approach.

241 While Corr is a good measurement of linear association between the forecast values
 242 and the observed values, it does not take forecast bias into account, and is sensitive (i.e.
 243 non-robust) to rare events. Compared to the RMSE, MAE, which measures the average
 244 magnitude of the forecast errors, is a cleaner, hence preferable, measure of the average error
 245 (Willmott and Matsuura, 2005). These scores for forecast lead times of 1 to 7 days are
 246 analyzed for all models built on different predictor data sets.

247 A skill score measures the accuracy of the forecasts y made by a model relative to that
 248 of a reference model. The reference model usually employs a naive forecasting method, such
 249 as climatology or persistence. A climatological forecast y_c simply issues the climatological
 250 mean of the target data y_d at that time of year, while a persistence forecast y_p simply issues
 251 the value of y_d (at time t when the forecast is made) for all future times ($t + 1, t + 2, \dots$).
 252 The NSE uses y_c as reference, i.e.

$$\text{NSE} = 1 - \frac{\sum (y - y_d)^2}{\sum (y_c - y_d)^2}. \quad (6)$$

253 Note $\text{NSE} = 1$ for a perfect model, and $\text{NSE} < 0$ when the model forecasts are worse than
 254 those from the reference model. Optimal linear combination of climatology and persistence
 255 is considered a good method to produce more accurate forecasts (Murphy, 1992), i.e.

$$y' = ky_p + (1 - k)y_c, \quad (7)$$

256 where the constant k ($0 \leq k \leq 1$) is derived from the autocorrelation of y_d . When y' is

257 used as the reference, the NSE becomes

$$\text{NSE} = 1 - \frac{\sum (y - y_d)^2}{\sum (y' - y_d)^2}. \quad (8)$$

258 To verify model skill at forecasting extreme events, PSS and EDS are defined by

$$\text{PSS} = \frac{a}{a + c} - \frac{b}{b + d}, \quad (9)$$

259

$$\text{EDS} = \frac{2 \ln(\frac{a+c}{n})}{\ln(\frac{a}{n})} - 1, \quad (10)$$

260 where a denotes the number of occasions in which the extreme event was forecast and
261 observed (hits), b the number of times the event was forecast but not observed (false
262 alarms), c the number of times the event was observed but not forecast (misses) and d the
263 number of times the event was neither forecast nor observed (correct rejections), with the
264 total number of cases being $n = a + b + c + d$. Both PSS and EDS scores range from -1
265 (worst forecast) to 1 (perfect forecast).

266 5 Results and discussion

267 As mentioned before, we tested four combinations of predictors, with the first combination
268 involving local observations only, i.e. T_{\max} (the daily maximum temperature), $T_{\max-\min}$ (the
269 difference between the daily maximum and minimum temperatures), precipitation P and
270 streamflow Q , all at time t , plus Q at the preceding day $t - 1$. Other combinations involve
271 predictors from the GFS outputs and the climate indices (Table 1). The MLR and machine
272 learning models were trained using data from 1983–1997, and their forecasts verified (i.e.
273 tested) using data from 1998–2001.

274 Four performance scores, i.e. Corr, MAE, RMSE, and NSE, evaluate the model forecasts
275 during the verification period. These scores for forecast lead times of 1 to 7 days are

276 illustrated in Figure 3 for the case with all potential predictors supplied (then selected by
277 stepwise linear regression). As shown, the three nonlinear models generally perform better
278 than MLR, with the BNN model (which uses the average of 30 ensemble members) tending
279 to attain slightly better scores (higher Corr and NSE, and lower MAE and RMSE) than the
280 other models, especially at the longer lead times. As $NSE < 0$ indicates the model forecasts
281 are worse than those from the reference model using climatological forecasts, MLR was able
282 to outperform climatology forecasts up to a lead time of 3 days, while the nonlinear models
283 were able to surpass climatology forecasts up to 5 days (Figure 3d).

284 Since BNN tended to slightly outperform SVRGA and GP in Figure 3, the BNN model
285 was selected to be trained using the four combinations of predictors. The NSE score in
286 Figure 4d shows that the case of using the local observations and GFS output as predictors
287 tended to be the best among the four at short lead times (1–4 days). At longer lead times
288 (5–7 days), climate indices are more helpful than GFS output, as seen in the Corr and
289 MAE scores in Figure 4, where having local observations and climate indices as predictors
290 tended to outperform the case of using local observations and GFS output. This shows the
291 advantage of incorporating climate signals in longer lead forecasts when the GFS output is
292 no longer accurate.

293 The climate indices selected as predictors through stepwise regression in streamflow
294 forecasting for lead times of 1–7 days (Table 2) are those for the El Niño-Southern Oscilla-
295 tion (ENSO), Pacific-North American teleconnection (PNA), Arctic Oscillation (AO), and
296 North Atlantic Oscillation (NAO). Seasonal streamflow is affected by the interannual vari-
297 ability of the accumulated snow in the watershed, associated with ENSO and other modes
298 of climate variability. Composites of the snow water equivalent anomalies (SWEA) showed
299 that SWEA were negative during high PNA years and El Niño years and positive during
300 La Niña years and low PNA years in the Columbia River Basin in British Columbia (Hsieh
301 and Tang, 2001). Coulibaly and Burn (2005) found correlation between seasonal stream-
302 flow and climate indices of ENSO in western and eastern Canada, and between streamflow

303 and NAO in western Canada for the period after 1950 for all seasons. Brown and Goodison
304 (1996) revealed that the North American winter snow cover has gradually increased in last
305 century and the spring snow cover has decreased. This change was related to the PNA
306 and to some extent ENSO and NAO. Shabbar and Bonsal (2004) found that ENSO and
307 AO influenced the frequency and duration of both cold spells and warm spells in Canadian
308 winter temperatures.

309 The forecasted streamflow by the BNN and MLR models are compared with the ob-
310 served data at 1-day lead time in Figure 5 and at 7-day lead time in Figure 6 for 400 days
311 at the end of the verification period 1998–2001. In Figure 5, both MLR and BNN forecasts
312 show generally good agreement with the observed streamflow in the study area, although
313 for some of the extreme flood events, the two models did not perform very well. Rapid tem-
314 perature change during short periods and consequent snowmelt, and rain on snow events
315 can be the reasons for the underestimation of extreme events by our models in the region.

316 Prediction intervals are calculated to estimate the uncertainty in the model predictions.
317 For the BNN ensemble model forecasts, the prediction intervals are computed according
318 to van Hinsbergen *et al.* (2009). For computing the MLR prediction interval, a Gaussian
319 distribution with conditional mean equal to the MLR predictions and constant variance
320 equal to the variance of the MLR residuals is assumed. The percentage of observed points
321 lying outside the 95 % prediction intervals for forecast lead times of 1 to 7 days for the BNN
322 model are 13.8, 3.4, 6.6, 8.1, 2.9, 3.7, and 2.9%, respectively. The corresponding values are
323 5.2, 5.0, 4.5, 3.5, 3.2, 3.2, and 3.1% for the MLR model. These percentages are not far from
324 the theoretical value of 5%, indicating that the prediction intervals have been estimated
325 reasonably well (Gneiting *et al.*, 2007) for both BNN and MLR. The average width of the
326 prediction intervals, which provides a measure of “sharpness” (Gneiting *et al.*, 2007), is 42.4
327 m^3s^{-1} for the BNN model and 54.4 m^3s^{-1} for the MLR, for 1-day lead forecasts, suggesting
328 that BNN is better than MLR in terms of probabilistic performance.

329 As mentioned earlier, there are several choices for the reference model when calculating

330 a skill score such as the NSE. In Figure 7, for a given lead time, when a particular reference
331 model is forecasting poorly, the NSE skill score (relative to the poor reference model) is high.
332 Hence for the lead time of 1 day, persistence forecasts outperform climatology forecasts,
333 since NSE with reference to persistence is lower than NSE with reference to climatology.
334 However, for lead times of 2–7 days, climatology forecasts outperform persistence. The
335 optimal linear combination of climatology and persistence outperforms both climatology
336 and persistence for lead times of 1–4 days, but is poorer than climatology for lead times of
337 5–7 days.

338 To measure how good the models are in forecasting the occurrence of extreme events in
339 streamflow, two skill score indices are used: Peirce Skill Score (PSS) and Extreme Depen-
340 dency Score (EDS) (Stephenson *et al.*, 2008). Streamflow values above the 95% percentile
341 in the streamflow data are considered to be the “extreme” events. Before calculating the
342 scores, since the predictions from the models tended to underestimate the amplitude of the
343 streamflow (see Figures 5 and 6), the forecast values were rescaled by matching the cumu-
344 lative distribution of the predicted streamflow to that of the observed streamflow based on
345 the training data.

346 Figure 8 illustrates the PSS and EDS for different combinations of predictors. For
347 linear and nonlinear models, using local observations and GFS output as predictors tended
348 to yield higher PSS and EDS than other choices of predictors. In contrast to Figure 4, the
349 addition of climate indices as predictors was not helpful at long lead times, indicating that
350 although climate indices contributed to the overall streamflow forecast scores (Corr, MAE,
351 RMSE and NSE), they failed to contribute in the extreme event scores like PSS and EDS.
352 At shorter lead times, BNN was performing better than MLR (Figure 8), but at 7-day lead
353 time, the advantage of BNN has disappeared, suggesting that at long lead time the signal
354 is too weak to allow the complicated nonlinear method to improve on the simple MLR. The
355 BNN model tended to slightly outperform the other nonlinear models (SVRGA and GP)
356 except for SVRGA at 3-day lead time (not shown) .

357 One may ponder why climate indices which are related to *seasonal* conditions are helpful
358 in *daily* streamflow forecasts at lead times of 5–7 days, but are not helpful for extreme
359 streamflow forecasts. Consider a La Niña winter where there is more accumulated snow in
360 the mountains, then the average daily streamflow in spring/summer will also be enhanced,
361 so a machine learning model using the information from an ENSO index to forecast a higher
362 average daily streamflow value will have higher skill than one which does not use the ENSO
363 index. In other words, using seasonal conditions to forecast slightly higher or lower average
364 daily streamflow values does improve daily forecast skills, but the effect is too small to
365 improve the forecasts of extreme events.

366 **6 Conclusions**

367 In this study, local observations, GFS reforecast output (as representing information from
368 weather forecasts) and climate indices were used as predictors for streamflow forecasting at
369 lead times of 1–7 days using three nonlinear machine learning methods and MLR. Machine
370 learning methods used in this study were the Bayesian neural network (BNN), support
371 vector regression (with genetic algorithm used for selecting hyperparameters and kernels)
372 (SVRGA), and Gaussian process (GP). The multiple local minima problem of BNN was
373 alleviated by using the average forecast of an ensemble of BNN models.

374 In terms of forecast scores, the nonlinear models generally outperformed MLR, and
375 BNN tended to slightly outperform the other nonlinear models. Local observations plus
376 the GFS output as predictors were generally best at shorter lead times, while local obser-
377 vations plus climate indices were best at longer lead times. Climate indices selected include
378 the equatorial Pacific SST in the Niño 3.4 region, the Pacific-North American (PNA) tele-
379 connection, the Arctic Oscillation (AO) and the North Atlantic Oscillation (NAO). For
380 both linear and nonlinear models, the PSS and EDS scores used to evaluate the binary
381 forecasts for extreme (high) streamflow events were best when the predictors were the local

382 observations plus GFS output. Climate indices as extra predictors were not beneficial in
383 the extreme event forecasts, even at the longer lead times. Hence climate indices contribute
384 to daily streamflow forecast scores during longer lead times of 5–7 days, but not to forecast
385 scores for extreme streamflow events for all lead times studied (1–7 days).

386 **Acknowledgements**

387 We are grateful to Dr. Ziad Shawwash for his helpful comments. The support from the
388 Natural Sciences and Engineering Council of Canada and the Canadian Foundation for
389 Climate and Atmospheric Sciences is gratefully acknowledged.

References

- 390 Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*. Springer, New York.
- 391 Bousquet, O., von U. Luxburg, and G. Ratesch, 2004. *Advanced Lectures on Machine*
392 *Learning*. Springer, Berlin, Heidelberg New York.
- 393 Breiman, L., 1996. Bagging predictions. *Machine Learning*, 24:123–40.
- 394 Brown, R. D. and B. E. Goodison, 1996. Interannual variability in reconstructed Canadian
395 snow cover, 1915-1992. *Journal of Climate*, 9(6):1299–1318.
- 396 Burn, D. H., 2008. Climatic influences on streamflow timing in the headwaters of the
397 Mackenzie River Basin. *Journal of Hydrology*, 352(1-2):225 – 238.
- 398 Cannon, A. J. and P. H. Whitfield, 2002. Downscaling recent streamflow conditions in
399 British Columbia, Canada using ensemble neural network models. *J. Hydrology*, 259(1-
400 4):136–51.
- 401 Chang, C.-C. and C.-J. Lin, 2001. LIBSVM: A library for support vector machines. Software
402 available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- 403 Cherkassky, V. and Y. Ma, 2004. Practical selection of SVM parameters and noise estima-
404 tion for SVM regression. *Neural Netw.*, 17(1):113–126.
- 405 Chiang, Y. M., L. C. Chang, and F. J. Chang, 2004. Comparison of static-feedforward
406 and dynamic-feedback neural networks for rainfall-runoff modeling. *J. Hydrology*, 290(3-
407 4):297–311.
- 408 Clark, M. P. and L. E. Hay, 2004. Use of medium-range numerical weather prediction
409 model output to produce forecasts of streamflow. *J. Hydrometeorol*, 5:15–32.
- 410 Coulibaly, P. and D. H. Burn, 2005. Spatial and temporal variability of Canadian seasonal
411 streamflows. *Journal of Climate*, 18(1):191–210.
- 412

- 413 Fleming, S. W., R. D. Moore, and G. K. C. Clarke, 2006. Glacier-mediated streamflow
414 teleconnections to the Arctic oscillation. *International Journal of Climatology*, 26:619–
415 636.
- 416 Fleming, S. W., P. H. Whitfield, R. D. Moore, and E. J. Quilty, 2007. Regime-dependent
417 streamflow sensitivities to Pacific climate modes cross the Georgia-Puget transboundary
418 eco-region. *Hydrological Processes*, 21(24):3264–3287.
- 419 Gneiting, T., F. Balabdaoui, and A. E. Raftery, April 2007. Probabilistic forecasts, cali-
420 bration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69:243–268.
- 421 Goldberg, D. E., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*.
422 Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- 423 Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006. Reforecasts: an important dataset
424 for improving weather predictions. *Bulletin of the American Meteorological Society*,
425 87(1):33–46.
- 426 Holland, J. H., 1992. *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge,
427 MA, USA.
- 428 Hsieh, W. W., 2009. *Machine Learning Methods in the Environmental Sciences - Neural*
429 *Networks and Kernels*. Cambridge.
- 430 Hsieh, W. W. and B. Tang, 2001. Interannual variability of accumulated snow in the
431 Columbia Basin, British Columbia. *Journal of Water Resources Research*, 37(6):1753–
432 1759.
- 433 Hurrell, J. W., 1995. Decadal trends in the North Atlantic oscillation: regional temperatures
434 and precipitation. *Science*, 269:676–679.
- 435 Kennedy, A. M., D. C. Garen, and R. W. Koch, 2009. The association between climate

- 436 teleconnection indices and upper Klamath seasonal streamflow: trans-niño index. *Hy-*
437 *drological Processes*, 23:973–984.
- 438 MacKay, D. J. C., 1992. Bayesian interpolation. *Neural Comput.*, 4(3):415–447.
- 439 Makkeasorn, A., N. Chang, and X. Zhou, 2008. Short-term streamflow forecasting with
440 global climate change implications - a comparative study between genetic programming
441 and neural network models. *Journal of Hydrology*, 352(3-4):336 – 354.
- 442 Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997. A pacific
443 interdecadal climate oscillation with impacts on salmon production. *Bulletin of the*
444 *American Meteorological Society*, 78(6):1069–1079.
- 445 McCuen, R. H., Z. Knight, and A. G. Cutter, 2006. Evaluation of the Nash–Sutcliffe
446 efficiency index. *Journal of Hydrologic Engineering*, 11(6):597–602.
- 447 Moradkhani, H., K. Hsu, H. V. Gupta, and S. Sorooshian, 2004. Improved streamflow
448 forecasting using self-organizing radial basis function artificial neural networks. *Journal*
449 *of Hydrology*, 295:246 – 262.
- 450 Murphy, A. H., 1992. Climatology, persistence, and their linear combination as standards
451 of reference in skill scores. *Weather and Forecasting*, 7(4):692–698.
- 452 Rasmussen, C. and C. K. I. Williams, 2006. *Gaussian Processes for Machine Learning*.
453 MIT Press, Cambridge, MA, USA.
- 454 Shabbar, A. and B. Bonsal, 2004. Associations between low frequency variability modes
455 and winter temperature extremes in Canada. *Atmos.-Ocean*, 42:127–40.
- 456 Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008. The extreme
457 dependency score: a non-vanishing measure for forecasts of rare events. *Meteorological*
458 *Applications*, 15(1):41–50.

- 459 Thompson, D. W. J. and J. Wallace, 1998. The Arctic Oscillation signature in the winter-
460 time geopotential height and temperature fields. *Geophys. Res. Lett.*, 25:1297–300.
- 461 van Hinsbergen, C., J. van Lint, and H. van Zuylen, 2009. Bayesian committee of neural
462 networks to predict travel times with confidence intervals. *Transportation Research Part*
463 *C: Emerging Technologies*, 17(5):498 – 509.
- 464 Vapnik, V., S. E. Golowich, and A. J. Smola, 1996. Support vector method for func-
465 tion approximation, regression estimation and signal processing. In *Advances in Neural*
466 *Information Processing Systems*, volume 9, pages 281–287.
- 467 Vapnik, V. N., 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York,
468 Inc., New York, NY, USA.
- 469 Willmott, C. and K. Matsuura, 2005. Advantages of the mean absolute error (MAE) over
470 the root mean square error (RMSE) in assessing average model performance. *Climate*
471 *Research*, 30(1):79–82.
- 472 Wu, C.-H., G.-H. Tzeng, and R.-H. Lin, 2009. A novel hybrid genetic algorithm for kernel
473 function and parameter optimization in support vector regression. *Expert Syst. Appl.*,
474 36(3):4725–4735.

Table 1: List of potential predictors. These are of three types, i.e. local observations, climate indices and GFS outputs. The GFS outputs are available twice a day at 00:00 UTC and 12:00 UTC.

No.	Variable	Index	Unit	Type
1	maximum temperature	T_{\max}	$^{\circ}\text{C}$	Obs.
2	temperature range	$T_{\max-\min}$	$^{\circ}\text{C}$	Obs.
3	precipitation	P	mm	Obs.
4	streamflow at day t	Q_t	m^3s^{-1}	Obs.
5	streamflow at day $t - 1$	Q_{t-1}	m^3s^{-1}	Obs.
6	seasonal phase	$\sin(\text{phase}_s), \cos(\text{phase}_s)$	-	-
7	Arctic Oscillation	AO	-	Clim.
8	Antarctic Oscillation	AAO	-	Clim.
9	North Atlantic Oscillation	NAO	-	Clim.
10	Pacific North American Pattern	PNA	-	Clim.
11	sea surface temperature (Niño 3.4)	SST	-	Clim.
12	Niño 3.4 anomaly	SSTA	-	Clim.
13	Pacific Decadal Oscillation	PDO	-	Clim.
14	accumulated precipitation	apcp	mm	GFS
15	heating	heating	J	GFS
16	air pressure at 500m	z_{500}	pa	GFS
17	mean sea level pressure	prmsl	pa	GFS
18	precipitable water	p_{wat}	-	GFS
19	relative humidity	rhum	%	GFS
20	2m air temperature	$t_{2\text{m}}$	K	GFS
21	temperature	temp ₇₀₀	K	GFS
22	wind amplitude at 10m	wind amp	ms^{-1}	GFS
23	wind phase at 10m	$\sin(\text{phase}), \cos(\text{phase})$	-	GFS

Table 2: Climate indices selected for the case with all potential predictors included, for lead times of 1–7 days.

Lead-time (day)	Climate indices
1	SST
2	PNA, SST
3	AO, PNA, SST
4	AO, PNA, SST
5	AO, SST
6	NAO, SST
7	NAO, SST

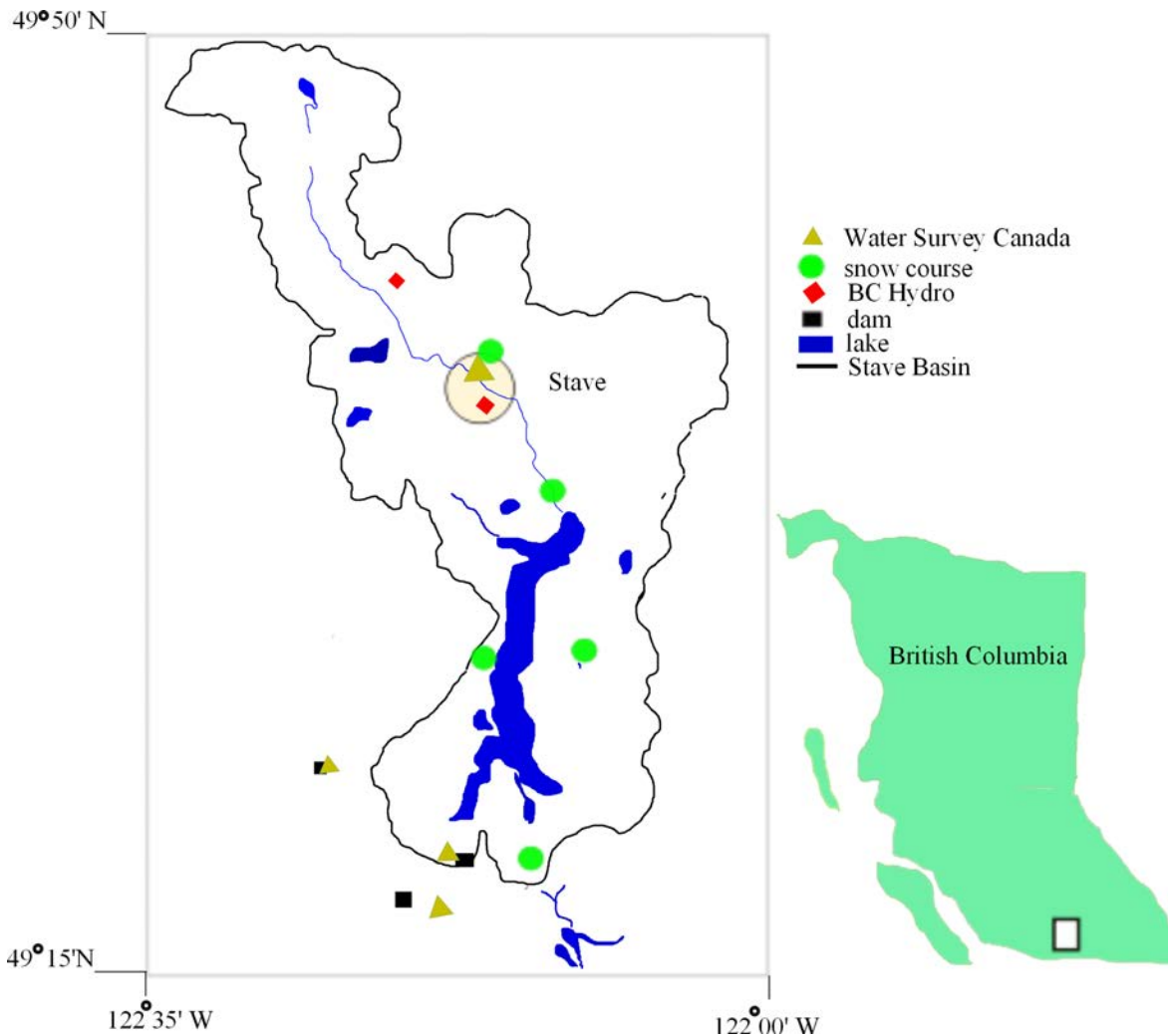


Figure 1: The Stave River Basin in southern British Columbia, Canada, with the inset map on the left. The black curve marks the boundary of the river basin. The golden triangles show the network of hydrometric stations run by the Water Survey of Canada, while the red diamonds show the hydro-meteorological stations of B.C. Hydro. The circular curve pinpoints the data used in this study, with streamflow measured at the triangle (hydrometric station 08MH147), and temperature and precipitation at the diamond (station Stave R. Upper DCP, where DCP = Data Collection Platform). Also indicated on the map are the dams (black squares) and the snow courses run by B.C. River Forecast Centre (green circles), where snow depth and snow water equivalent are measured.

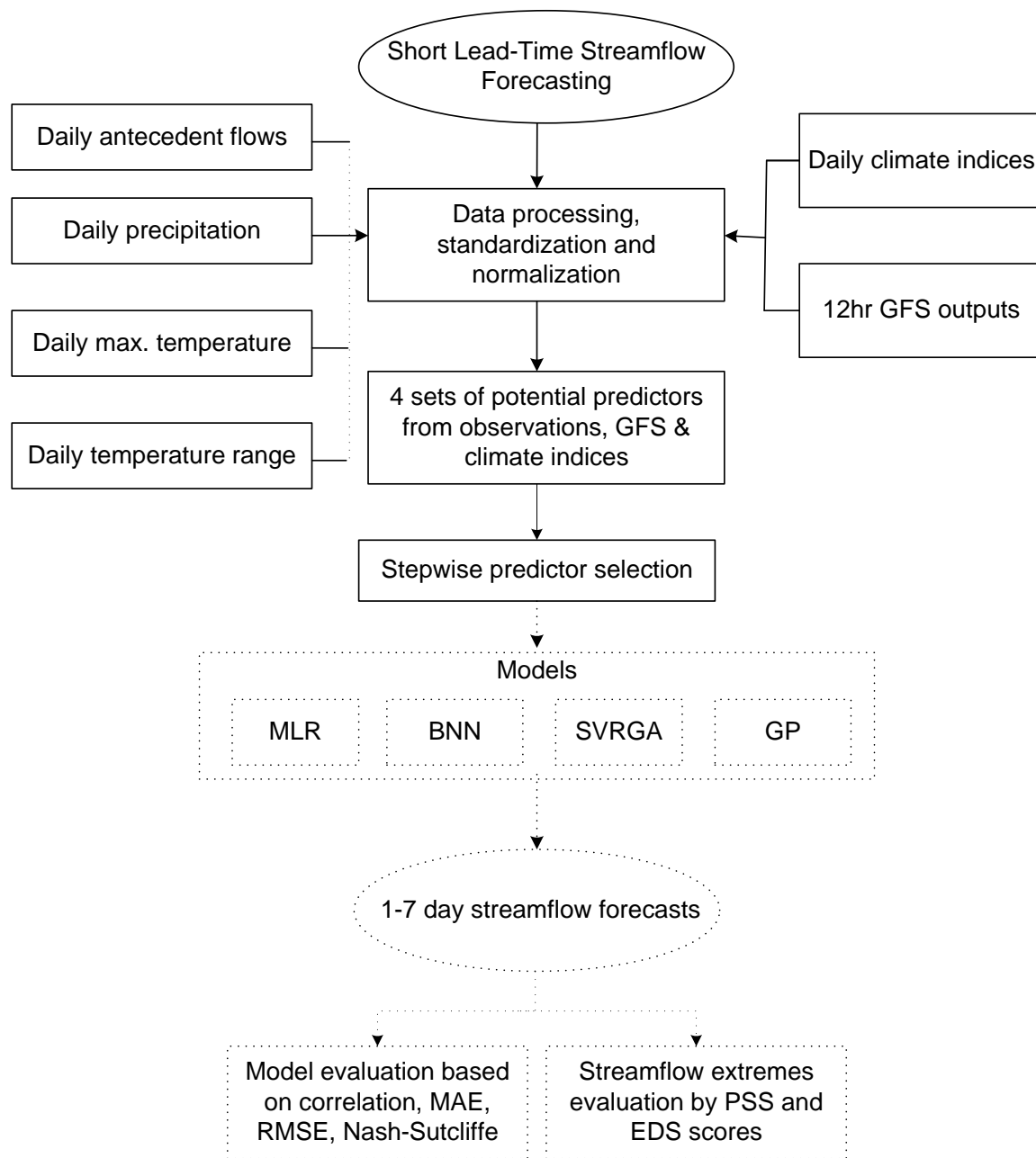


Figure 2: Flowchart of the streamflow forecasting. The 4 sets of potential predictors are (i) observations, (ii) observations and climate indices, (iii) observations and GFS output, and (iv) observations, climate indices and GFS output.

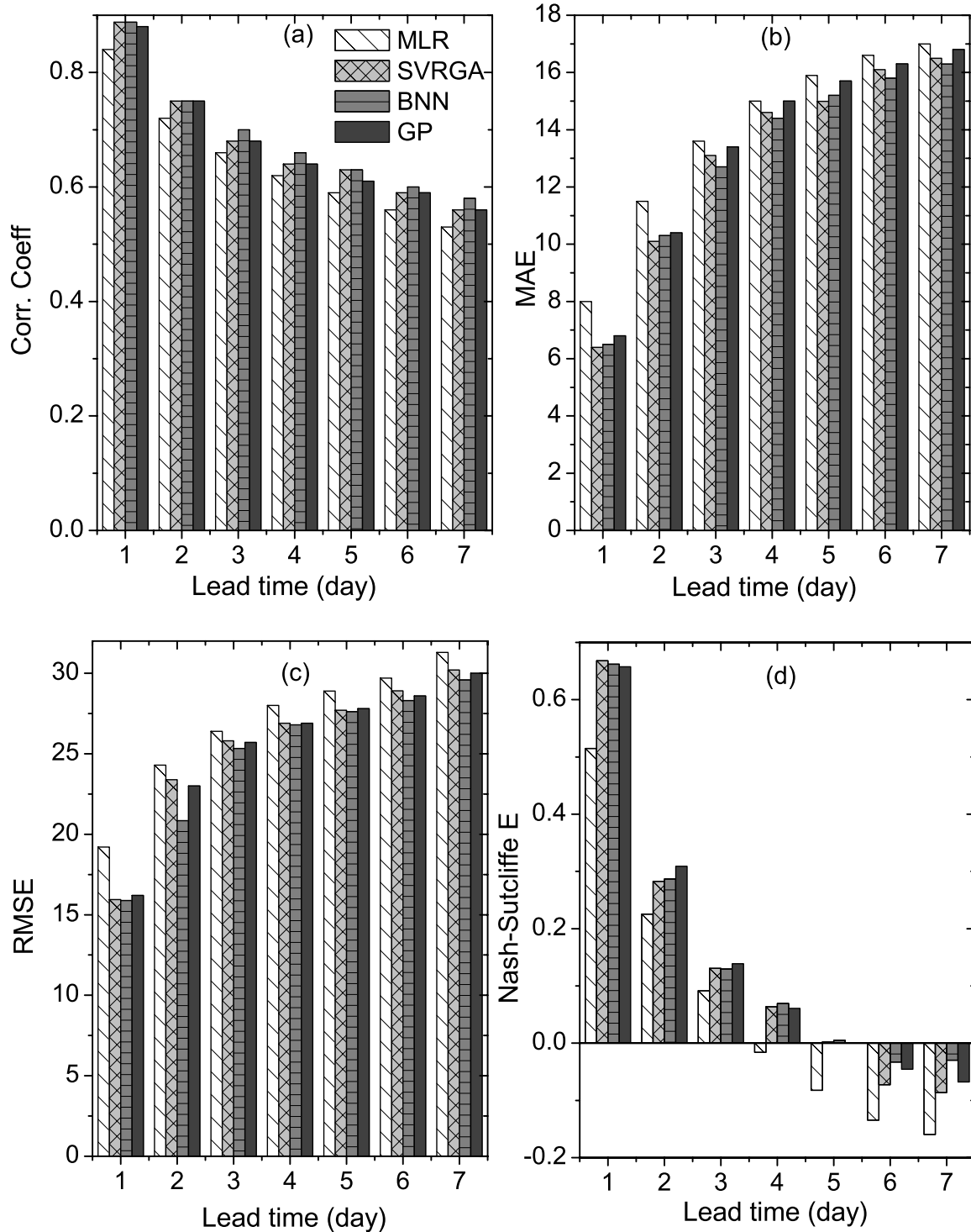


Figure 3: Streamflow forecast scores computed over the verification (i.e. test) period of 1998–2001 for MLR and three non-linear machine learning models (SVRGA, BNN and GP) with all potential predictors considered, as evaluated by the (a) Pearson correlation coefficient, (b) mean absolute error, (c) root mean squared error, and (d) Nash-Sutcliffe efficiency coefficient (with climatology as reference).

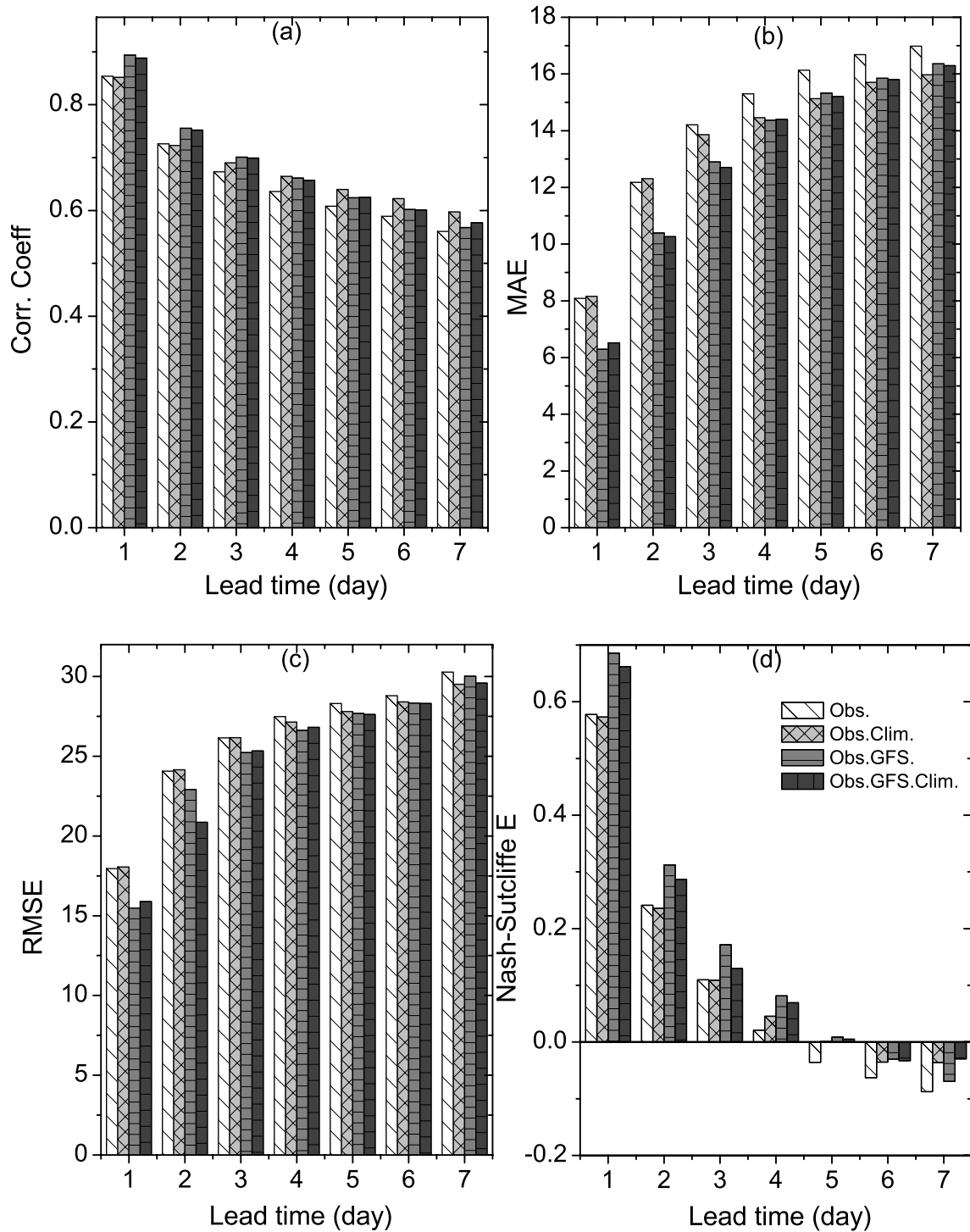


Figure 4: Streamflow forecast scores (computed over the verification period) for the BNN model for the 4 cases where predictors are (i) local observations only, (ii) local observations and climate indices, (iii) local observations and GFS output, and (iv) local observations, GFS output and climate indices, as evaluated by the (a) Pearson correlation coefficient, (b) mean absolute error, (c) root mean squared error, and (d) Nash-Sutcliffe efficiency.

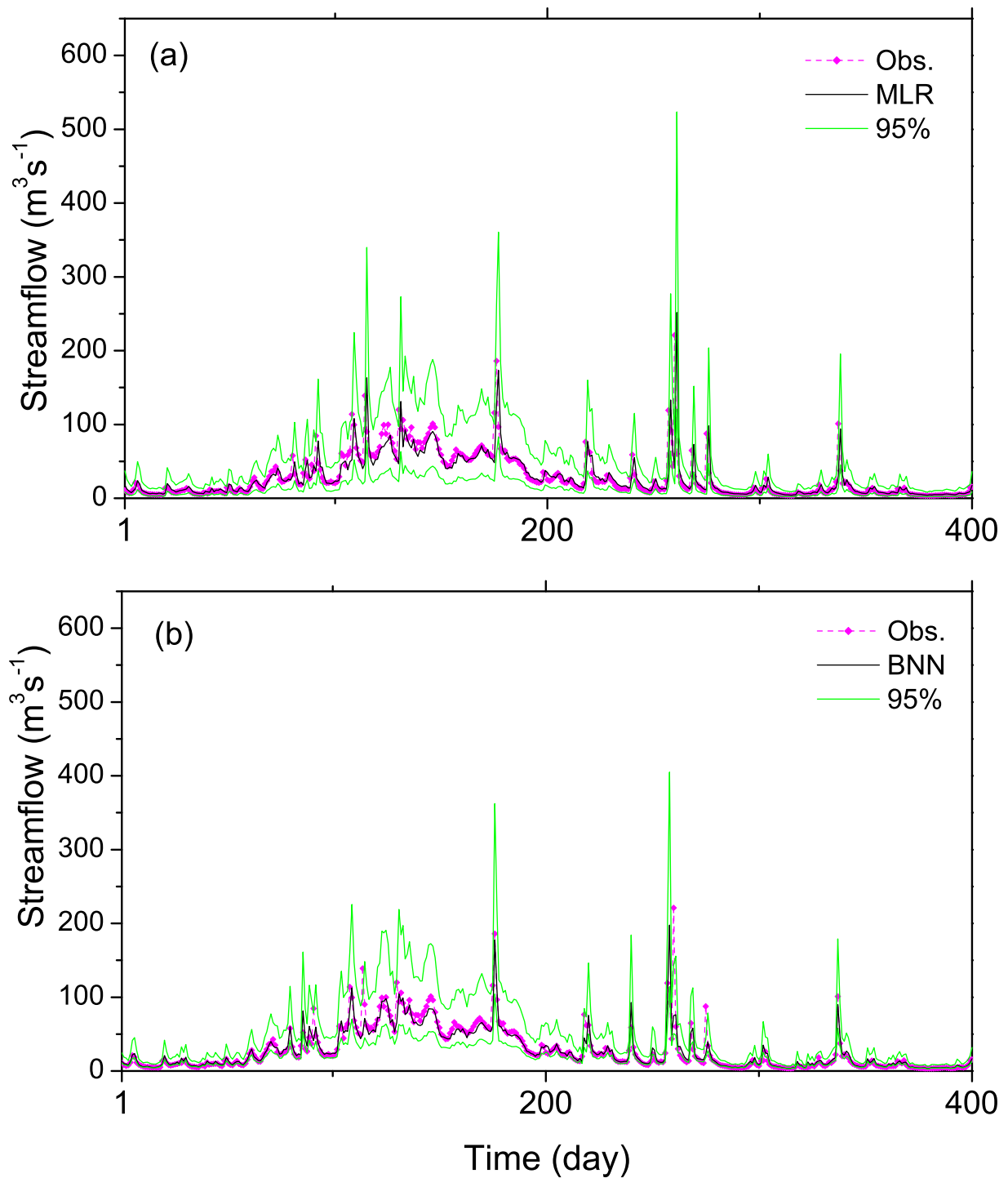


Figure 5: One-day lead time streamflow forecasts by the (a) MLR and (b) BNN models with all predictors considered, for 400 days (mainly covering the year 2001 during the verification period). The magenta diamonds show the observed streamflow, the solid black line shows the model forecasts, and light green lines show the 95% prediction intervals.

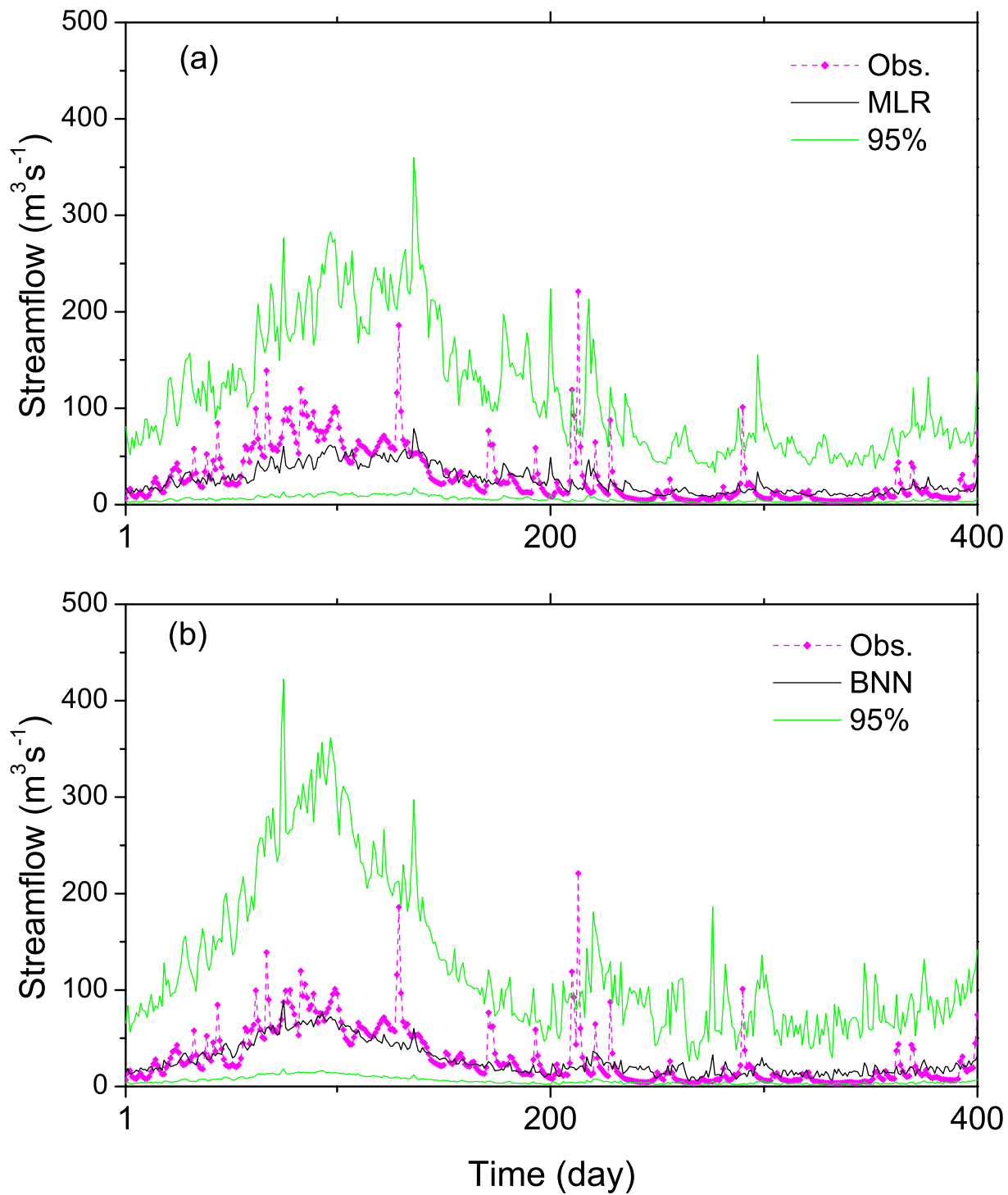


Figure 6: Seven-day lead time streamflow forecasts by the (a) MLR and (b) BNN models with all predictors considered. The magenta diamonds show the observed streamflow, the solid black line shows the model forecasts, and light green lines show the 95% prediction intervals.

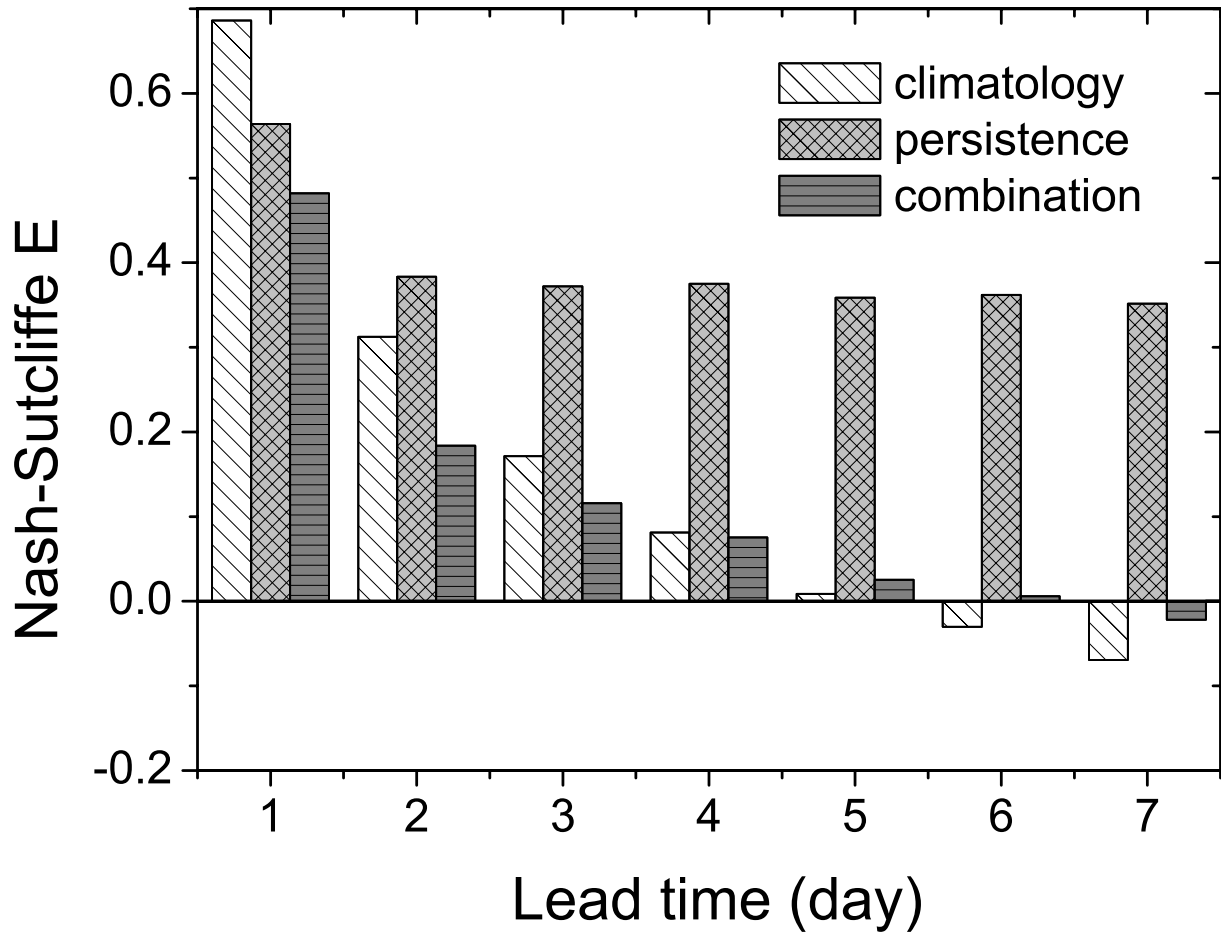


Figure 7: The Nash-Sutcliffe efficiency skill score (computed over the verification period), with the reference model being climatology, persistence, or the optimal combination of both. Only local observations and GFS output were considered for predictors.

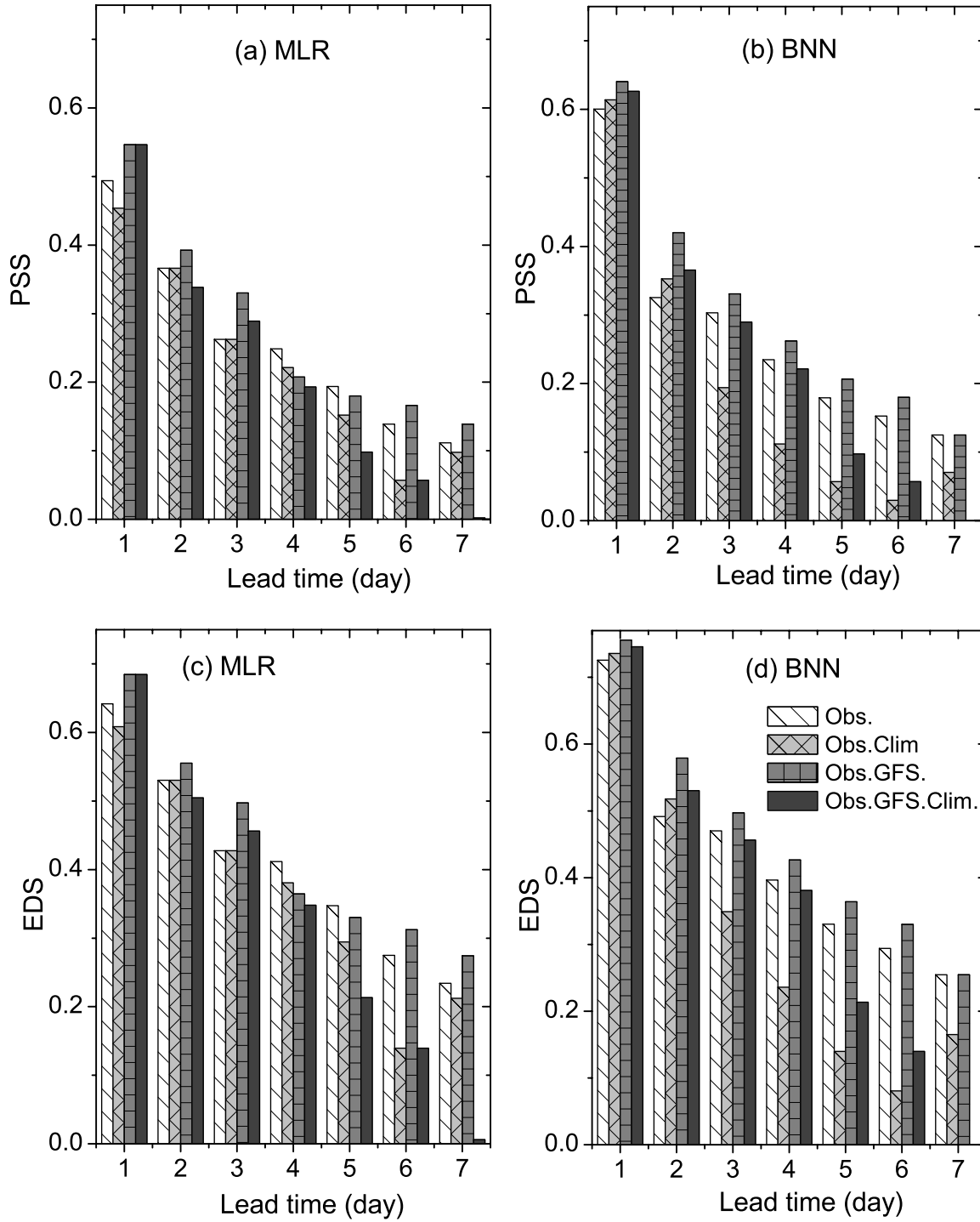


Figure 8: PSS forecast scores for extreme streamflow events from (a) MLR and (b) BNN, and EDS forecast scores from (c) MLR and (d) BNN, at lead times of 1–7 days over the verification period, where the predictors were (i) local observations only, (ii) local observations and climate indices, (iii) local observations and GFS output, and (iv) local observations, GFS output and climate indices.