The New England Journal of Statistics in Data Science Volume 1, 62–68 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS3

Dynamic Continuous Flows on Networks

Justina ZOU, Yi GUO, and David BANKS

Abstract

There are many cases in which one has continuous flows over networks, and there is interest in predicting and monitoring such flows. This paper provides Bayesian models for two types of networks—those in which flow can be bidirectional, and those in which flow is unidirectional. The former is illustrated by an application to electrical transmission over the power grid, and the latter is examined with data on volumetric water flow in a river system. Both applications yield good predictive accuracy over short time horizons. Predictive accuracy is important in these applications—it improves the efficiency of the energy market and enables flood warnings and water management.

KEYWORDS AND PHRASES: Autoregressive models, Bayesian hierarchical models, Network flows.

1. INTRODUCTION

Problems concerning the statistical modeling and monitoring of dynamic network flows arise naturally in areas such as energy, hydrology, and transportation. Our goal is to extend previous methodology for Bayesian dynamic flow models of discrete data [3] to the modeling of continuous flows. Rather than modeling Poisson counts, we use Normal and Gamma models for real-valued and positive flows, respectively. Our example applications concern the electrical power grid and hydrology, but the methodology applies to many more situations.

The first focus is the United States energy network, where key problems arise in managing electricity distribution. An excess of supply leads to lost revenue while an excess of demand leads to outages. Balancing authorities (BAs) help coordinate the demand and supply through interchange and power generation.

Most BAs produce electricity within their balancing authority area and can directly serve consumers. In addition, BA systems manage the flow of electricity in or out of their system to other BAs with which they are networked. This flow is called interchange. Electricity flow is directed, and power transfer occurs between two BA systems when one agrees to sell electricity to the other. The exact commercial value is determined by bids and offers within the market, and payments must be made by the end of the next business day. BAs which can accurately forecast demand over short horizons will be financially advantaged in these transactions.

We seek to model the dynamic flows between BAs as well as the self-flow within a BA. Each vertex in the network represents a balancing authority. We use data from the U.S. Energy Information Administration (EIA), which features hourly interchange, generation, and demand for 64 balancing authorities in the U.S. See Fig. 1 for a visualization of the BA network. Our data cover the time period from May 1 to May 31, 2020.

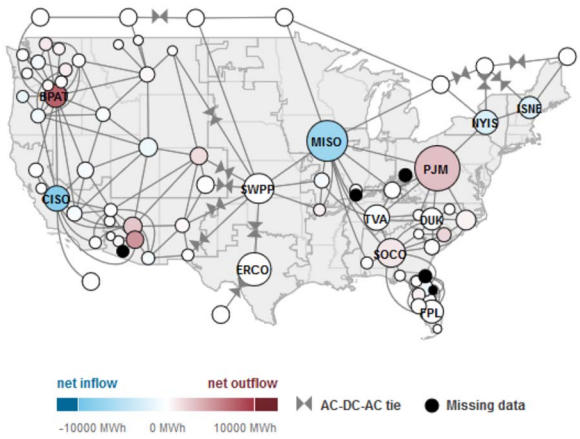

Figure 1: The BA network. Vertices are balancing authorities and edges are interchange values. The size of the node corresponds to BA size. Source: U.S. Energy Information Administration (May 2020).

Our second contribution is an application to hydrology. Hydrologic problems arise in environmental studies and flood modeling. Flooding occurs when the volume of water exceeds the capacity of a channel. To monitor river status, stream gauges at various points on a river measure the water flow and other properties such as precipitation and river height. Streamflow, or discharge, is the volume of water passing through rivers and other channels. Runoff from rainfall and evaporation are two of many mechanisms that can increase or decrease streamflow.

Hydrologic modeling has a long history [16]. Most previous statistical modeling of hydrologic activity focuses upon estimating the probability of extreme events, such as floods

Figure 2: Map of the Eel River in California. The seven black dots represent the approximate locations of the stream gauge stations. Adapted from Kmusser, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons.

[8], but there have also been efforts to model flows using stochastic processes [4]. We use Bayesian modeling to make near-term horizon forecasts of changing flows, which is important for reservoir management and flood control (e.g., decisions on dam retention and spillway operation).

In the hydrology model, vertices represent stream gauges and the edges represent directed, downstream flow from one part of the river to the next. We used the dataRetrieval package from R to obtain instantaneous discharge data in 15 minute intervals from the U.S. Geological Survey for the time period from January 1 to May 31, 2020 for the Eel River tributary system in northern California. See Fig. 2 for a map of the waterway with stream gauge locations. Discharge is calculated as the water velocity multiplied by the area in a cross section of the river. We obtain temperature and precipitation data from the National Oceanic and Atmospheric Administration.

The BA model easily extends to other applications with bidirectional continuous flows, such as banking networks or internet traffic. The hydrology model extends to networks with unidirectional continuous flows, such as gas pipelines or sewer systems. (Arguably, some of these flows might be discrete pennies or packets, but these are approximately continuous and neither satisfies the Poisson assumption used in Chen et al. [3].) Nonetheless, any new application will probably require some handfitted modification of the methods we describe.

In both case studies, we are interested in modeling seasonality and day-of-week effects. Electricity demand is generally regular; the demand on the previous Monday is likely similar to the demand of the current Monday. Similarly, precipitation and evaporation follow annual patterns that affect water levels. We take advantage of discounting and conjugacy to provide a general structure for sequential and seasonal analysis. Our work is able to characterize the node dynamics that are inherent in the networks. In addition, the model presents opportunities for Bayesian monitoring and intervention.

Sections 2 and 3 describe the statistical analyses for the electricity model and the hydrology model, respectively. Section 4 concludes.

2. THE ELECTRICAL GRID

Power demand forecasting has been done almost since the beginning of electrification. Previous research addresses different time horizons: annual usage drives the creation of new power plants [11], seasonal demand generally reflects weather trends [18], and hourly fluctuations are central to the energy market maintained by the BAs [17]. Our work focuses on the short horizon case.

Ghalehkhondabi et al. [5] reviews methodology for demand forecasting that was introduced between 2005 and 2015, especially for short term prediction. The methods listed include time series analysis, fuzzy logic, artificial neural networks, genetic algorithms, and hybrid methods. The review does not mention Bayesian methods, but these exist. Wang et al. [19] presents a Bayesian hierarchical regression model for predicting residential demand, and Bassamzadeh and Ghanem [2] presents a Bayesian network formulation. Our paper proposes a different Bayesian approach, one that ensures computational speed through decoupling and conjugacy, and which respects the connectivity of the electrical grid.

Suppose we have a connected network with v vertices and n edges. Let y^*_{ijt} denote the electrical flow in megawatt hours (MWh) between vertices i and j at time t; a positive flow runs from i to j; a negative flow runs from j to i. A flow y^*_{iit} denotes electricity created or used within the ith balancing authority (BA). Physics requires that $\sum_{j=1}^{n} y^*_{ijt} = 0$. This constraint will be enforced during the recoupling.

Exploratory data analysis confirmed the need to transform the data. We compared the standard logarithmic transformation to Fisher's arctanh transformation. The latter showed markedly better performance at stabilizing the variance. This transformation requires us to rescale the observations to lie in $[-1, 1]$, and to avoid infinite values, we rescaled to $(-1, 1)$. Specifically, let m_{ij1} be the minimum flow between BAs i and j, m_{ij2} be the second smallest value, M_{ij1} be the maximum flow, and M_{ij2} be the second largest value. We then applied the transformation

$$y_{ijt} = \operatorname{arctanh}\left(\frac{y^*_{ijt} - 2m_1 + m_2}{2M_1 - M_2 - 2m_1 + m_2}\right).$$

This avoids infinities by replacing the sample maximum and minimum with pseudovalues: $2M_1 - M_2$ is slightly larger than the maximum, $2m_1 - m_2$ is slightly smaller than the minimum, and these pseudovalues preserve the observed tail behavior.

We shall compare three models for these data. One is autoregressive with lag 168 (one week earlier) and covariate x_{jt}, which is the centered temperature at the jth BA at time t. Experts believe that there are weekly patterns in electricity usage and temperature is also a major driver of power demand [10]. The second model is autoregressive with lags 1 and 168. Temperature is implicitly present in the electrical demand during the previous hour. The third model is autoregressive with lag 1. As in Wilke [20], we shall compare these models in terms of predictive mean squared error for the next time step. Also, we use the first week of data to initialize the model for the autoregression using lag $T = 168$ and one hour of data to initialize for the simple lag 1 model.

The centering is done so that x_{jt} has mean zero, to prevent nonidentifiability due to confounding with the mean of the level process ϕ_{ijt} described below. Specifically, if the temperature at the jth BA at time t is K_{jt}, then $x_{jt} = K_{jt} - \bar{K}_j$.

The y_{ijt} is a time series with $y_{ijt}|\phi_{ijt} \sim N(\phi_{ijt}, \sigma^2_{ijt})$ that is conditionally independent over $t = 1, 2, \ldots$, where ϕ_{ijt} is a latent level process and σ^2_{ijt} is the variance of y_{ijt} at time t. The ϕ_{ijt} process evolves via a Markov model. To account for the influence of temperature on the electricity demand at destination node j as well as the day of the week effect, we have

$$\phi_{ijt} = \phi_{i,j,t-T} + \beta_{ij} x_{jt} + \epsilon_{ijt}$$

where $T = 168$ and ϵ_{ijt} is an innovation term. The first innovation is $\epsilon_{ij0} \sim N(0, \sigma^2_{ij0})$. The posterior updates of ϵ_{ijt} are $N(0, \gamma_t)$ where

$$\gamma_t = \left(\frac{k_{t-1}\delta_{t-1}}{\mathrm{Var}(\epsilon_{i,j,t-1})} + \frac{1}{\sigma^2_{ijt}} \right)^{-1}$$

with k_{t-1} the prior weight at time t such that $k_t = k_{t-1} * \delta_{t-1} + 1$ and $\delta_t \in (0, 1)$ is the discount factor that controls dependence upon the past. The innovation term ϵ_t is independent of ϵ_s and ϕ_s for $s < t$. The discount factor is determined as $\delta_t = d_i + (1 - d_i)\exp(-l k_{t-1})$, where d_i is the constant baseline discount factor for node i and $l > 0$ is a specified constant that determines how close δ_t is to d_i when information is high.

The second autoregressive model is like the first, except

$$\phi_{ijt} = \alpha_{ij} + \beta_{ij1}\phi_{ijt-1} + \beta_{ijT}\phi_{ijt-T} + \epsilon_{ijt}$$

where ϵ_{ijt} is the innovation term defined previously. The β_{ij} are not dynamic; we can determine a value for the β_{ij} by maximizing the model marginal likelihood with a normal prior on β such that $\beta \sim N(\beta_0, \sigma^2_\beta)$,

$$p(\beta|y_{0:t}, x_{1:t}) \propto \prod_{s=T+1:t} p(y_s|y_{T+1:s}, x_{1:s}, \beta) \times \prod_{l=0:T} p(y_l|\beta, x_{1:l})p(\beta).$$

Using the specification below, we have

$$p(\beta|y_{0:t}, x_{1:t}) \propto \prod_{s=T+1:t} t_{2r_t}(y_s|m_s, \frac{c_s(k_s+1)}{k_s r_s}) \times \prod_{l=0:T} N(y_l|\phi_l + \beta x_l, \sigma^2_l) N(\beta|\beta_0, \sigma^2_\beta).$$

Let ϕ and σ^2 have the usual normal-inverse gamma priors, which provides conjugacy and enables rapid computation. The hyperparameters for the mean and variance incorporate time series information on historical weekly and seasonal trends in electrical demand among the BAs. Specifically, at time $t = 0$, specify the priors as $\phi_0 \mid \sigma^2_0 \sim \mathrm{Normal}(m_0, \sigma^2_0/k_0)$ and $\sigma^2_0 \sim \mathrm{InverseGamma}(r_0, c_0)$, where $m_0 \in \mathbb{R}$, $k_0 > 0$, $r_0 > 0$, $c_0 > 0$ are known.

The time $t \to t + 1$ update/evolve steps are:

1. Time t prior:

$$\phi_t \mid x_{0:t-1}, \sigma^2_t \sim \mathrm{Normal}(m_{t-1}, \sigma^2_t/(k_{t-1}\delta_{t-1}))$$
$$\sigma^2_t \mid x_{0:t-1} \sim \mathrm{InverseGamma}(r_{t-1}, c_{t-1}).$$

2. Updates to the posterior:

$$\phi_t \mid x_{0:t}, \sigma^2_t \sim \mathrm{Normal}(m_t, \sigma^2_t/k_t)$$
$$\sigma^2_t \mid x_{0:t} \sim \mathrm{InverseGamma}(r_t, c_t)$$

where

$$m_t = \frac{k_{t-1}\delta_{t-1}m_{t-1} + x_{t-1}}{k_{t-1}\delta_{t-1} + 1}$$
$$k_t = k_{t-1}\delta_{t-1} + 1$$
$$r_t = r_{t-1} + 1/2$$
$$c_t = c_{t-1} + \frac{1}{2}\left(\frac{k_{t-1}\delta_{t-1}}{k_{t-1}\delta_{t-1} + 1}(x - m_{t-1})^2 \right).$$

3. Evolution of the time $t + 1$ prior:

$$\phi_{t+1} \mid x_{0:t}, \sigma^2_{t+1} \sim \mathrm{Normal}\left(m_t, \sigma^2_{t+1}/(k_t\delta_t)\right)$$
$$\sigma^2_{t+1} \mid x_{0:t} \sim \mathrm{InverseGamma}(r_t, c_t).$$

The third autoregressive model simply depends on the flow in the preceding hour. Aside from dropping the lag at $T = 168$, the model specification is exactly the same as for the immediately previous model.

We fit these three models with prior parameters $m_{ij,0:T}$ initialized with the first week of data, $k_{ij,0:T}, c_{ij,0:T}, r_{ij,0:T} =$

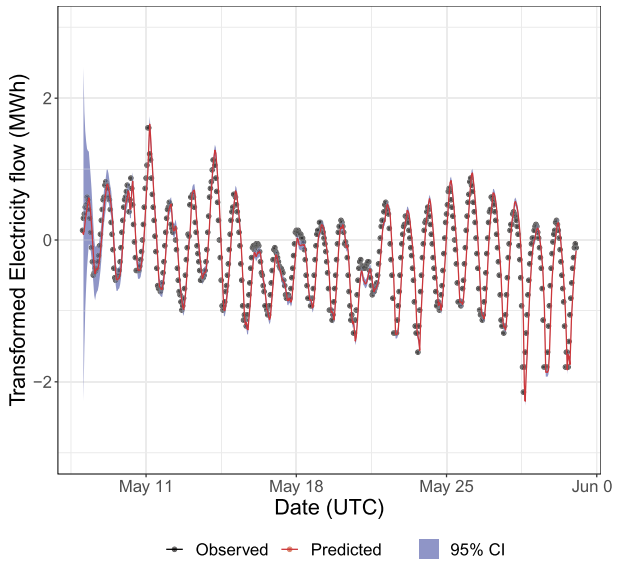

Figure 3: This figure shows the observed transformed power needs and the transformed power needs predicted by the model with two lags (the best model we considered) for the City of Homestead BA. It also shows a pointwise 95% confidence band (which becomes so tight that it is nearly invisible as time advances).

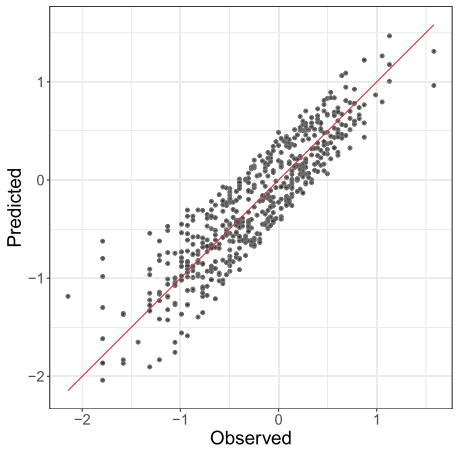

Figure 4: This figure is a scatterplot of the predicted transformed hourly demand against the actual transformed demand for the City of Homestead BA.

1, $\beta_0 = 0$, and $\sigma_\beta^2 = 1$. We compared the results in terms of one-step-ahead predictive squared error (OPSE). The autoregressive model that included temperature had an OPSE of 0.186, the autoregressive model with two lags had an OPSE of 0.100, and the model that depended only on the previous hour had OPSE equal to 0.281. The AIC value for the two-lag model is -1083.52 and the AIC for the one-lag model is 1125.84, so the former is preferred.

Figure 3 shows the predicted and observed values from the two-lag model, along with a 95% pointwise confidence band, for the City of Homestead BA. We selected that BA because it was easy to gather its temperature data and because it had relatively few edges in the network. Based on similar plots made for other BAs, we believe it is typical.

Similarly, Fig. 4 plots the model's predicted values against the actual values. It serves as a useful diagnostic. For example, note the tendency towards higher variation among the smaller values.

Using the two-lag model, we calculated the flows between all connected pairs of the 64 BAs. It took a total of 40 minutes to perform the computation using a laptop with a 2.5GHz CPU and 12 GB RAM. There are 296 edges in the network, and our decoupling, which considers only the flows between linked pairs of the BAs, enables massive parallelism. Further speed accrues from the conjugacy in our model.

This application seeks to advise BAs on how to manage their energy markets more efficiently, by accurately forecasting future needs and enabling detection of changepoints. So the sum-to-zero constraint that arises from the physics of energy transmission is not essential. (In fact, because of power loss due to resistance and other factors, the constraint only holds approximately). Nonetheless, it is interesting to address that case.

For the predicted values $\hat{y}_{ijt}$, let $\hat{y}^*_{ijt} = \tanh(\hat{y}_{ijt})$ so that the measurements are back on their original scale. Let $\mathcal{N}_i$ be the set of indices corresponding to the BAs to which the ith BA is connected. Let $\hat{y}^+_{ijt}$ denote the modified estimates that satisfy the sum-to-zero constraint, so that:

$$\forall i,\ y^+_{iit} + \sum_{j\in\mathcal{N}_i} y^+_{ijt} + \sum_{j\in\mathcal{N}_i} y^+_{jit} = 0. \tag{2.1}$$

This does not have a unique solution, so we add the condition that the y^+_{ijt} minimize

$$\sum_i \sum_j |\hat{y}^*_{ijt} - y^+_{ijt}|$$

at each value of t. This formulation is a linear equation with a convex constraint, so it is easily solved with standard software.

Ideally, this application would close with a comparison the predictive accuracy of the model described to the predictive accuracy of the forecasts made by BAs using commercial software (and, of course, show that our methodology is superior). Unfortunately, such software is proprietary and our efforts to work with relevant practitioners have so far been unsuccessful. Consequently, this publication represents what we believe to be the first statistically sophisticated model for predicting power markets among BAs in the literature.

3. THE HYDROLOGIC NETWORK

Let y_{ijt} denote the river flow in cubic feet (ft^3) from node i to node j at time t. In this study, we focus on the flows

between nodes rather than self-loops. Hence, we are not applying sum-to-zero constraints on the overall hydrologic network. The y_{iit} denotes the change in volume at node i at time t.

Rivers do not reverse their flow directions unless affected by tides or by rare geological activities. Thus, $y_{ijt} \geq 0$ if i and j are connected. To account for the non-negativity of the flows, we follow [13, section 5.5] and use a gamma distribution. Specifically, we have $y_{ijt}|\phi_{ijt} \sim \text{Gamma}(\alpha_{ij}, \alpha_{ij}/\phi_{ijt})$, where α_{ij} is a fixed shape parameter specified for the river between nodes i and j and $E(y_{ijt}|\phi_{ijt}) = \phi_{ijt}$. Here ϕ_{ijt} is a latent level process, which evolves via the following Markov model:

$$\begin{aligned}
&\frac{\alpha_{ij}}{\phi_{ijt}} = \frac{\alpha_{ij}}{\phi_{i,j,t-1}} \frac{\eta_{ijt}}{\delta_{ijt}}, \\
&\eta_{ijt} \sim \text{Beta}(\delta_{ijt} r_{ijt}, (1-\delta_{ijt}) r_{ijt}), \\
&\eta_{ijt} \perp \eta_{ijs}, \phi_{ijs} \text{ for } s < t.
\end{aligned}$$

The discount factor is determined as $\delta_t = d_i + (1 - d_i)\exp(-l r_{t-1})$, where d_i is the constant baseline discount factor for node i and $l > 0$ is a specified constant that determines how close δ_t is to d_i when information is high.

A generalized beta prime distribution is formed by compounding two gamma distributions:

$$\beta'(x; \alpha, \beta, 1, q) = \int_0^\infty G(x; \alpha, r) G(r; \beta, q) dr.$$

If a random variable $X \sim \beta'(\alpha, \beta, p, q)$ then $kX \sim \beta'(\alpha, \beta, p, kq)$, which is useful in accelerating computation.

Hence, we have the following equations by setting α_{ij} as the value that maximizes the model marginal likelihood,

$$\begin{aligned}
p(y_{ij,1:t} \mid y_{ij0}, \phi_{ij,1:t}) &\propto \prod_{s=1:T} p(y_{ijs} \mid y_{ij,0:s-1}, \phi_{ijs}) \\
&\propto \prod_{s=1:T} \beta'(\alpha_{ij}, r_{ij,s-1}, 1, c_{ij,s-1}/\alpha_{ij})
\end{aligned}$$

where the generalized beta prime distribution (or inverted beta distribution) has density

$$f(x|\alpha, \beta, p, q) = \frac{p \left(\frac{x}{q}\right)^{\alpha p - 1} \left(1 + \left(\frac{x}{q}\right)^p\right)^{-\alpha - \beta}}{q B(\alpha, \beta)}$$

for $0 \leq x < \infty$, $\alpha, \beta, p, q > 0$, and $B(\cdot, \cdot)$ is the usual beta function.

It is possible to include covariates x_{jt}, such as precipitation and temperature, at node j and time t to the model. As before, we center the covariates. The resulting model is

$$y_{ijt} - \widehat{\phi_{ijt}} = \boldsymbol{\beta_{ij}^T x_{jt}} + \epsilon_{ijt},$$

where $\epsilon_{ijt} \sim \text{Normal}(0, \sigma_{ij}^2)$.

At time $t = 0$, specify the prior as $\frac{1}{\phi_{ij0}} \mid y_{ij0} \sim \text{Gamma}(r_0, c_0)$, where $r_0 > 0$ and $c_0 > 0$ are known. Then the time $t \to t+1$ update/evolution steps are:

1. Time t prior:

$$\frac{1}{\phi_{ijt}} \mid y_{ij,t-1} \sim \text{Gamma}(\delta_{ijt} r_{ij,t-1}, \delta_{ijt} c_{ij,t-1})$$

2. Updates to the posterior:

$$\frac{1}{\phi_{ijt}} \mid y_{ijt} \sim \text{Gamma}(r_{ijt}, c_{ijt}),$$

   where $r_{ijt} = \alpha_{ij} + r_{ij,t-1}$ and $c_{ijt} = \alpha_{ij} y_{ijt} + c_{ij,t-1}$.
3. Evolves to the $t+1$ prior:

$$\frac{1}{\phi_{ij,t+1}} \mid y_{ijt} \sim \text{Gamma}(\delta_{ij,t+1} r_{ijt}, \delta_{ij,t+1} c_{ijt}).$$

For the Eel River hydrology data, we fit three models on the log of river flow: one without covariates, one with the log of precipitation, and one with the natural logarithm of precipitation and temperature. The logarithmic transformations were chosen based upon a preliminary exploratory data analysis. We set priors $\frac{1}{\phi_{ij0}} \sim \text{Gamma}(r_{ij0}, c_{ij0})$ with $r_{ij0} = c_{ij0} = 1$. The model without covariates had an OPSE of 0.00187, the model with one covariate had an OPSE of 0.00185, and the two-covariate model had an OPSE of 0.00185. The AIC for the zero covariate model is -8133.64, for the precipitation-only model it is -10571.28, and for the model with precipitation and temperature it is -10573.36. Since the two-covariate model has a lower AIC, it is preferred.

Figure 5 shows the predicted and observed river flow values for the two-covariate model for the flow at the gauge near Fortuna. We see that the predicted flow (red) closely follows the observed flow, as expected. Plots of the other stream gauges also show similar results.

Figure 6 shows the predicted flow against the observed flow. From the plot, it appears that the model tends to underestimate more than overestimate river discharge. This is reasonable since our analysis did not use relative humidity as a covariate, which, in that area during this time period, should reduce the evaporation rate.

Hydrology researchers measure the performance of a model in several ways, including root mean squared error and the percent bias in runoff ratio [6], and measures of percent bias in model predictions [12]. But the most traditional metric is the Nash-Sutcliffe model efficiency coefficient (NSE) [14]. This measure is equivalent to the coefficient of determination or r^2 value between predicted values and observations.

A second criterion important to the hydrological community is model agility [7]. Model agility refers to the ability of the same model to apply with high accuracy to different river segments and across different river systems. For many years, hydrology tried to micromodel the physical processes that drove water flow, including such quantities as leaf reflectance, maximum rate of carboxylation at 25°, a monthly leaf area index, and a plethora of other quantities

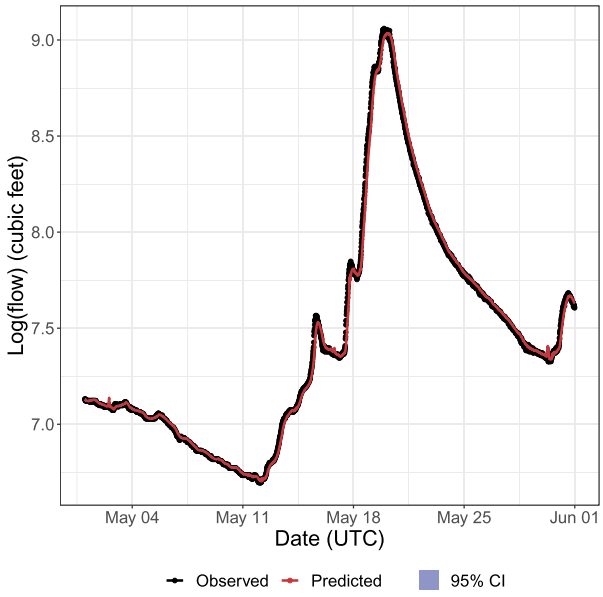

Figure 5: This figure shows the observed log river flow and the log river flow predicted by the model with log precipitation and temperature as covariates (the best model we considered) for the stream gauge closest to Fortuna. It also shows a pointwise 95% confidence band (which is so tight that it is nearly invisible).

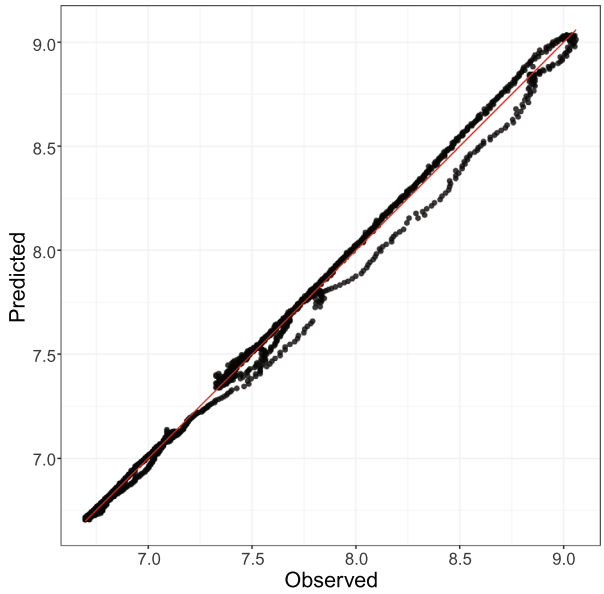

Figure 6: This figure shows the observed log river flow against the log river flow predicted by the model with log precipitation and temperature as covariates (the best model we considered) for the stream gauge closest to Fortuna.

that encoded beliefs about the physical mechanisms that drive evaporation, runoff, and rainfall. Our review of the literature did not discover any discussion of the Curse of

Table 1. This table shows the NSE for each gauge in the Eel River system. Gauge a_{123} is the closest to Fortuna, and it flows into the ocean. The notation is such that if d_3 flows to to c_3, then $d_3 \to c_3$. We have $d_3 \to c_3 \to b_3 \to a_{123}$, $c_2 \to b_2 \to a_{123}$, and $b_1 \to a_{123}$.

Gauge	NSE
a_{123}	0.9971
b_1	0.9909
b_2	0.9970
b_3	0.9966
c_2	0.9952
c_3	0.9967
d_3	0.9937

Dimensionality [cf. 9, section 2.5], but surely that would an issue in these models too.

Researchers found that these physics-based models lacked agility [12, 15]. They pushed for simpler models that worked well in general. To study the performance of our model with respect to agility and NSE, we computed the NSE for every gauge in the Eel River system. We see that the values are all very close to 1, demonstrating that our models have agility. Table 1 shows the results.

Values of NSE close to 1 correspond to good performance, and values near 0 indicate poor performance. Table 1 shows good performance across all the river segments, indicating that, at least within the Eel River system, our model demonstrates agility.

We initially considered a model that was lagged to reflect the time required for a bolus of water to move between adjacent gauges. That model did not (with one gauge pair exception) outperform the model without distance-based lagging. One possible reason for this lack of dependence is that the Eel River system is only 111 miles long, so rainfall, evaporation, and runoff are similar across the entire geography. Another issue is that flow rates are not constant. Typically they are slower at the headwaters and faster at the coast, but they also depend upon volume. After a heavy rainfall, rivers flow faster. For these reasons, we chose to use the simpler model which made no assumptions about travel time between gauges.

4. CONCLUSION

This paper addresses the problem of modeling continuous flows in networks, which arise in many physical and engineering situations. The models are widely applicable, although they will surely need to be tailored to specific applications. The examples in this paper provide some guidance on how to do such tailoring, through transformations, the use of covariates, and model selection based on such criteria as the AIC. The paper treats both directed and bidirectional flows.

The models use decoupling to enable massively parallel solution, which is essential because the number of edges in a

network often scale combinatorially. For the relatively sparse networks in our examples, it was possible to do the computation upon a single laptop, but for many applications it would be infeasible.

Similarly, we leverage conjugacy in the modeling to achieve rapid computation. Such conjugacy is not always faithful to the physical reality of the problem, but it is a convenient starting point for the analysis. In our examples, the conjugacy approximation performs well.

The electrical grid example has an interesting sum-to-zero constraint that should be respected when the decoupled analyses are recoupled. In contrast, the hydrology example does not, since precipitation runoff and evaporation are poorly measured and highly variable, and can differ over even relatively short geographic distances.

Accepted 7 March 2022

REFERENCES

[1] Andreadis, K. M. and Lettenmaier, D. P. (2006). Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources* **29**(6) 872–886. ISSN 0309-1708. URL. https://www.sciencedirect.com/science/article/pii/S0309170805002058.

[2] Bassamzadeh, N. and Ghanem, R. (2017). Multiscale stochastic prediction of electricity demand in smart grids using bayesian networks. *Applied energy* **193**. 369–380.

[3] Chen, X., Irie, K., Banks, D., Haslinger, R., Thomas, J. and West, M. (2018). Scalable bayesian modeling, monitoring, and analysis of dynamic network flow data. *Journal of the American Statistical Association* **113**(522) 519–533. https://doi.org/10.1080/01621459.2017.1345742.

[4] Clarke, R. T. (1988). *Stochastic processes for water scientists: development and applications*. John Wiley & Sons Ltd.

[5] Iman Ghalehkhondabi, Ardjmand, E., Weckman, G. R. and Young, W. A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems* **8**(2) 411–447.

[6] Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F. (2009). Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of hydrology* **377**(1–2) 80–91.

[7] Vijai Gupta, H., Sorooshian, S. and Yapo, P. O. (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of hydrologic engineering* **4**(2) 135–143.

[8] Haberlandt, U. and Radtke, I. (2014). Hydrological model calibration for derived flood frequency analysis using stochastic rainfall and probability distributions of peak flows. *Hydrology and Earth System Sciences* **18**(1) 353–365.

[9] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

[10] Hor, C. -L., Watson, S. J. and Majithia, S. (2005). Analyzing the impact of weather variables on monthly electricity demand. *IEEE transactions on power systems* **20**(4) 2078–2085.

[11] Hyndman, R. J. and Fan, S. (2009). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems* **25**(2) 1142–1153.

[12] Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G. and Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research* **51**(1) 716–728.

[13] Mauro, N. (2017). *Fundamentals of statistical hydrology*. Springer.

[14] Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology* **10**(3) 282–290.

[15] Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B. and Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology* **18**(8) 2215–2225.

[16] Singh, V. P. (2018). Hydrologic modeling: progress and future directions. *Geoscience letters* **5**(1) 1–18.

[17] Son, H. and Kim, C. (2017). Short-term forecasting of electricity demand for the residential sector using weather and social variables. *Resources, conservation and recycling* **123**. 200–207.

[18] Taylor, J. W. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting* **19**(1) 57–70.

[19] Wang, S., Sun, X. and Lall, U. (2017). A hierarchical bayesian regression model for predicting summer residential electricity demand across the usa. *Energy* **140**. 601–611.

[20] Wilke, U. (2013). Probabilistic bottom-up modelling of occupancy and activities to predict electricity demand in residential buildings. Technical report, EPFL.

Justina Zou. Department of Statistical Science, Duke University, USA.
E-mail address: justina.zou@duke.edu

Yi Guo. Department of Statistical Science, Duke University, USA.
E-mail address: yg109@duke.edu

David Banks. Department of Statistical Science, Duke University, USA.
E-mail address: dlbanks@duke.edu