Supplementary Materials for

Seasonal hydroclimatic ensemble forecasts anticipate nutrient and suspended sediment loads using a dynamical-statistical approach

Sanjib Sharma¹, Heather Gall², Jorge Gironás³⁴⁵⁶, and Alfonso Mejía¹

¹Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA
²Department of Agricultural and Biological Engineering, The Pennsylvania State University, University Park, PA
³Departamento de Ingeniería Hidráulica y Ambiental, Pontificia Universidad Católica de Chile,
⁴Centro Nacional de Investigación para la Gestión Integrada de Desastres Naturales (CIGIDEN)
⁵Centro de Desarrollo Urbano Sustentable (CEDEUS)
⁶Centro Interdisciplinario de Cambio Global, Pontificia Universidad Católica de Chile

The Supplementary Material section includes:

Supplementary Text 1: Statistical preprocessor
Supplementary Text 2: Spatially distributed hydrological model
Supplementary Text 3: Statistical postprocessor
Supplementary Figure 1: Map of study the area showing the gaged locations and the land cover distribution. The call out boxes show the NSE values for the streamflow, total nitrogen, total phosphorus, and total suspended sediment load simulations. The inset map shows the location of the basin in the northeast US.
Supplementary Figure 2: Time series plots of monthly accumulated streamflow for the selected forecast locations.
Supplementary Figure 3: Time series plots of monthly total nitrogen load for the selected forecast locations.
Supplementary Figure 4: Time series plots of monthly total phosphorous load for the selected forecast locations.
Supplementary Figure 5: Time series plots of monthly total suspended sediment load for the selected forecast locations.

References to Supplementary Materials

Supplementary Text 1: Statistical preprocessor

We use logistic regression to preprocess the CFSv2 forecasts (Messner et al 2014a-b, Yang et al 2017). Different logistic regression models are used for near-surface temperature and precipitation forecasts to take into consideration that precipitation is a censored variable, i.e., it can only take on values ≥ 0. To explain the logistic regression models used, we start with the model developed by Hamill et al (2004) which is given by

\[ P(y \leq q \mid x) = \Lambda[\delta(x)] \]

where \( \Lambda(.) \) denotes the cumulative distribution function of the standard logistic distribution, \( y \) is the forecast variable, \( q \) is a specified threshold, \( x \) is a predictor variable that depends on the forecast members, and \( \delta(x) \) is a linear function of the predictor variable \( x \).

Equation (1) requires separate logistic regressions to be fitted to each threshold of interest (Wilks 2009). This results in logistic regressions that can cross each other which, in turn, implies the occurrence of nonsense negative probabilities. To avoid these inconsistencies, Wilks (2009) included a transformation of the predictand thresholds as an additional predictor variable in the logistic regression

\[ P(y \leq q \mid x) = \Lambda[\phi(q) - \delta(x)] \]

where \( \phi(.) \) is a monotone nondecreasing function. In addition to avoiding negative probabilities, equation (2) has the advantage that fewer parameters need to be estimated; instead of having a linear function \( \delta(x) \) for each threshold, \( \delta(x) \) is now the same for all the thresholds. This is especially advantageous for small training datasets.

Furthermore, to appropriately utilize the uncertainty information in the ensemble spread, Messner et al (2014a) proposed the heteroscedastic extended logistic regression (HXLR) postprocessor. HXLR uses an additional predictor vector \( \phi \) to control the dispersion of the logistic predictive distribution,

\[ P(y \leq q \mid x) = \Lambda\left\{ \frac{\phi(q) - \delta(x)}{\exp[\eta(\phi)]} \right\}, \]

where \( \eta(.) \) is another linear function of the predictor variable \( \phi \).
The functions $\delta(.)$ and $\eta(.)$ are defined as
\[
\delta(x) = a_0 + a_1 x, \\
\eta(\phi) = b_0 + b_1 \phi,
\]
where $a_0$, $a_1$, $b_0$, and $b_1$ are parameters that need to be estimated.

Maximum likelihood estimation with the log-likelihood function is used to estimate the parameters associated with equation (3) (Messner et al 2014a-b). One variation of the HXLR that can accommodate nonnegative variables that are continuous for positive values and have a natural threshold at zero, such as precipitation amounts, is heteroscedastic censored logistic regression (HCLR) (Messner et al 2014b). To implement the HCLR model, the predicted probability or likelihood $\pi_i$ of the $i^{th}$ observed outcome, accounting for the fact that $y \geq 0$, is determined as follows
\[
\pi_i = \begin{cases} 
\lambda \left[ \frac{a(0) - \delta(x)}{\exp[\eta(\phi)]} \right], & y_i = 0 \\
\lambda \left[ \frac{a(y_i) - \delta(x)}{\exp[\eta(\phi)]} \right], & y_i > 0,
\end{cases}
\]
where $\lambda[.]$ denotes the likelihood function of the standard logistic function. In essence, HCLR fits a logistic error distribution with point mass zero to the transformed predictand. To transform the precipitation ensembles, we use a square root transformation, which has been shown to work well under various conditions (Yang et al 2017; Sharma et al 2018). Consequently, the predictor variables $x$ and $\phi$ are the mean and standard deviation of the square root transformed precipitation ensemble forecasts.

HXLR is implemented using the daily-average near-surface temperature forecasts, whereas HCLR is employed using daily precipitation accumulations. To verify the preprocessed CFSv2 forecasts, however, biweekly and monthly values are used since these are the timescales needed to generate the streamflow, nutrient load, and suspended sediments load forecasts. Both preprocessor models are applied to each CFSv2 grid cell in the JRB and each lead time ranging from 1 to 90 days. At each grid cell, the preprocessor models are implemented for the period 2003-2016 using a leave-one-out cross-validation approach. For this, we select a stationary period of 14-years for training and the remaining 1-year is used for verification purposes. This is repeated for each season and year, until all the 15 years of data are preprocessed and verified independently of the training period. Thus, to forecast a given season and specific lead time, we use $\sim 10,080$ forecasts (i.e., 8 members $\times$ 90 days per season $\times$ 14 years). Finally, the Schaake shuffle approach (Clark et al 2004) is applied to each individual forecast lead time to maintain the observed space-time variability in the preprocessed CFSv2 near-surface temperature and precipitation forecasts.

**Supplementary Text 2: Spatially distributed hydrological model**

NOAA’s National Weather Service Hydrology laboratory-Research Distributed Hydrologic Model (HL-RDHM) is used as the spatially distributed hydrological model (Koren et al 2004). Within HL-RDHM, the Sacramento Soil Moisture Accounting model with Heat Transfer (SAC-HT) is used to model hillslope runoff generation, and the SNOW-17 submodel is used to model snow accumulation and melting. The hillslope runoff generated at each grid cell by SAC-HT and SNOW-17 is routed to the stream network using a nonlinear kinematic wave approach (Koren et al 2004). Similarly, flows in the stream network are routed downstream using a nonlinear kinematic wave algorithm that accounts for parameterized stream cross-section shapes (Koren et al 2004, Smith et al 2012).

We run HL-RDHM in a fully distributed manner at 4-km spatial resolution and daily time step. HL-RDHM is used to both simulate daily flows and generate ensemble streamflow forecasts. To run HL-RDHM in forecasting mode, the preprocessed precipitation and near-surface temperature forecasts from the CFSv2 are used as the climate forcing. With HL-RDHM we generate raw ensemble streamflow forecasts for lead times of up to 90 days, initialized every 10 days. The simulated streamflows are generated by forcing HL-RDHM with gridded observed precipitation and near-surface temperature data.

To calibrate HL-RDHM, we select 10 out of the 17 SAC-HT parameters associated with baseflow, percolation, evaporation, storm runoff and the channel routing process. The most sensitive parameters are found to be the upper and lower soil zones transport and storage parameters, as well as the stream routing parameters. Note that when calibrating HL-RDHM we adjust the parameter fields rather than the actual parameter values at each grid cell using a multiplier approach (Kuzmin et al 2008). To calibrate HL-RDHM, a-priori parameter estimates based on previous studies (Anderson et al 2006, Reed et al 2004) are first manually adjusted. Once the manual changes do not yield
noticeable improvements in model performance, the parameter values are tuned up using an automatic technique, Stepwise Line Search (SLS) (Kuzmin et al. 2008; Kuzmin 2009). With SLS, the following objective function is optimized:

$$OF = \sqrt{\sum_{i=1}^{m} [o_{i} - s_{i}(\Gamma)]},$$

(7)

where $o_{i}$ and $s_{i}$ denote the daily observed and simulated flows at time $i$, respectively; $\Gamma$ is the parameter vector being estimated; and $m$ is the total number of days used for calibration. Four years (2002-2005) of streamflow data are used to calibrate the HL-RDHM for the selected basins, with the first year (year 2002) being used to warm-up the model.

**Supplementary Text 3: Statistical postprocessor**

Quantile Regression (QR) (Koenker 2005) is employed to both i) statistically postprocess the ensemble streamflow forecasts, and ii) generate the ensemble nutrients and suspended sediments load forecasts. The QR model is implemented separately for each lead time and basin. The QR model estimates lead time $t$-specific quantiles of a conditional distribution of streamflow,

$$\Phi_{t} = \{V_{t,\tau_{1}}, V_{t,\tau_{2}}, \ldots, V_{t,\tau_{T}}\},$$

(8)

where $T$ is the number of quantiles $\tau$ considered. In the case that $T$ is sufficiently large and the quantiles $\tau$ cover the entire domain of $[0, 1]$, we consider $\Phi_{t}$ to be a continuous distribution. For every lead time $t$ considered, and for every quantile $\tau$, the QR model between the verification $V$ and predictor variable, i.e., forecast $F$, is developed as:

$$V'_{t} = c'_{t} + d'_{t}F'_{t}.$$  

(9)

In equation (9), $c'_{t}$ and $d'_{t}$ are the regression parameters for quantile interval $\tau$ at time $t$. These parameters are determined by minimizing the sum of the residuals from a training dataset as follows:

$$\arg \min_{c',d'} \sum_{j=1}^{T} \Omega'_{j}[V'_{t,j} - (c'_{t} + d'_{t}F'_{j})].$$

(10)

$v'_{t,j}$ and $f'_{j}$ are the $j^{th}$ paired forecast-verification samples from a total of $J$ samples, $\Omega'_{j}$ is the QR function for the $\tau^{th}$ quantile defined as:

$$\Omega'_{j}(\psi'_{j}) = \begin{cases} (t-1)\psi'_{j} & \text{if } \psi'_{j} \leq 0 \\ \psi'_{j} & \text{if } \psi'_{j} > 0 \end{cases},$$

(11)

and $\psi'_{j}$ is the residual term computed as the difference between the verification and the linear QR estimate $c'_{t} + d'_{t}f'_{j}$ for any quantile. The resulting minimization problem in equation (10) is solved using linear programming via the interior point method.

To postprocess the raw ensemble streamflow forecasts generated from HL-RDHM with the QR model, we use equation (9) with streamflow observations as the verification $V$ and the ensemble mean forecast as the predictor variable $F$. Optimal parameter values in equation (9) are then computed using equations (10) and (11) from the training dataset. Finally, during the verification period, these optimal regression parameters are used to fit equation (9) and generate calibrated ensemble streamflow forecasts for any quantile. The QR postprocessor is employed with biweekly accumulated streamflow forecasts. The postprocessor is implemented for the years 2003-2016, using the same leave-one-out approach used for the preprocessor.

We also use the QR postprocessor to obtain TN, TP, and TSS load simulations and forecasts. This is done by using the monthly accumulated streamflow, together with the associated nutrient and suspended sediment load observations, as the training dataset. To generate the simulated nutrient and suspended sediment loads, we use equation (9) in log-linear form with the load observations as the verification $V$ and the simulated streamflow as the predictor variable $F$. To generate the forecasted nutrient and suspended sediment loads, we use instead the raw ensemble streamflow forecast mean as the predictor variable $F$ in equation (9).
Supplementary Figure 1: Map of study the area showing the gaged locations and the land cover distribution. The call out boxes show the NSE values for the streamflow, total nitrogen, total phosphorus, and total suspended sediment load simulations. The inset map shows the location of the basin in the northeast US.
Supplementary Figure 2: Time series plots of monthly accumulated streamflow for the selected forecast locations.
Supplementary Figure 3: Time series plots of monthly total nitrogen load for the selected forecast locations.
Supplementary Figure 4: Time series plots of monthly total phosphorous load for the selected forecast locations.
Supplementary Figure 5: Time series plots of monthly total suspended sediment load for the selected forecast locations.
References Supplementary Materials

Anderson R M, Koren V I and Reed S M 2006 Using SSURGO data to improve Sacramento Model a priori parameter estimates Journal of Hydrology 320 103–116

Clark M, Gangopadhyay S, Hay L, Rajagopalan B and Wilby R 2004 The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields Journal of Hydrometeorology 5 243-262

Hamill T M, Whitaker J S and Wei X 2004 Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts Monthly Weather Review 132 1434-1447

Koenker R 2005 Quantile regression, Cambridge University Press, Cambridge, 38, https://doi.org/10.1017/CBO9780511754098

Koren V, Reed S, Smith M, Zhang Z and Seo D J 2004 Hydrology laboratory research modeling system (HL-RMS) of the US national weather service Journal of Hydrology 291 297-318

Kuzmin V 2009 Algorithms of automatic calibration of multi-parameter models used in operational systems of flash flood forecasting Russian Meteorology and Hydrology 34 473–481

Kuzmin V, Seo D J and Koren V 2008 Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search Journal of Hydrology 353 109-128

Messner J W, Mayr G J, Zeileis A and Wilks D S 2014a Heteroscedastic extended logistic regression for postprocessing of ensemble guidance Monthly Weather Review 142 448-456

Messner J W, Mayr G J, Wilks D S and Zeileis A 2014b Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored Monthly Weather Review 142 3003-3014

Reed S et al 2004 Overall distributed model intercomparison project results Journal of Hydrology 298 27-60

Sharma S, Siddique R, Reed S, Ahnert P, Mendoza P and Mejia A 2018 Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system Hydrology and Earth System Sciences 22 1831-1849

Smith M B et al 2012 The distributed model intercomparison project – Phase 2: Motivation and design of the Oklahoma experiments J. Hydrol. 418 3–16

Wilks D S 2009 Extending logistic regression to provide full-probability-distribution MOS forecasts Meteorological Applications 16 361-368

Yang X, Sharma S, Siddique R, Greybush S J and Mejia A 2017 Postprocessing of GEFS Precipitation Ensemble Reforecasts over the US Mid-Atlantic Region Monthly Weather Review 145 1641-1658