Leveraging gauge networks and strategic discharge measurements to aid development of continuous streamflow records

Michael J. Vlah[1], Matthew R. V. Ross[2], Spencer Rhea[1], Emily S. Bernhardt[1]

[1]Duke University
[2]Colorado State University

Corresponding Author: MJV

MRVR, ESB, and MJV originated the project and identified its goals and methods. MJV carried out all analyses and drafted the manuscript. SR assisted in data collection. All authors took part in steering the project and editing the manuscript.

Grant sponsor: National Science Foundation, MacroSystems, 1926420

Abstract

Q~~Streamflow, or discharge, is an essential measure in the study of rivers and streams. However, q~~uantifying continuous discharge can be difficult, especially for nascent monitoring efforts, due to the challenges of establishing gauging locations, sensor protocols, and installations. Some continuous discharge series generated by the National Ecological Observatory Network (NEON) during its pre- and early-operational phases (2015-present) are marked by anomalies related to sensor drift, gauge movement, and incomplete rating curves. Here, we investigate the potential ~~for both simple and complex models to accurately estimate continuous discharge (at least daily estimates), using only discrete manual measurements of streamflow.~~ to estimate continuous discharge when discrete streamflow measurements are available at the site of interest. ~~We were inspired to do this work because some continuous discharge series generated by the National Ecological Observatory Network (NEON) during its pre- and early-operational phases (2015-present) are marked by anomalous data due to sensor drift, gauge movement, and incomplete rating curves.~~ Using field-measured discharge as truth, we reconstructed continuous discharge for all 27 NEON stream gauges ~~over this period~~ via linear regression on nearby donor gauges and/or prediction from neural networks trained on a large corpus of established gauge data. Reconstructions achieved median efficiencies of 0.83 (Nash-Sutcliffe, or NSE) and 0.81 (Kling-Gupta, or KGE) across all sites, and improved KGE at 11 sites versus published data,~~.~~ with linear regression generally outperforming deep learning approaches due to the use of target site data for model fitting, rather than evaluation only. Estimates from this analysis inform ~199 site-months of missing data in the official record, and can be used jointly with NEON data to enhance the descriptive and predictive value of NEON's stream data products. We provide 5-minute composite discharge series for each site that combine the best estimates across modeling approaches and NEON's published data. The success of this effort demonstrates the potential to establish "virtual gauges," or sites at which continuous streamflow can be accurately estimated from discrete measurements, by transferring information from nearby donor gauges and/or large collections of training data.

Introduction

Discharge, or streamflow, is a fundamental measure in hydrology, biogeochemistry, and river science more broadly. A measure of water volume over time, discharge is used to infer theoretical watershed runoff (depth of water “blanketing” the land surface, or depth over time), which in turn is integral to understanding watershed processes such as chemical weathering (White & Blum 1995). Accurate, and at least daily, discharge estimates are essential components of nearly any quantitative study of physical or chemical watershed or river processes at the ecosystem scale. Determination of solute fluxes (Bukaveckas et al. 1998), gas exchange rates (Hall, 2016), ecosystem metabolism (Odum 1956), and sediment transport (Graf 1984) all require well constrained estimates of discharge.

Despite its centrality to so many fields of study, discharge is a notoriously difficult metric to capture on a regular basis, especially in free-flowing systems, as it may vary greatly with annual cycles and weather events (Turnipseed & Sauer 2010). Established institutions like the USGS (USA), ECCC (Canada), and ANA (Brazil) have honed their instrumentation, methods, and monitoring locations over decades to generate reasonable discharge estimates even under extreme conditions (Benson & Dalrymple 1967; Costa 2004); however, nascent and/or small-budget monitoring efforts face several challenges. Critically, hundreds of these efforts are constantly occurring within academic research groups, municipalities, counties, and other entities building smaller gauge networks, with much less expertise, support, and budget than gauging programs supported by dedicated national programs.

Not including purely model-based methods for discharge prediction (Manning 1891; Hsu et al. 1995, Durand et al. 2022), automated discharge estimation requires the careful construction of an empirical “rating curve,” by which discharge can be continuously inferred from water level, or “stage” (but see Shen 1981). To build such a relationship, technicians must sample discharge and stage at points covering the range of observable flow, ideally including flood stage. In dynamic systems, this rating curve must be regularly updated. Point estimates of discharge can be collected using Acoustic Doppler current profiling (Moore et al. 2017), manual flow meter profiling, or light-based methods (Wang 1988) to determine average cross-sectional velocity, or via conservative tracer injections (Tazioli 2011). In many streams, two or more of these methods must be employed, depending on conditions (Turnipseed & Sauer 2010). During 10-year or 100-year floods, no method may be viable or safe. Even under regular storm conditions, a technician may be unable to mount a sampling effort quickly enough to capture peak flow, or may produce an inaccurate measurement. As a result, rating curves may remain in a state of insufficiency for years, during which time high discharge estimates are unreliable, especially where they are made by extrapolating beyond observed maximum flow.

Gauge placement presents another obstacle to the rapid deployment of discharge monitoring stations (Isaacson & Coonrod 2011). Stage measured via pressure transduction is susceptible to bias and nonlinearity under turbulent flow conditions (Horner et al. 2018). Sensors placed in a depositional area may be buried by sediment, and installations in forested watersheds or debris flow regions may be destroyed during floods. Often, equipment must be relocated at least once before a new gauge site can be properly established. Even an established stage-discharge rating curve must be regularly updated and maintained because the bed of the river can change as sediment is deposited or excavated, altering the relationship between stage and flow.

For some studies aiming to quantify stream or watershed processes that require continuous discharge time series, establishment of a high-quality monitoring station may be infeasible. Where co-location of the site of interest with an existing stream gauge~~Where co-location with an existing stream gauge~~ is also infeasible~~not possible~~, record extension (Hirsch 1982; Nalley et al. 2020) and gap-filling (Harvey et al. 2012; Arriagada et al. 2021) techniques cannot be employed, as these rely on prior knowledge of the statistical properties of the discharge time series being augmented. In such scenarios, streamflow reconstruction or prediction techniques are suitable, as these may proceed a priori or from minimal observation. Reconstruction typically involves methods that leverage the correlation between a partially measured target site and nearby "donor" (predictor) gauges. Discharge may also be quantified in the absence of direct measurements at the target location via statistical (Chokmani & Ouarda 2004), mechanistic (Regan et al. 2019), or machine learning (Kratzert et al. 2022) modeling techniques.

Here, we use both linear regression (OLS, L2/~~R~~ridge, segmented) and deep learning (LSTM-RNN) approaches to reconstruct discharge from the early operational phase (2015-2022) of the National Ecological Observatory Network (NEON), a time during which site ~~location~~ selection issues and rating curve development rendered potentially unreliable many site-months of discharge estimates (Rhea et al. 2023a). Our goal was to achieve Kling-Gupta Efficiency (KGE) scores greater than those of the official NEON continuous discharge product at as many sites as possible. A secondary goal was to improve temporal coverage of the official record where it contains gaps. For researchers intending to use NEON continuous discharge data between 2015 and 2022, the results of this effort, as well as efforts by Rhea et al. (2023a), can ensure that data gaps and questionable periods in the official record are replaced by high-quality estimates wherever possible. We provide composite discharge series~~,~~ for all 27 NEON stream gauge locations, built from the best NEON-published estimates and the best estimates generated by this study (https://doi.org/10.6084/m9.figshare.c.6488065). Composite series can be visualized at https://macrosheds.org/data/vlah_etal_2023_composites/.

The success of this effort demonstrates the viability of "virtual gauges" (*sensu* Philip & McLaughlin 2018; not to be confused with the "virtual staff gauges" of Seibert et al. 2019). In this study, we use the term to describe sites at which discrete discharge observations can be used to fit or evaluate models that generate continuous flow. For accurate results, field measurement campaigns should prioritize characterizing the distribution of possible flow conditions, rather than achieving any particular threshold number of observations. ~~In this study, we use the term to describe sites at which discrete discharge is measured at least 35 times across a wide range of flows, and can be used to fit or evaluate models that generate continuous flow.~~ Methods like those presented could be used to reduce the cost and simplify the process of establishing streamflow monitoring sites, especially in river networks that are already partially gauged.

Methods

Data selection, acquisition, and processing

We used the “neonUtilities” package (Lunch et al. 2022) in R to retrieve NEON discharge data. Officially released (NEON 2023b) and provisional (NEON 2023c) field measurements ~~(NEON 2023b; NEON 2023c; accessed 2023-01-23~~) were used to fit linear regression models and evaluate all models, as these data were collected directly by NEON technicians, using a combination of state-of-the-art methods including acoustic Doppler current profiling (ADCP; Moore et al. 2017), conservative salt tracer releases (Tazioli 2011), and flow meter measurements (Pantelakis et al. 2022). We used quality-controlled “finalQ” values where available, or “totalQ” values (taken directly from the flowmeter) in their absence. We refer to NEON’s discharge field measurements hereafter as e.g. “the response variable”, or “response discharge time series,” in the context of linear regression, or as the “target” variable in the context of machine learning. In either context, we refer to the 27 NEON sites for which discharge predictions were generated as “target sites” or “target gauges” (Table 1).

Continuous discharge data (NEON 2023a~~; 2023 release accessed 2023-05-01~~) were also retrieved via neonUtilities. These were used to finetune a subset of site-specific neural network models, and to construct composite discharge series. Provisional continuous discharge data were not used. Evaluation results used to distinguish likely reliable vs. potentially unreliable subsets of NEON’s RELEASE-2023 continuous discharge time series, per site-month, were provided by Rhea et al. (2023a) and accessed through HydroShare (Rhea 2023). Continuous elevation of surface water data are available, but approximately one third of all site-months are marked by disagreement between reported surface elevation and measured stage, or by likely sensor drift (Rhea et al. 2023a). We therefore chose not to use surface elevation to inform our models, though it no doubt contains predictive value.

Donor gauge data for linear regression analysis were acquired primarily from the US Geological Survey’s National Water Information System (NWIS), using the “dataRetrieval” package (DeCicco et al. 2022) in R. NWIS gauge ID numbers are provided in cfg/donor_gauges.yml at the GitHub and Zenodo links below. Additional donor gauge data from Niwot Ridge LTER and Andrews Forest LTER were retrieved from the MacroSheds dataset (Vlah et al. 2023) via package “macrosheds” (Rhea et al. 2023b), and from the EDI data portal (Johnson et al. 2020), respectively.

We used the original CAMELS dataset (Newman et al. 2014; Addor et al. 2017), the USGS National Hydrologic Model with Precipitation-Runoff Modeling System (NHM-PRMS; hereafter NHM; Regan et al. 2019), and the MacroSheds dataset as training data for neural network simulations of discharge data at each target site. CAMELS watershed attributes were generated for MacroSheds and NHM sites using the code provided at https://github.com/naddor/camels, except where otherwise indicated in Table 2, and daily Daymet meteorological forcings (Thornton et al. 2022; *sensu* Newman et al. 2015) were retrieved via Google Earth Engine (Gorelick et al. 2017). All code for this project can be found on GitHub, at https://github.com/vlahm/neon_q_sim, or in the Zenodo archive at https://doi.org/10.5281/zenodo.7976251. All data sources and links are provided in Table S2.

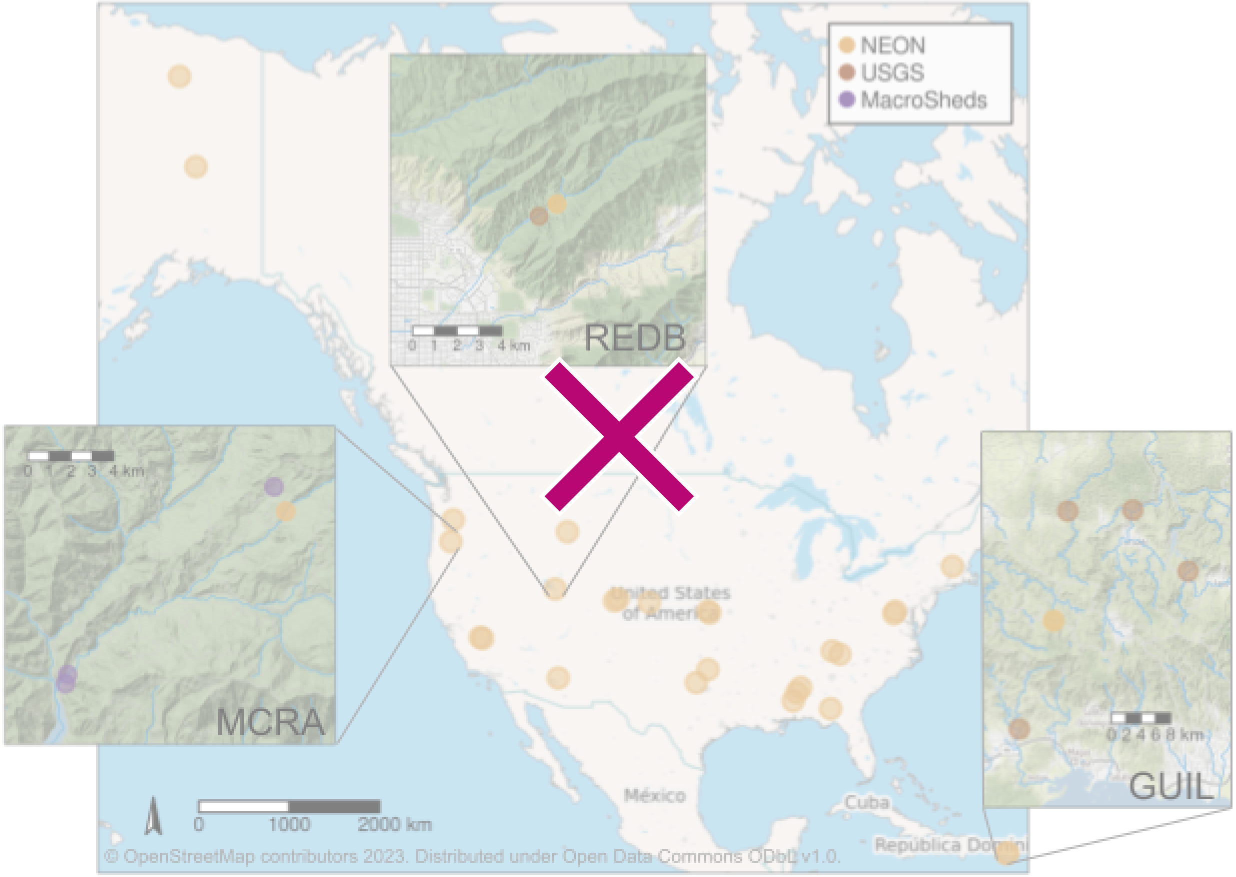
Donor Gauge Selection

Candidate donor gauges were identified by visually examining an interactive map of NEON gauges, USGS gauges, and MacroSheds gauges (https://macrosheds.org/ms_usgs_etc_reference_map/megamap.html), generated with package “mapview”

(Appelhans et al. 2022) in R. We also used the National Water Dashboard of the USGS (https://dashboard.waterdata.usgs.gov/app/nwd/en/?aoi=default) to identify active gauges in Alaska, USA. For each target site, up to four donor gauge candidates were selected on the basis of spatial proximity and geographic similarity to the target site (Figure 1). Generally, no greater than this number of gauges were even remotely reasonable candidates (i.e. within 50 km of the target site; not in an urban area; not downstream of a reservoir), but for one target site (MCRA) we had ten nearby candidate gauges to select from–all associated with the Andrews Experimental Forest in western Oregon State, USA. In this case we chose three candidate sites representing a catchment upstream of the target site (GSWS08), downstream of the target site on the MCRA mainstem (GSLOOK), and downstream on a tributary of MCRA (GSWS01).

Barring gauges on reaches that are subject to overt human influence, the exact methods used to choose donor gauges are of little consequence, so long as informative donor gauges are not overlooked. In practice, there will usually be just a few, if any, potential donor gauges available for a given location. If multiple donor gauges are included in a regression, L2 regularization (ridge regression) should be used to account for their covariance (see “Linear regression and model selection” subsection below.)

Target sites



© OpenStreetMap contributors 2023. Distributed under Open Data Commons ODbL v1.0.

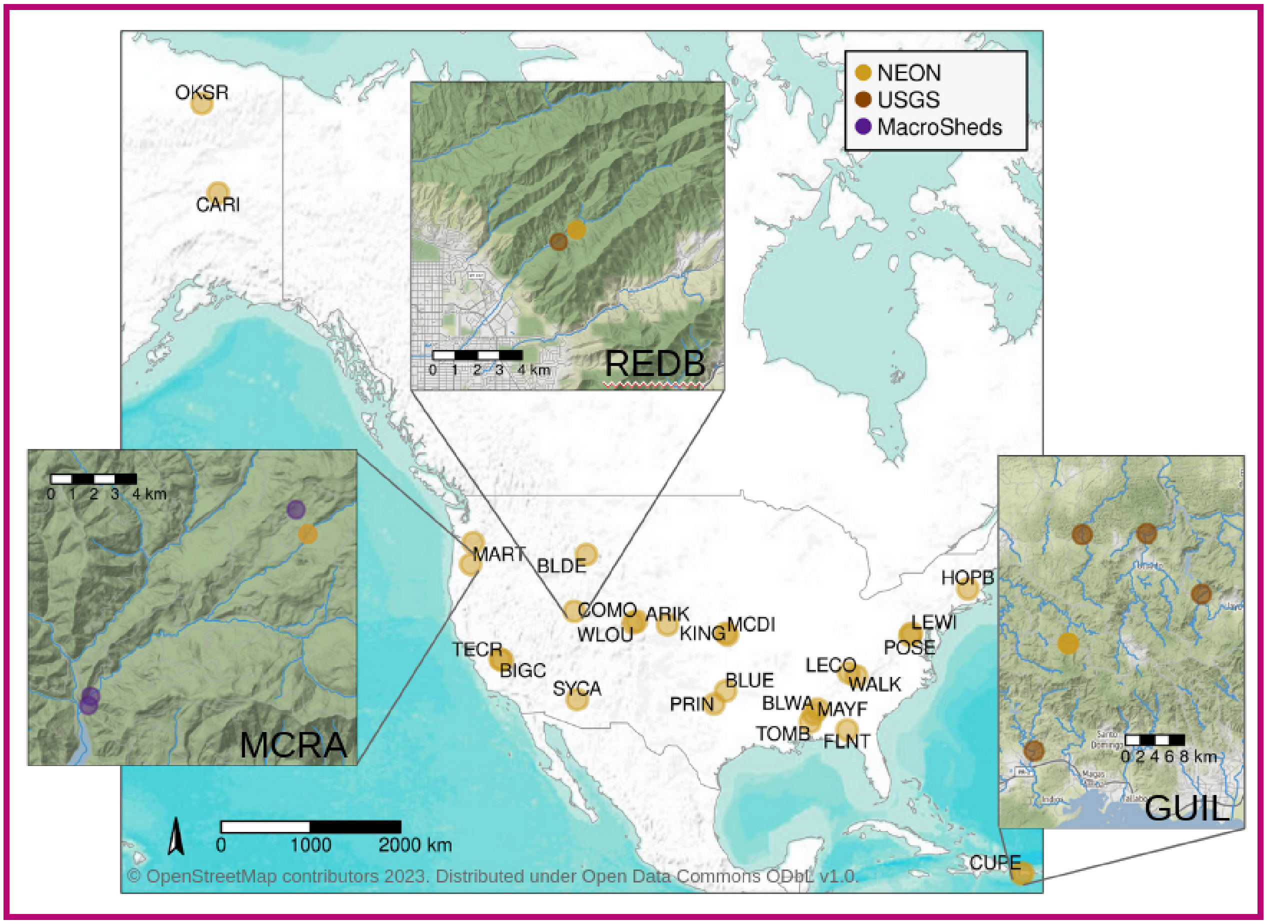


Figure 1. Map of target sites (NEON) and donor gauge candidates for three target sites: MCRA = McRae Creek, state of Oregon; REDB = Red Butte Creek, state of Utah; GUIL = Rio Guilarte, Puerto Rico.

All 27 lotic (flowing) aquatic sites associated with NEON were included as target sites for discharge prediction in this study (Figure 1). Sites TOMB, BLWA, and FLNT are installed on major rivers, downstream of hydropower dams. All other sites have been free of dam influence since 2012 at the latest, and are designated "wadeable streams" by NEON. In addition to the three sites above, hydrology at BLUE, GUIL, KING, MCDI, and ARIK may be influenced by agricultural activity, especially in the relatively arid Midwest (i.e. states KS, CO, OK). Continuous discharge data for TOMB are provided by a nearby gauge of the U.S. Geological Survey's National Water Information System, and are given at hourly intervals, rather than NEON's customary 1-minute intervals.

Table 1. Target sites for discharge prediction. See https://www.neonscience.org/field-sites for more information.

Site code	**Full name**	**State (USA)**	**Watershed area (km2)**	**Mean watershed elevation (m)**
TOMB	Lower Tombigbee River	AL	47085.3	20
BLWA	Black Warrior River	AL	16159.4	22

FLNT	Flint River	GA	14999.4	30
ARIK	Arikaree River	CO	2631.8	1179
BLUE	Blue River	OK	322.2	289
SYCA	Sycamore Creek	AZ	280.3	645
OKSR	Oksrukuyik Creek	AK	57.8	766
PRIN	Pringle Creek	TX	48.9	253
BLDE	Blacktail Deer Creek	WY	37.8	2053
CARI	Caribou Creek	AK	31.0	225
MCDI	McDiffett Creek	KS	22.6	396
REDB	Red Butte Creek	UT	16.7	1694
MAYF	Mayfield Creek	AL	14.4	77
KING	Kings Creek	KS	13.0	324
HOPB	Lower Hop Brook	MA	11.9	203
LEWI	Lewis Run	VA	11.9	152
BIGC	Upper Big Creek	CA	10.9	1197
GUIL	Rio Guilarte	PR	9.6	551
LECO	LeConte Creek	TN	9.1	579
MART	Martha Creek	WA	6.3	337
WLOU	West St Louis Creek	CO	4.9	2908
CUPE	Rio Cupeyes	PR	4.3	157
MCRA	McRae Creek	OR	3.9	876
COMO	Como Creek	CO	3.6	3021
TECR	Teakettle Creek - Watershed 2	CA	3.0	2011
POSE	Posey Creek	VA	2.0	276

WALK	Walker Branch	TN	1.1	264

Linear regression and model selection

~~Candidate donor gauges were identified by visually examining an interactive map of NEON gauges, USGS gauges, and MacroSheds gauges (https://macrosheds.org/ms_usgs_etc_reference_map/megamap.html), generated with package "mapview" (Appelhans et al. 2022) in R. We also used the National Water Dashboard of the USGS (https://dashboard.waterdata.usgs.gov/app/nwd/en/?aoi=default) to identify active gauges in Alaska, USA. For each target site, up to four donor gauge candidates were selected on the basis of spatial proximity and geographic similarity to the target site (Figure 1). Generally, no greater than this number of gauges were even remotely reasonable candidates (e.g. within 50 km of the target site; not in an urban area; not downstream of a reservoir), but for one target site (MCRA) we had ten nearby candidate gauges to select from–all associated with the Andrews Experimental Forest in western Oregon State, USA. In this case we chose three candidate sites representing a catchment upstream of the target site (GSWS08), downstream of the target site on the MCRA mainstem (GSLOOK), and downstream on a tributary of MCRA (GSWS01).~~ All donor and response discharge time series were neglog transformed (Equation 1; Whittaker et al. 2005) before fitting linear regression models.

Equation 1: $x_{neglog} = sign(x)log(|x| + 1)$

Series were scaled by 1000 before transformation, in order to reduce the disproportionate impact of adding one to every value. Response observations were synchronized to the interval of the predictor series by approximate datetime join, allowing forward or backward time-shifts of up to 12 hours if necessary.

One of three forms of linear regression was employed at each site, depending on the number and location of donor gauges, and the donor-target gauge relationships. For sites with a single donor gauge (REDB, HOPB, BLUE, SYCA, LECO), considered predictors were: discharge from the donor gauge, a 4-season categorical variable, and their interaction. Additionally, an intercept parameter could be estimated, or not, for each specification. Thus, up to six models were fit using Ordinary Least Squares (OLS) regression (Galton 1886), ensuring at least 15 observations per model parameter. At LECO, an additional dummy variable was included to address an intercept change due to a wildfire in November of 2016. The best model was selected via 10-fold cross-validation, minimizing mean squared error (MSE). MSE, being a squared-error term, disproportionately penalizes inaccurate prediction of high discharge values, and helps to balance against the relative rarity of high discharge measurements in the field data. At site SYCA, the log-log relationship between discharge at the target gauge and a single donor gauge exhibited a distinct breakpoint, and segmented least-squares regression was used (R package "segmented"; Muggeo 2008). At all other sites (19 in total), predictors included discharge series from 2-4 donor gauges, season, and all interactions. To control overfitting and shrink covarying coefficients toward zero, we used L2 regularization (~~R~~ridge regression; Gruber 2017) via R package "glmnet" (Friedman et al. 2010). As with the other regression approaches, 10-fold cross-validation and MSE loss were used for model parameter selection–in this case for the value of the penalty hyperparameter λ, which was set to the mean across folds of λ producing minimum cross-validated error. Unlike OLS and segmented regression, r~~R~~idge regression uses biased estimators that complicate calculation of prediction intervals. We generated 95% prediction intervals for ridge regression~~glmnet~~ discharge estimates using the 95th percentiles of 1000 bootstrap predictions at each prediction point, generated from 1000 resamples of the fitting data, stratified by season. We emphasize that these prediction intervals should be conservative estimates of the true uncertainty, as they do not fully account for uncertainty due to bias (Goeman et al. 2012).

For each site, we fit two sets of models as described above, one with discharge scaled by watershed area (i.e. “specific discharge” in the surface water hydrology sense) prior to transformation, and one without areal scaling. Only one model from each set was ultimately selected for each target site, on the basis of Kling-Gupta efficiency (KGE; Gupta et al. 2009), a composite model efficiency metric that incorporates measures of correlation, variance, and bias. We also report percent bias and Nash-Sutcliffe efficiency (NSE; Nash & Sutcliffe 1970), a measure of predictive accuracy that implicitly compares predictions to a mean-only reference model.

Predictions were generated for all time points during which data were available at the selected donor gauges. At target site COMO, a secondary model omitting one donor gauge was able to produce 36% more predictions than the selected model, so our predicted discharge at COMO is a composite of both models, preferring the better model’s predictions where available. We were unable to locate sub-daily donor gauge data near COMO, so regression predictions for this site are at a daily interval. Regression predictions for all other sites were generated at sub-daily intervals matching the coarsest interval across predictor gauges–generally 15 minutes, though note that in most cases these predictions were interpolated to five minutes for our composite discharge product.

Neural network setup and operation

Supplementing the linear regression methods described above, we simulated discharge data at all 27 target sites using long short-term memory recurrent neural networks (LSTM-RNN; hereafter “LSTM”; Hochreiter & Schmidhuber 1997). Four LSTM strategies were employed, all of which involved training on a large and diverse corpus of stream discharge data (Table 3). Two of these strategies included further finetuning to the time-series dynamics of each target site in turn. Due to the relative scarcity of field-measured discharge observations (between 39 and 213 per site; mean 122), none were used in LSTM training. Instead, these measurements were used only to evaluate predictions. LSTMs trained in this study are intended only for discharge prediction within the temporal and spatial bounds of NEON’s early operational phase, not for forecasting or application to other sites. Therefore, all available, daily training data were used as such; no validation set was kept for hyperparameter tuning, and no holdout set of daily estimates was kept for evaluation (note that split-sample designs may be undesirable more generally: Arsenault et al. 2018; Guo et al. 2018; Shen et al. 2022). See Kratzert et al. (2019b) and Read et al. (2019) for split-sample considerations in the context of a generalist and process-guided generalist LSTM, respectively.

After a hyperparameter search routine, described below, potentially skilled models were identified as those achieving at least 0.5 KGE and 0.4 NSE. The best performing, potentially skilled LSTM for each site (if applicable) was then re-trained 30 times, forming an ensemble. Ensembles were trained for 18 of

27 sites. LSTM predictions included in our composite discharge product are means taken across the distributions of ensemble point predictions. Uncertainty bounds were computed as the 2.5 and 97.5% quantiles of these distributions. LSTM skill was evaluated on the basis of mean ensemble efficiency (KGE) with respect to field-measured discharge (Table S1).

Daily discharge time series (training data) and field-measured discharge were scaled by watershed area. For each predicted day, LSTMs received 5 dynamic Daymet meteorological forcing variables and 11 static watershed attribute summary statistics (Table 2). Multitask learning (Caruana 1998; Sadler et al. 2022) was found to improve discharge prediction broadly in a preliminary analysis, so Daymet minimum air temperature was used as a secondary target variable. Kratzert et al. (2019a) found that a maximum of about 150 preceding days were able to influence LSTM output on a similar prediction problem, so we set the input sequence length to 200 days to ensure full utilization of available information. In other words, for each day of prediction, the model was able to leverage information from the preceding 200 days.

We employed four different training pipelines described in Table 3. Of the 671 CAMELS watersheds (i.e. basins), we used a subset of 531 with undisputed areas less than 2000 km^2 (Newman et al. 2017). For finetuning data, we used version 1 of the MacroSheds dataset (Vlah et al. 2023). We excluded MacroSheds sites outside North America, or with coastal or urban hydrological influence, for a total of 133 sites out of the 169 that are currently available. We chose MacroSheds sites for finetuning because the MacroSheds and NEON datasets focus primarily on small watersheds, often smaller than 10 km^2 in area, while only eight CAMELS watersheds are smaller than 10 km^2 and most are larger than 100 km^2 (Vlah et al. 2023). Daily mean discharge computed from NEON's continuous discharge product, only for those site-months deemed Tier 1 or Tier 2 by Rhea et al (2023a), was used alongside MacroSheds data for finetuning.

For the process-guided strategies, we used NHM estimates for all reaches coinciding with a CAMELS or MacroSheds gauge, for a total of 551 reaches. Only nine target sites on relatively high-order streams were amenable to the process-guided specialist approach, as these sites are on reaches large enough to be modeled by the NHM. The most recent version of the NHM at the time of this writing provides discharge estimates beginning in 1980, and ending in 2016, just before the installation of most NEON target sites.

Table 2. LSTM input data. * = Attribute tested as an afterthought, but not included in this study due to negligible improvement in trial parameter search.

Meteorological forcing data (watershed-average time series)	
Maximum air temp	2-meter daily maximum air temperature (°C)
Precipitation	Mean daily precipitation (mm/day)
Solar radiation	Daily surface-incident solar radiation (W/m2)
Vapor pressure	Near-surface daily average vapor pressure (Pa)
PET	Potential evapotranspiration (mm); estimated using Priestley-Taylor

	formulation with gridded alpha product (Aschonitis et al. 2017)
Watershed attributes (statistics computed over full record)	
Precipitation mean	Mean daily precipitation (mm/day)
PET mean	Mean daily potential evapotranspiration (mm/day); estimated using Priestley-Taylor formulation with gridded alpha product (Aschonitis et al. 2017)
Aridity index	Ratio of PET mean to Precipitation mean
Precip seasonality	Seasonality of precipitation; estimated by representing annual precipitation and temperature as sine waves. Positive values indicate summer peaks, while negative values indicate winter peaks. Values near 0 indicate uniform precipitation throughout the year.
Snow fraction	Fraction of precipitation falling on days with temp < 0 °C
High precipitation frequency	Frequency of high precipitation days (days with ≥ 5x mean daily precipitation)
High precip duration	Average duration of high precipitation events (number of consecutive days ≥ 5x mean daily precipitation)
Low precip frequency	Frequency of dry days (days with precipitation < 1 mm/day)
Low precip duration	Average duration of dry periods (number of consecutive days with precipitation < 1 mm/day)
Elevation	Catchment mean elevation (m)
Slope	Catchment mean slope (m/km)
Area	Catchment area (km^2)
Source*	Binary indicator for NHM estimates–process-guided LSTMs only.
Target data (time series)	
Discharge	Specific discharge, or discharge normalized by watershed area. The same quantity may be referred to as “runoff” in other studies (mm/day).
Minimum air temp	2-meter daily minimum air temperature (°C)

Table 3. LSTM model training pipelines used in the simulation of discharge at target sites. Here, “NEON” refers to NEON’s continuous discharge product, RELEASE-2023~~2023 release~~, with quality-flagged estimates and < Tier-2 site-months (according to Rhea et al. 2023a) removed.

Model type	**Phase 1**	**Phase 2**	**Phase 3**
Generalist	Pretrain on CAMELS	Finetune on MacroSheds + NEON	N/A
Specialist	Pretrain on CAMELS	Finetune on MacroSheds + NEON	Finetune on NEON target site
Process-guided generalist	Pretrain on CAMELS + CAMELS-NHM	Finetune on MacroSheds + MacroSheds-NHM + NEON + NEON-NHM	N/A
Process-guided specialist	Pretrain on CAMELS + CAMELS-NHM	Finetune on MacroSheds + MacroSheds-NHM + NEON + NEON-NHM	Finetune on NHM estimates for target site

LSTMs were configured in R, and trained~~, validated, and tested~~ using v1.3.0 of the NeuralHydrology library in Python (Kratzert et al. 2022; Van Rossum & Drake 2009) on the Duke Compute Cluster at Duke University, Durham NC, USA. All trained models used the Adam optimizer (Kingma & Ba 2014) and NeuralHydrology’s “NSE loss” function, after an initial evaluation in which we compared it to MSE and root mean squared error (Table 4). Learning was annealed using series of three fixed rates for pretraining and for round one of finetuning, according to:

Equation 2:
$$r = \begin{cases} a, & e \in \left\{0, \cdots, \left\lfloor \frac{E}{3} \right\rfloor\right\} \\ \frac{a}{10}, & e \in \left\{\left\lceil \frac{E}{3} \right\rceil, \cdots, \left\lfloor \frac{2E}{3} \right\rfloor\right\} \\ \frac{a}{100}, & e \in \left\{\left\lceil \frac{2E}{3} \right\rceil, \cdots, E\right\} \end{cases}$$

Where r is the learning rate, a is any power of 10 between 0.1 and 10^{-7}, and E is the number of training epochs. Learning rate was annealed using series of two fixed rates for round two of finetuning, according to:

Equation 3:
$$r = \begin{cases} \frac{a}{10}, & e \in \left\{0, \cdots, \left\lfloor \frac{E}{2} \right\rfloor\right\} \\ \frac{a}{100}, & e \in \left\{\left\lceil \frac{E}{2} \right\rceil, \cdots, E\right\} \end{cases}$$

Learning rate and other hyperparameters were selected via an inexhaustive (pseudo) grid search (Table 4), i.e. we specified a sequence of possible values for each hyperparameter and randomly selected from them to specify 30 models for each generalist. For each site, one specialist model was then configured to further finetune each of the 30 generalists, again using partial grid search to define any mutable hyperparameters. Otherwise, hyperparameters were inherited from the previous training period (Table 4). Due to our incomplete hyperparameter search procedure, better combinations probably exist. We elected not to exhaustively pursue optimal hyperparameter combinations due to the computational demand of a full grid search, and a lack of access via NeuralHydrology to callback methods necessary for implementation of true random search (Bergstra & Bengio 2012).

Table 4. LSTM hyperparameter search space for all model types, and selected values (bold, italic) used for pretraining. These were observed to allow for both malleability and high performance of subsequent finetuning iterations over nearly 2000 exploratory LSTM trials. The ditto mark "``" indicates that a finetuning parameter is inherited from the preceding training iteration. The relationship of *a* to the learning_rate is defined in Equations 2 and 3. See the NeuralHydrology documentation for parameter definitions: https://neuralhydrology.readthedocs.io/en/latest/usage/config.html.

LSTM parameter	**Pretrain**	**Finetune 1**	**Finetune 2 (specialists only)**
hidden_size	20, ***30***, 40, 50	``	``
output_dropout	0.1, 0.2, 0.3, 0.4, ***0.5***, 0.6	0.2, 0.3, 0.4, 0.5	``
learning_rate *a*	10^{-2}, $\mathbf{10^{-3}}$, 10^{-4}, 10^{-5}	10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}	10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}
batch_size	32, 64, 128, 256, ***512***, 1024	32, 64, 128, 256, 512	``
epochs	20, ***30***, 40, 50, 60	20, 30, 40	10, 20, 30
finetune_modules	N/A	head, lstm, head & lstm	head, lstm
target_variables	discharge, ***discharge & min air temp***	``	``
loss	***NSE***, MSE, RMSE	``	``

All LSTM models were outfitted with fully connected, single-layer embedding networks to efficiently encode inputs as fixed-length numerical vectors (Arsov & Mirceva 2019). Separate embedding networks were used for static and dynamic inputs, with 20 neurons for static inputs and 200 neurons for dynamic inputs. All embedding neurons used the hyperbolic tangent activation function. Another advantage of embedding networks in the context of the NeuralHydrology library is that they provide one of few opportunities to introduce dropout, which can improve training efficiency and reduce overfitting (Srivastava et al. 2014).

Composite discharge data product

This study generated time-series predictions of discharge for each lotic NEON site using up to three distinct processes: linear regression on absolute discharge, linear regression on specific discharge, and one of four LSTM strategies. We provide regression predictions wherever applicable (24 of 27 sites). LSTM predictions are provided only for sites that had promising model performance after a hyperparameter search, and for which ensemble models were therefore trained (18 of 27). All model outputs and results from this study are archived at https://dx.doi.org/10.6084/m9.figshare.22344589.

In addition to predictions from individual modeling strategies, we provide an analysis-ready discharge dataset for all 27 sites that splices the best available predictions across methods, including published NEON estimates (NEON 2023a~~; accessed 2023-05-01~~), into composite series (https://dx.doi.org/10.6084/m9.figshare.23206592), which can be visualized interactively at https://macrosheds.org/data/vlah_etal_2023_composites/. Composite series for each NEON site begin at the start of site operation and extend to at most September 30, 2021, the last date included in the 2023 release of NEON's continuous discharge product. We also provide individual model predictions extending through 2022. A complete list of products from this study, and their links, can be found in Table S3.

To construct composite series, we first distinguished as "good" site-months of NEON discharge estimates categorized as Tier 1 or Tier 2 by Rhea et al. (2023a). For a NEON site-month to meet the requirements for at least Tier 2, four requirements must be met. The linear relationship between stage, determined from pressure transducer readings, and field-measured gauge height must score at least 0.9 NSE. The transducer-derived stage series must also pass a drift test, relative to gauge height, but only if sufficient data exist to perform such a test. The rating curve used to relate stage to discharge must score at least 0.75 NSE, and fewer than 30% of predicted discharge values may exceed the range of measured discharge used to build the curve. See Rhea et al. (2023a) for further details.

Although only 50% of NEON's RELEASE-2023~~2023-release~~ estimates are classified as Tier 1 or Tier 2, the remainder may still be of high analytical value if NEON's quality control indicators and uncertainty bounds are observed. We also stress that NEON rating curves and protocols have improved over the course of its early operational phase, and continue to do so.

We then ranked the available predictions for each site, assigning rank 1 either to predictions from linear regression, or to NEON's continuous data product, depending on overall KGE and NSE against field measured discharge. KGE was considered first, and used to determine preference except in cases where the difference between NSE scores was greater than that between KGE scores, and opposite in sign. Rank 2 predictions were then used to fill gaps of 12 or more hours in the rank 1 series, but only "good" NEON site-months were included. Only after this first round of gap-filling were the remaining NEON data incorporated, with site-years achieving at least 0.5 KGE and 0.5~~4~~ NSE against field-measured discharge being used to fill still-remaining gaps. Finally, daily LSTM predictions (placed at 12:00:00 UTC on the day of prediction) were used to fill any recalcitrant gaps, but only if produced by an ensemble model achieving at least 0.5 KGE and 0.5 NSE across all field discharge observations. Note that while such benchmarks are in common use (Moriasi et al. 2015), the efficiency that any model can or should achieve varies substantially with the hydroclimate and watershed characteristics of a given site (Seibert et al. 2018). We provide all data and code for modifying the composite discharge product in accordance with alternative benchmarks as users see fit. After visual examination of composite series plots, we chose to prefer NEON predictions to linear regression predictions at site ARIK, "good" or not, due to frequent sharp disjoints between the two predicted series. See Table S1 for an account of linear regression and LSTM methods used in the construction of ensemble series.

The prevailing interval varies across data sources used to assemble our composite discharge product, from one minute (NEON) to one day (LSTM predictions; regression predictions at site COMO). Regression predictions were primarily generated at 15-minute intervals, and their timestamps are always divisible by 15 minutes. Around the prevailing NEON interval there is considerable variation due to data gaps and sensor reconfigurations, both across sites and across the temporal ranges of each site's record. To reduce the complexity associated with irregular time-series analysis, we synchronized the interval across data sources to five minutes. Regression estimates were linearly interpolated to five minutes, though gaps larger than 15 minutes were not interpolated. NEON estimates were first smoothed with a triangular moving average window of 15 minutes to remove unrealistic minute-to-minute noise associated with Bayesian error propagation. They were then interpolated the same way as the regression estimates, and finally downsampled to five minutes, with some timestamps being shifted by up to two minutes. For example, a duration of 30-minute sampling, with a sample taken at 00:03:00, would be shifted by two minutes, by rounding each timestamp up to the nearest minute divisible by five.

Results ~~and Discussion~~

A performance comparison of linear regression on discharge from donor gauges, and four LSTM strategies, is shown in Figure 2 and Figure S1, and detailed in Table S1. Via linear regression, we were able to produce 15-minute discharge estimates at 11 sites with overall KGE scores higher than those of published series (Figure 2). At four of the same sites, we achieved higher KGE via LSTM methods, which generated daily discharge series. Of the ten sites at which published discharge KGE was less than 0.8, we improved five to above that mark (mean 0.932, n = 5).

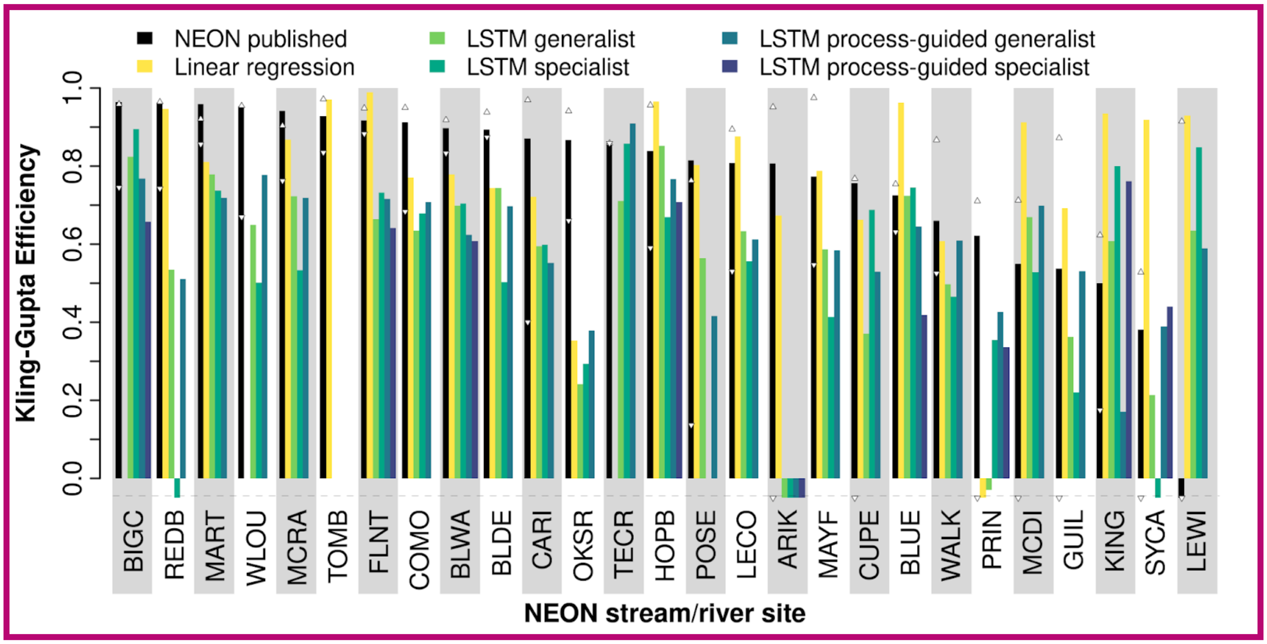


Figure 2. Efficiency of five stream discharge prediction methods and NEON's published continuous discharge product at 27 NEON gauge locations, versus field-measured discharge. Small, white triangles represent max/min KGE of published discharge by water year (Oct 1 through Sept 30) with at least 5 field measurements (or 2 for site OKSR). KGE was computed on all available observation-estimate pairs except those with quality flags (dischargeFinalQF or dischargeFinalQFSciRvw of 1). For the best performing LSTM method, at all sites except TECR, FLNT, REDB, WALK, POSE, and KING, displayed KGE is averaged over 30 ensemble runs with identical hyperparameters. For the sites just named,

performance of a chosen method, after ensembling, dropped below that of at least one other method's optimal KGE from parameter search. For all other LSTM site-method pairs, which were not ensembled, displayed performance is that of the best model trained during the parameter search phase. Sites are ordered by the KGE of NEON continuous discharge. See Table 3 for LSTM model definitions. KGE of 1 is a perfect prediction, while KGE of -0.41 is similar in skill to prediction from the mean. Negative values are truncated at -0.05 in this plot to improve visualization.

For 12 of 27 sites, linear regression on specific discharge (i.e. scaled by watershed area) provided the most accurate discharge predictions, while linear regression on absolute discharge performed better at the other 12 sites with donor gauges. LSTM models (as proper ensembles) outperformed linear regression at only 2 sites. In general, linear regression provided more accurate predictions than all LSTM methods. Linear regression on absolute discharge produced estimates with median NSE of 0.848 and median KGE of 0.806, across sites (*n* = 24; Table 5). Linear regression on specific discharge produced similar median scores (Table 5), but with deviations of up to 0.05 NSE and 0.08 KGE at individual sites.

Table 5. Performance of five stream discharge prediction methods, and official continuous discharge time-series data, across *n* of 27 NEON gauge locations (final column). For both the Nash-Sutcliffe and Kling-Gupta Efficiency coefficients, a value of 1 indicates perfect prediction. A value of 0 NSE indicates that predictive skill is equivalent to prediction from the mean, while negative NSE is worse than mean prediction. This threshold lies at approximately -0.41 for KGE (Knoben et al. 2019). "Linreg" = linear regression on donor gauge discharge series, and "scaled" means predictor and response discharge were scaled by their respective watershed areas.

	NSE				**KGE**				
Model/Data	**Median**	**Mean**	**Min**	**Max**	**Median**	**Mean**	**Min**	**Max**	***n***
Official record	0.880	0.417	-9.95	0.989	0.839	0.711	-1.50	0.964	27
Linreg	0.848	0.760	-0.038	0.993	0.806	0.746	-0.697	0.988	24
Linreg scaled	0.847	0.757	-0.037	0.993	0.807	0.743	-0.695	0.989	24
Generalist LSTM	0.473	-18.8	-498	0.904	0.634	-0.220	-20.2	0.852	26
Specialist LSTM	0.477	-12.6	-307	0.920	0.556	-0.256	-15.7	0.895	25
Process-guided generalist LSTM	0.434	-31.3	-824	0.848	0.618	-0.453	-26.4	0.869	26
Process-guided specialist LSTM	0.329	-92.0	-831	0.749	0.652	-2.40	-26.5	0.866	9

Linear regression was not applicable at sites TECR, BIGC, or WLOU due to the lack of donor gauges contemporary with target gauge data. Donor gauges associated with Kings River Experimental Watersheds exist within close proximity to TECR and BIGC, but we were unable to access up-to-date discharge records for these gauges.

The process-guided specialist LSTM yielded predictions on par with those of the other LSTM strategies in terms of KGE, (median 0.652; n = 9), but performed worst of the four in terms of NSE (median 0.329; n = 9). Conversely, the specialist performed better than the generalist in terms of NSE, but not KGE. The process-guided specialist LSTM strategy was viable at nine sites for which discharge estimates were available from the National Hydrologic Model.

In addition to improvements in accuracy, estimates from this study inform ~5,981 site-days (75%) of missing data in the official discharge record (Figure 3), though note that they also omit ~4,486 site-days otherwise present in NEON's official record. Omissions occur wherever observations are missing from the records of one or more donor gauges, and LSTM methods did not achieve desired efficiencies. Approximately 1,221 site-days are missing from the official record and from our reconstructions.

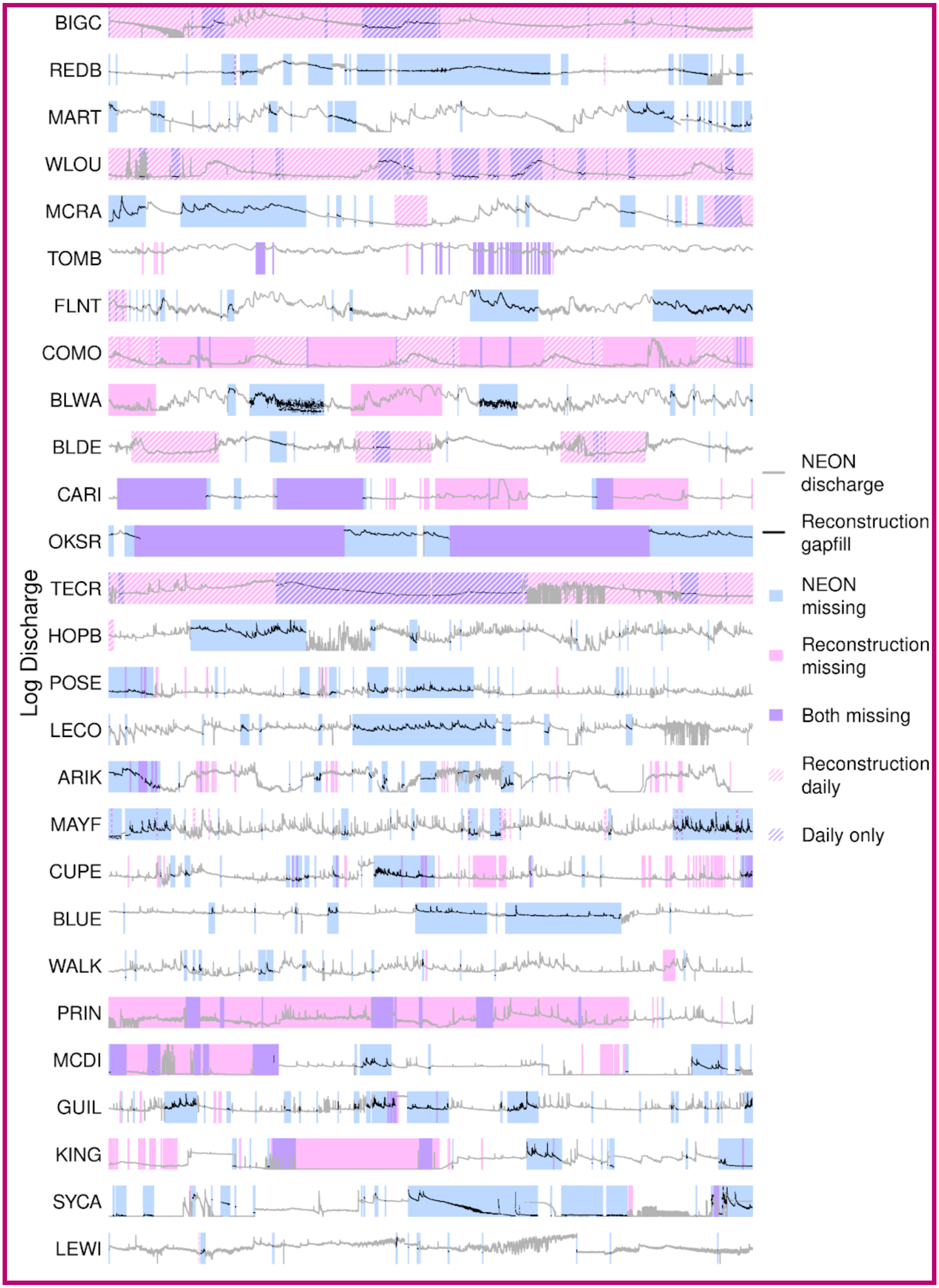

BIGC
REDB
MART
WLOU
MCRA
TOMB
FLNT
COMO
BLWA
BLDE
CARI
OKSR
TECR
HOPB
POSE
LECO
ARIK
MAYF
CUPE
BLUE
WALK
PRIN
MCDI
GUIL
KING
SYCA
LEWI
Log Discharge
NEON discharge
Reconstruction gapfill
NEON missing
Reconstruction missing
Both missing
Reconstruction daily
Daily only

Figure 3. Durations of missing values (gaps) in NEON's 2023 release of continuous discharge time series, illustrating gaps filled or informed by estimates from this analysis. All officially published values are shown, including those with quality control flags. Sites are ordered as in Figure 2. Gaps smaller than six hours are not indicated. Figure S10 is the same, but with a fixed and labeled x-axis.

Estimated discharge time series from this study are of practical value for any researcher using NEON continuous discharge data, especially for those sites and site-months at which published data from NEON's early operational phase may be unreliable (Rhea et al. 2023a). Figure 4 shows that official records at sites REDB and LEWI are compromised by disagreement (erratic sections of gray lines) between pressure transducer stage readings and manual gauge height recordings, discussed in Rhea et al (2023a). Red lines show improved estimates via linear regression on discharge from donor gauges. Sites FLNT and WALK show generally close agreement between NEON discharge and our regression estimates, but note uncertainty associated with high discharge values.

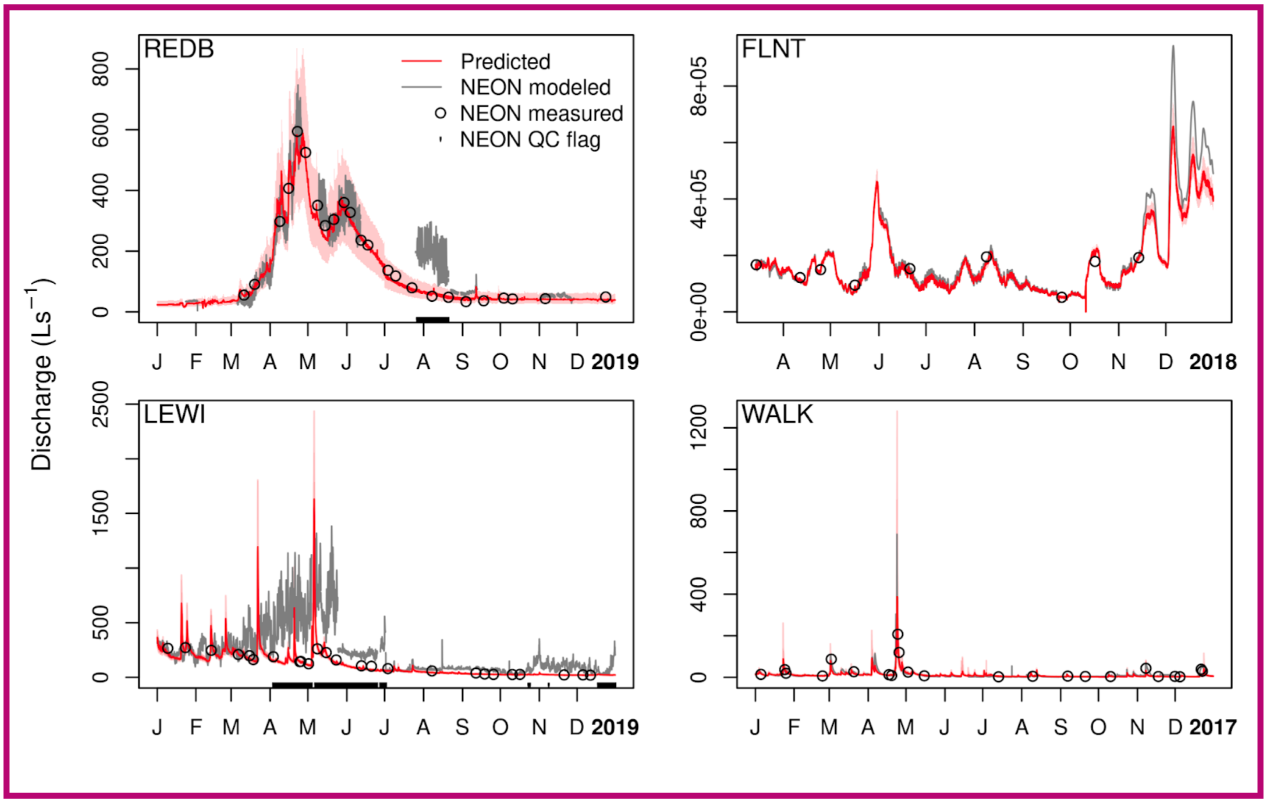


Figure 4. Best linear regression predictions of continuous discharge for four NEON gauge-years, compared with official NEON discharge data. All officially published values are shown, including those with quality control flags, indicated by black marks on lower border. Light red polygons represent 95% prediction intervals. NEON uncertainty is not shown.

Discussion

This study was designed to produce high-quality estimates of continuous discharge for NEON stream gauges, especially at ten gauges for which the KGE of published continuous discharge was lower than 0.8, over the full record, when compared to field-measured discharge. A secondary goal was to improve temporal coverage of the official discharge record where possible.

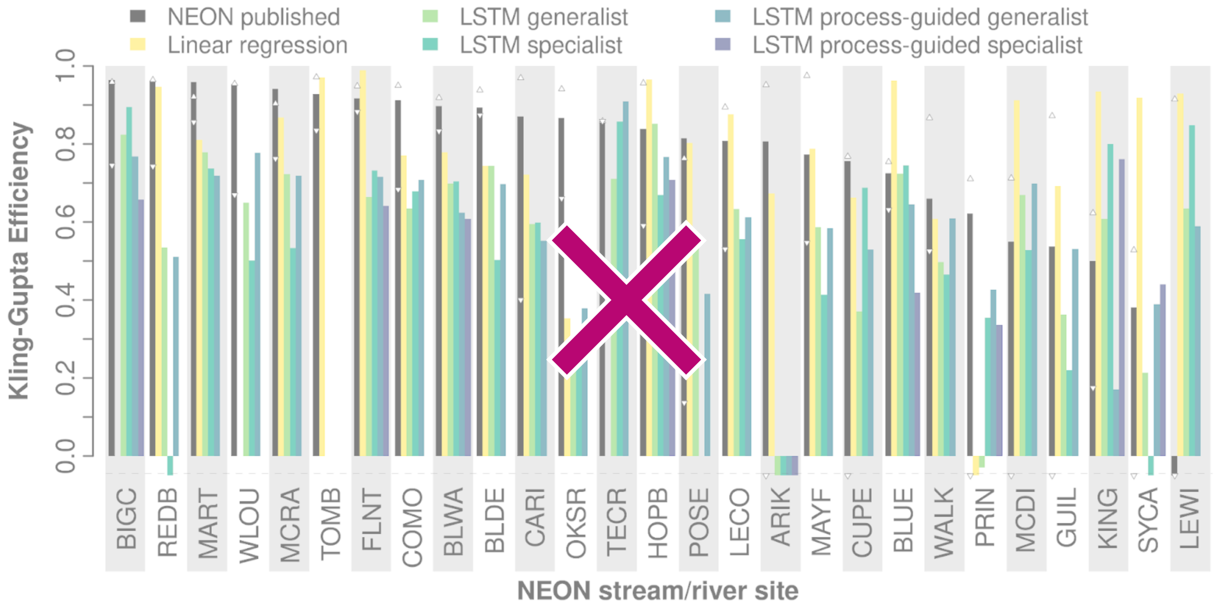
We treat NEON field-measured discharge as truth, which means there are 39-213 observations for each target site. Although these numbers represent a tremendous investment of time and technical effort, they do not meet the high data volume requirements for most machine learning approaches, so we used field discharge only to evaluate, rather than train, LSTM models. By contrast, in linear regression, regardless of the details of any particular method, we ultimately fit a line to the relationship between donor gauge data and field measurements at each target site. Because the linear regression models are allowed to "see" all of the target site data (after a model is selected via cross-validation), they have a powerful advantage over the LSTM approaches, which in this context must essentially treat target watersheds as if they are ungauged. Furthermore, whereas the LSTM models must parameterize each day of prediction individually, the regression models need only parameterize relationships between flow regimes. Still, if given enough training data, including examples of watersheds and streams similar to each of those modeled in this study, the LSTM approaches would eventually close the performance gap. See Figures S2, S3, S4, S5, S7, and S8 for linear regression diagnostics.

~~This study was designed to produce high-quality estimates of continuous discharge for NEON stream gauges, especially at ten gauges for which the KGE of published continuous discharge was lower than 0.8, over the full record, when compared to field-measured discharge. A secondary goal was to improve temporal coverage of the official discharge record where possible. Via linear regression on discharge from donor gauges, we were able to produce 15-minute discharge estimates at 11 sites with overall KGE scores higher than those of published series (Figure 2). At four of the same sites, we achieved higher KGE via LSTM methods, which generated daily discharge series. Of the ten sites at which published discharge KGE was less than 0.8, we improved five to above that mark (mean 0.932, n = 5).~~

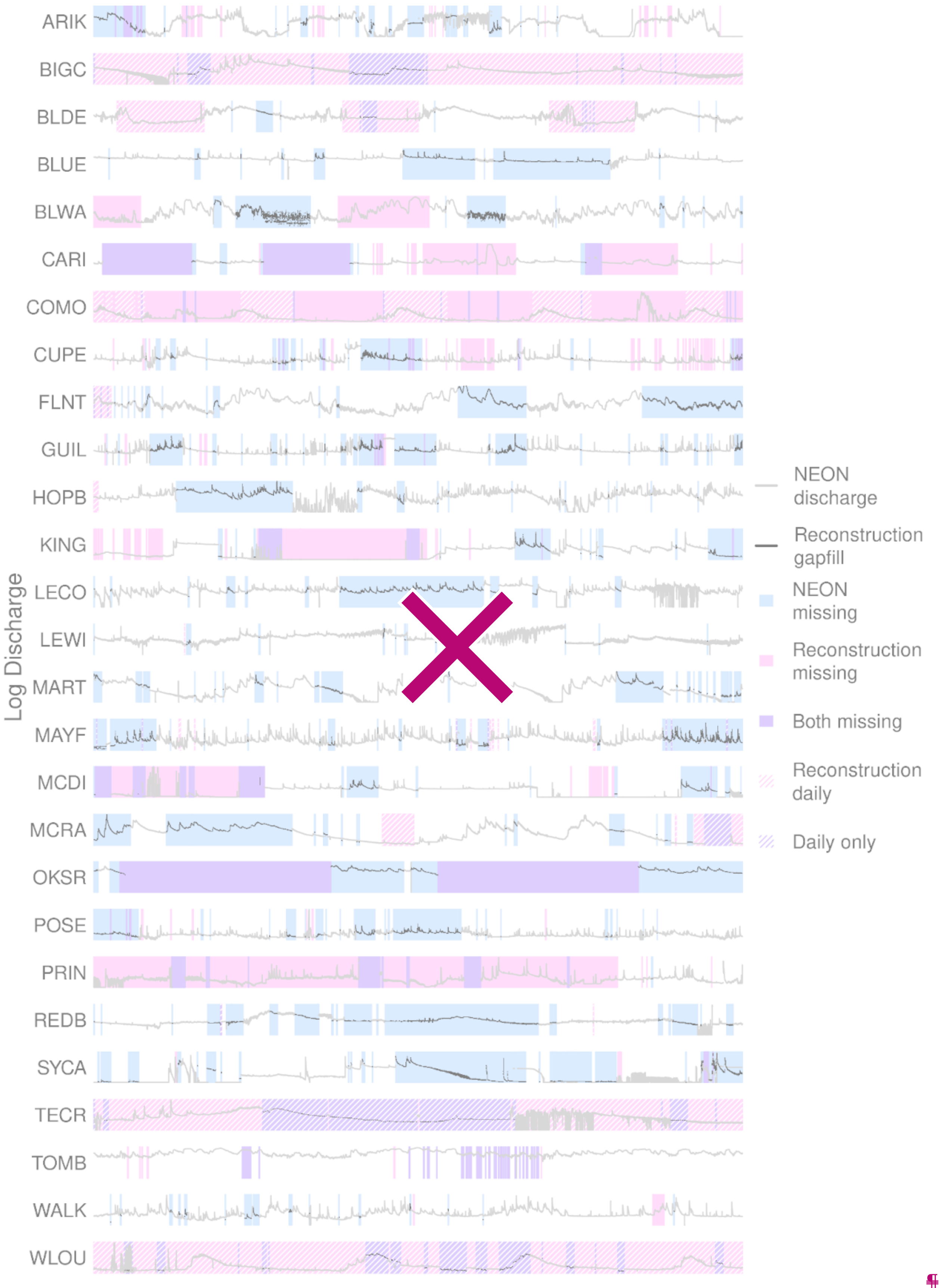
~~We caution that evaluation scores for both NEON's published estimates and ours are computed on a small fraction of each series for which both an estimate and a direct field measurement are available (39-213 per site), and that measurements tend to be collected disproportionately at low flow. Therefore, users of our composite discharge product should observe uncertainties associated with estimates from all methods.~~

~~In addition to improvements in accuracy, estimates from this study inform ~5,981 site-days (75%) of missing data in the official discharge record (Figure 3), though note that they also omit ~4,486 site-days otherwise present in NEON's official record. Omissions occur wherever observations are missing from the records of one or more donor gauges, and LSTM methods did not achieve desired efficiencies. Approximately 1,221 site-days are missing from the official record and from our reconstructions.~~¶

¶



~~Figure 2. Efficiency of five stream discharge prediction methods and NEON's published continuous discharge product at 27 NEON gauge locations, versus field-measured discharge. Small, white triangles represent max/min KGE of published discharge by water year (Oct 1 through Sept 30) with at least 5 field measurements (or 2 for site OKSR). NEON estimates with quality flags (dischargeFinalQF or dischargeFinalQFSciRvw of 1) were not included in KGE calculation. For the best performing LSTM method, at all sites except TECR, FLNT, REDB, WALK, POSE, and KING, displayed KGE is averaged over 30 ensemble runs with identical hyperparameters. For the sites just named, performance of a chosen method, after ensembling, dropped below that of at least one other method's optimal KGE from parameter search. For all other LSTM site-method pairs, which were not ensembled, displayed performance is that of the best model trained during the parameter search phase. Sites are ordered by the KGE of NEON continuous discharge. See Table 3 for LSTM model definitions. KGE of 1 is a perfect prediction, while KGE of -0.41 is similar in skill to prediction from the mean. Negative values are truncated at -0.05 in this plot to improve visualization.~~¶


Log Discharge
ARIK
BIGC
BLDE
BLUE
BLWA
CARI
COMO
CUPE
FLNT
GUIL
HOPB
KING
LECO
LEWI
MART
MAYF
MCDI
MCRA
OKSR
POSE
PRIN
REDB
SYCA
TECR
TOMB
WALK
WLOU
NEON discharge
Reconstruction gapfill
NEON missing
Reconstruction missing
Both missing
Reconstruction daily
Daily only

~~Figure 3. Durations of missing values (gaps) in NEON's 2023 release of continuous discharge time series, illustrating gaps filled or informed by estimates from this analysis. All officially published values are shown, including those with quality control flags. Gaps smaller than six hours are not indicated.¶~~

~~A performance comparison of linear regression and four LSTM strategies is shown in Figure 2 and Figure S1, and detailed in Table S1. For 12 of 27 sites, linear regression on specific discharge (i.e. scaled by watershed area) provided the most accurate discharge predictions, while linear regression on absolute discharge performed best at the other 12 sites with donor gauges. LSTM models (as proper ensembles) outperformed linear regression at only 2 sites. In this study, we treat NEON field-measured discharge as truth, which means there are 39-213 observations for each target site. Although these numbers represent a tremendous investment of time and technical effort, they do not meet the high data volume requirements for most machine learning approaches, so we used field discharge only to evaluate, rather than train, LSTM models. By contrast, in linear regression, regardless of the details of any particular method, we ultimately fit a line to the relationship between donor gauge data and field measurements at each target site. Because the linear regression models are allowed to "see" all of the target site data (after a model is selected via cross-validation), they have a powerful advantage over the LSTM approaches, which in this context must essentially treat target watersheds as if they are ungauged. See Figures S2, S3, S4, S5, S7, and S8 for linear regression diagnostics.~~

~~Linear regression was not applicable at sites TECR, BIGC, or WLOU due to the lack of donor gauges contemporary with target gauge data. Donor gauges associated with Kings River Experimental Watersheds exist within close proximity to TECR and BIGC, but we were unable to access up-to-date discharge records for these gauges.¶~~

~~¶~~

~~In general, linear regression provided more accurate predictions than all LSTM methods. Linear regression on absolute discharge produced estimates with median NSE of 0.848 and median KGE of 0.806, across sites (*n* = 24; Table 5). Linear regression on specific discharge (i.e. scaled by watershed area) produced similar median scores (Table 5), but with deviations of up to 0.05 NSE and 0.08 KGE at individual sites.¶~~

~~¶~~

~~The process-guided specialist LSTM strategy yielded predictions on par with those of the other LSTM strategies in terms of KGE, (median 0.652; *n* = 9), but performed worst of the four in terms of NSE (median 0.329; *n* = 9),~~

In this study, discharge estimates produced by linear regression were more accurate than those generated by LSTM models in 21 of 23 comparisons (Figure 2). This demonstrates the value of existing gauge networks in advancing discharge estimation at newly or partially gauged locations; however, there is a limit to the predictive potential of linear regression methods, as they depend on strong correlation between streamflow at target and donor gauges. In principle, there is no such limit for machine learning approaches, which are instead limited by the quality and quantity of training data.

The process-guided specialist LSTM yielded predictions on par with those of the other LSTM strategies in terms of KGE, but performed worst of the four in terms of NSE, possibly indicating that information gleaned from NHM estimates helped this strategy to accurately capture discharge variance and reduce prediction bias, without ultimately improving the correlation between predictions and observations. Unlike KGE, NSE only explicitly captures this latter metric (Nash & Sutcliffe 1970; Gupta et al. 2009). Conversely, the specialist performed better than the generalist in terms of NSE, but not KGE, ~~Conversely, the specialist performed better than the generalist in terms of NSE, but not KGE,~~ suggesting information contained in NEON's continuous discharge product~~estimates~~ was of disproportionate predictive value relative to each of correlation, variance, and bias, favoring correlation.

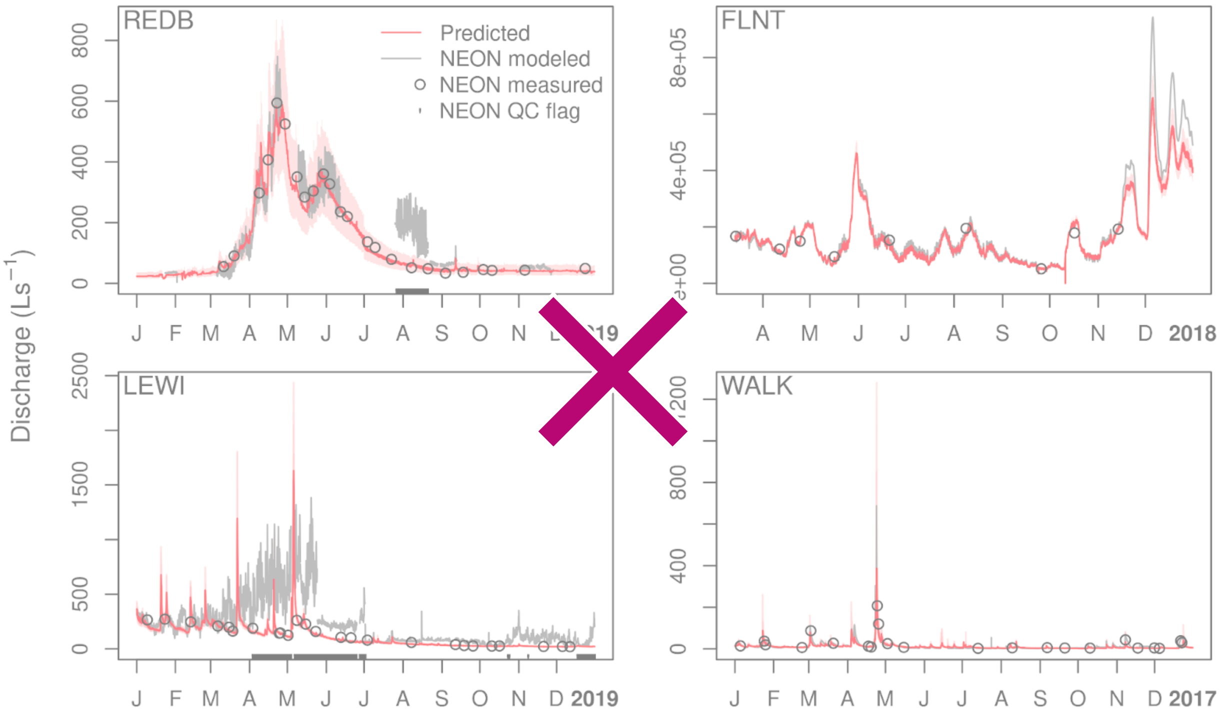
The specialist may have ~~also~~ been affected by data filtering choices. After filtering NEON continuous discharge for rating curve issues, drift, and quality flags, relatively few daily estimates were available for some sites (47-1642). Annual and seasonal variation in meteorological forcings and discharge in NEON sites' generally small, often mountainous watersheds may be large enough that finetuning a pretrained LSTM on a few hundred days of site-specific data reduces its ability to generalize at that site. Our specialist LSTM strategy in particular might be improved with a broader hyperparameter search, especially one that explores smaller learning rates. Ideally, site-specific finetuning should enable better prediction by allowing the network to assimilate information unique to the target site without corrupting previously learned generalities. For validation plots of all ensembled LSTMs, see Figure S6.

~~Table 5. Performance of five stream discharge prediction methods, and official continuous discharge time-series data, across *n* of 27 NEON gauge locations (final column). For both the Nash-Sutcliffe and Kling-Gupta Efficiency coefficients, a value of 1 indicates perfect prediction. A value of 0 NSE indicates that predictive skill is equivalent to prediction from the mean, while negative NSE is worse than mean prediction. This threshold lies at approximately -0.41 for KGE (Knoben et al. 2019). "Linreg" = linear regression on donor gauge discharge series, and "scaled" means predictor and response discharge were scaled by their respective watershed areas.~~

	~~NSE~~				**~~KGE~~**				
~~Model/Data~~	**~~Median~~**	**~~Mean~~**	**~~Min~~**	**~~Max~~**	**~~Median~~**	**~~Mean~~**	**~~Min~~**	**~~Max~~**	***~~n~~***
~~Official record~~	~~0.880~~	~~0.417~~	~~-9.95~~	~~0.989~~	~~0.839~~	~~0.711~~	~~-1.50~~	~~0.964~~	~~27~~
~~Linreg~~	~~0.848~~	~~0.760~~	~~-0.038~~	~~0.993~~	~~0.806~~	~~0.746~~	~~-0.697~~	~~0.988~~	~~24~~
~~Linreg scaled~~	~~0.847~~	~~0.757~~	~~-0.037~~	~~0.993~~	~~0.807~~	~~0.743~~	~~-0.695~~	~~0.989~~	~~24~~
~~Generalist LSTM~~	~~0.473~~	~~-18.8~~	~~-498~~	~~0.904~~	~~0.634~~	~~-0.220~~	~~-20.2~~	~~0.852~~	~~26~~
~~Specialist LSTM~~	~~0.477~~	~~-12.6~~	~~-307~~	~~0.920~~	~~0.556~~	~~-0.256~~	~~-15.7~~	~~0.895~~	~~25~~
~~Process-guided generalist LSTM~~	~~0.434~~	~~-31.3~~	~~-824~~	~~0.848~~	~~0.618~~	~~-0.453~~	~~-26.4~~	~~0.869~~	~~26~~
~~Process-guided specialist LSTM~~	~~0.329~~	~~-92.0~~	~~-831~~	~~0.749~~	~~0.652~~	~~-2.40~~	~~-26.5~~	~~0.866~~	~~9~~

~~The process-guided specialist LSTM strategy was viable at nine sites for which discharge estimates were available from the National Hydrologic Model.~~ The process-guided specialist LSTM strategy was viable at nine sites for which discharge estimates were available from the National Hydrologic Model. By using a mechanistic (i.e. process-based) model with higher spatial resolution than the NHM, it should be possible to apply this process-guided approach at more of the NEON sites. A potentially stronger process-guided approach would use mechanistic model predictions as features (predictors), rather than training targets, but that would require mechanistic model predictions concurrent with discharge series at target sites, whereas NHM predictions at the time of this writing are available only through the year 2016. For a summary of process-guided deep learning strategies, see the “Integrating Design” subsection of Appling et al. (2022).

~~Estimated discharge time series from this study are of practical value for any researcher using NEON continuous discharge data, especially for those sites and site-months at which published data from NEON’s early operational phase may be unreliable (Rhea et al. 2023a). Figure 4 shows that official records at sites REDB and LEWI are compromised by disagreement (erratic sections of gray lines) between pressure transducer stage readings and manual gauge height recordings, discussed in Rhea et al (2023a). Red lines show improved estimates via linear regression on discharge from donor gauges. Sites FLNT and WALK show generally close agreement between NEON discharge and our regression estimates, but note uncertainty associated with high discharge values.~~¶



¶

~~Figure 4. Best linear regression predictions of continuous discharge for four NEON gauge-years, compared with official NEON discharge data. All officially published values are shown, including those with quality control flags, indicated by black marks on lower border. Light red polygons represent 95% prediction intervals. NEON uncertainty is not shown.~~¶

We caution that evaluation scores for both NEON's published estimates and ours are computed on a small fraction of each series for which both an estimate and a direct field measurement are available (39-213 per site), and that measurements tend to be collected disproportionately at low flow. This often occurs for practical reasons such as site access and technician safety, but may also reflect a need to characterize the low-flow variability of the stage-discharge relationship in streams with unstable low-flow hydrologic controls, such as unconsolidated bed material.

Whatever the reason for less sampling at high flow~~For practical reasons, field discharge measurements are often collected disproportionately in low flow conditions. As a result~~, any model attempting to use field measurements to reconstruct continuous discharge will estimate with greater~~morehigher~~ uncertainty at high flow than at low, and~~. Therefore~~ users of our composite discharge product should observe uncertainties associated with estimates from all methods. Mechanistic models that proceed from physical principles, or data-driven approaches that can generalize from prior observations, do not in principle suffer this disadvantage, as they do not depend on observations from a target site. However, these approaches may not reliably generate strong predictions at all sites or under all conditions (Razavi & Coulibaly 2013; Kratzert et al. 2019b), and may produce erratic point estimates where conditions diverge from past observations. Hybrid approaches that successfully leverage field measurements, as well as physical principles or learned relationships, are likely to yield well-constrained predictions where our efforts did not.

This study demonstrates that, in proximity to established streamflow gauges, even simple statistical methods can be used to generate accurate, continuous discharge at "virtual gauges," where discrete discharge has been measured. The number of field measurements across sites in this study varies from 39 to 213, but the number required for virtual gauging may be substantially smaller even than the minimum of this range. If the discharge relationships between a target site and all donor gauges were perfectly linear or log-linear, they could in principle be established with only two precise measurements at the target site. More important than the quantity is the distribution of measurements across flow conditions, which should be sufficient to fully characterize all modeled discharge relationships and their linearity or lack thereof (Sauer 2002; Zakwan et al. 2017). Concretely, we advocate for "storm chasing," or disproportionately seeking to sample discharge under high-flow~~extreme~~ conditions, and during both rising and falling limbs of storm events, rather than routine sampling. Observed NEON flow conditions relative to predicted discharge can be seen in Figure S9. See Philip & McLaughlin (2018) for further commentary on establishing a virtual gauge network, and Seibert & Beven (2009) and Pool & Seibert (2021) for information on the number and statistical properties of discharge samples required to establish strong stage-discharge or discharge-discharge relationships.

Conclusions

Using linear regression on donor gauge data and LSTM-RNNs, we reconstructed continuous discharge at 5-minute and/or daily frequency for the 27 stream and river monitoring locations of the National Ecological Observatory Network (NEON) over the water years 2015-2022. Relative to field-measured discharge as ground truth, our estimates achieve higher Kling-Gupta efficiency than NEON's official continuous discharge at 11 sites. We also provide continuous discharge estimates for ~199 site-months for which no official values have been published. Estimates from this study can be used in conjunction with officially released NEON continuous discharge data to enhance the analytical potential of NEON's river and stream data products during its early operational phase. Toward that end, we provide composite discharge series for each site, incorporating the best available estimates across all methods used in this study and NEON's published estimates. Considering the lag of up to 2.5 years before provisional discharge data become fully quality controlled and officially released by NEON, our methods may also be used to increase the rate at which discharge-associated stream chemistry, dissolved gas, and water quality products become fully usable by the community. All data and results from this study can be downloaded from the Figshare collection at https://doi.org/10.6084/m9.figshare.c.6488065. Composite series can be visualized interactively at https://macrosheds.org/data/vlah_etal_2023_composites/. All code necessary to reproduce this analysis is archived at https://doi.org/10.5281/zenodo.7976251. A complete list of products and URLs can be found in Table S3.

In general, linear regression methods produced more accurate discharge estimates (median KGE: 0.79; median NSE: 0.81; *n* = 24 sites) than LSTM approaches due to the fact that regression models were able to fully leverage available field measurements as well as highly informative donor gauge data. Nonetheless, LSTM methods achieved median ensemble KGE of 0.71 and NSE of 0.56 across 18 sites, making their estimates a valuable supplement. Although LSTM-generated discharge series are of daily frequency, some users will prefer them to higher resolution regression estimates, as the latter may be subject to error in the event of highly localized precipitation events affecting either donor or target gauges, but not both.

Improvements to our design could be made in several ways. LSTM models could be exposed to additional training data, such as the recently published Caravan compendium of CAMELS offshoots (Kratzert et al. 2023) or future expansions of the MacroSheds dataset (Vlah et al. 2023). Neural networks trained on sub-daily inputs might be better equipped to exploit atmospheric-hydrological dynamics that respond to both daily and annual cycles. Linear regression methods too might be improved with the use of additional predictors, such as continuous water level or precipitation.

The success of simple statistical methods in generating high-quality continuous discharge time series demonstrates the viability of "virtual gauges," or locations at which a small number of field discharge measurements, in proximity to one or more established gauges, provide a basis for continuous discharge estimation in lieu of a gauging station. Virtual gauges have the potential to greatly expand the spatial coverage of continuous discharge data throughout the USA and any richly gauged region of the world.

Acknowledgements

The authors are grateful to the NeuralHydrology team for their efforts in democratizing deep learning for the hydrology community. We thank NEON, NCAR, NWIS, Niwot Ridge LTER, Andrews Forest LTER, and the USGS for generating the data, and the National Science Foundation for providing the funding that made this analysis possible. Special thank-yous to Dr. Parker Norton of the USGS for extracting all NHM-PRMS outputs used in this study.

The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle. This material is based in part upon work supported by the National Science Foundation through the NEON Program.

References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences 21, 5293–5313.

Appelhans, T., Detsch, F., Reudenbach, C., Woellauer, S., 2022. mapview: Interactive Viewing of Spatial Data in R.

Appling, A.P., Oliver, S.K., Read, J.S., Sadler, J.M., Zwart, J., 2022. Machine learning for understanding inland water quantity, quality, and ecology.

Arriagada, P., Karelovic, B., Link, O., 2021. Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. Journal of Hydrology 598, 126454.

Arsenault, R., Brissette, F., Martel, J.-L., 2018. The hazards of split-sample validation in hydrological model calibration. Journal of hydrology 566, 346–362.

Arsov, N., Mirceva, G., 2019. Network Embedding: An Overview. https://doi.org/10.48550/ARXIV.1911.11726

Aschonitis, V.G., Papamichail, D., Demertzi, K., Colombani, N., Mastrocicco, M., Ghirardini, A., Castaldelli, G., Fano, E.-A., 2017. High resolution global grids of revised Priestley-Taylor and Hargreaves-Samani coefficients for assessing ASCE-standardized reference crop evapotranspiration and solar radiation, links to ESRI-grid files. Supplement to: Aschonitis, VG et al. (2017): High-resolution global grids of revised Priestley-Taylor and Hargreaves-Samani coefficients for assessing ASCE-standardized reference crop evapotranspiration and solar radiation. Earth System Science Data, 9(2), 615-638, https://doi.org/10.5194/essd-9-615-2017. https://doi.org/10.1594/PANGAEA.868808

Benson, M.A., Dalrymple, T., 1967. General field and office procedures for indirect discharge measurements. US Govt. Print. Off.,.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of machine learning research 13.

Bukaveckas, P., Likens, G., Winter, T., Buso, D., 1998. A comparison of methods for deriving solute flux rates using long-term data from streams in the Mirror Lake watershed. Water, Air, and Soil Pollution 105, 277–293.

Caruana, R., 1998. Multitask learning. Springer.

Chokmani, K., Ouarda, T.B., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resources Research 40.

DeCicco, L.A., Lorenz, D., Hirsch, R.M., Watkins, W., Johnson, M., 2022. dataRetrieval: R packages for discovering and retrieving water data available from U.S. federal hydrologic web services. U.S. Geological Survey, Reston, VA. https://doi.org/10.5066/P9X4L3GE

Durand, M., Gleason, C.J., Pavelsky, T.M., de Moraes Frasson, R.P., Turmon, M.J., David, C.H., Altenau, E.H., Tebaldi, N., Larnier, K., Monnier, J., others, 2022. A framework for estimating global river discharge from the Surface Water and Ocean Topography satellite mission. Authorea Preprints.

Friedman, J., Tibshirani, R., Hastie, T., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33, 1–22. https://doi.org/10.18637/jss.v033.i01

Galton, F., 1886. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland 15, 246–263.

Goeman, J., Meijer, R., Chaturvedi, N., 2012. L1 and L2 penalized regression models. cran. r-project. or.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2017.06.031

Graf, W.H., 1984. Hydraulics of sediment transport. Water Resources Publication.

Gruber, M., 2017. Improving efficiency by shrinkage: The James–Stein and Ridge regression estimators. Routledge.

Guo, D., Johnson, F., Marshall, L., 2018. Assessing the potential robustness of conceptual rainfall-runoff models under a changing climate. Water Resources Research 54, 5030–5049.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology 377, 80–91.

Hall Jr, R.O., 2016. Metabolism of streams and rivers: Estimation, controls, and application, in: Stream Ecosystems in a Changing Environment. Elsevier, pp. 151–180.

Harvey, C.L., Dixon, H., Hannaford, J., 2012. An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. Hydrology Research 43, 618–636.

Hirsch, R.M., 1982. A comparison of four streamflow record extension techniques. Water Resources Research 18, 1081–1088.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H., Pierrefeu, G., 2018. Impact of stage measurement errors on streamflow uncertainty. Water Resources Research 54, 1952–1976.

Hsu, K., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall-runoff process. Water resources research 31, 2517–2530.

Isaacson, K., Coonrod, J., 2011. USGS streamflow data and modeling sand-bed rivers. Journal of Hydraulic Engineering 137, 847–851.

Johnson, S.L., Rothacher, J.S., Wondzell, S.M., 2020. Stream discharge in gaged watersheds at the HJ Andrews Experimental Forest, 1949 to present. https://doi.org/10.6073/PASTA/0066D6B04E736AF5F234D95D97EE84F3

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Knoben, W.J., Freer, J.E., Woods, R.A., 2019. Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences 23, 4323–4331.

Kratzert, F., Gauch, M., Nearing, G., Klotz, D., 2022. NeuralHydrology — A Python library for Deep Learning research in hydrology. Journal of Open Source Software 7, 4050. https://doi.org/10.21105/joss.04050

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., Klambauer, G., 2019a. NeuralHydrology–interpreting LSTMs in hydrology. Explainable AI: Interpreting, explaining and visualizing deep learning 347–362.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019b. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research 55, 11344–11354.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., others, 2023. Caravan-A global community dataset for large-sample hydrology. Scientific Data 10, 61.

Lunch, C., Laney, C., Mietkiewicz, N., Sokol, E., Cawley, K., NEON (National Ecological Observatory Network), 2022. neonUtilities: Utilities for Working with NEON Data.

Manning, R., 1891. On the flow of water in open channels and pipes 20, 161–207.

Moore, S.A., Jamieson, E.C., Rainville, F., Rennie, C.D., Mueller, D.S., 2017. Monte Carlo approach for uncertainty analysis of acoustic Doppler current profiler discharge measurement by moving boat. Journal of Hydraulic Engineering 143, 04016088.

Moriasi, D., Gitau, M., Pai, N., Daggupati, P., 2015. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. Transactions of the ASABE (American Society of Agricultural and Biological Engineers) 58, 1763–1785. https://doi.org/10.13031/trans.58.10715

Muggeo, V.M.R., 2008. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News 8, 20–25.

Nalley, D., Adamowski, J., Khalil, B., Biswas, A., 2020. A comparison of conventional and wavelet transform based methods for streamflow record extension. Journal of Hydrology 582, 124503.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology 10, 282–290.

NEON (National Ecological Observatory Network) ~~(NEON)~~, 2023a. Continuous discharge (DP4.00130.001), RELEASE-2023. https://doi.org/10.48443/H2ZE-2F12. Data accessed from https://data.neonscience.org/data-products/DP1.00130.001/RELEASE-2023 on May 5, 2023.

NEON (National Ecological Observatory Network) ~~(NEON)~~, 2023b. Discharge field collection (DP1.20048.001), RELEASE-2023. https://doi.org/10.48443/TYS0-ZE83. Data accessed from https://data.neonscience.org/data-products/DP1.20048.001/RELEASE-2023 on January 31, 2023.

NEON (National Ecological Observatory Network), 2023c. Discharge field collection (DP1.20048.001), PROVISIONAL. Data accessed from https://data.neonscience.org/data-products/DP1.20048.001/RELEASE-2023 on January 31, 2023. Data archived at https://dx.doi.org/10.6084/m9.figshare.22344589.

Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J., others, 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences 19, 209–223.

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R., Blodgett, D., 2014. A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. UCAR/NCAR: Boulder, CO, USA.

Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B., Nearing, G., 2017. Benchmarking of a physically based hydrologic model. Journal of Hydrometeorology 18, 2215–2225.

Odum, H.T., 1956. Primary production in flowing waters 1. Limnology and oceanography 1, 102–117.

Pantelakis, D., Doulgeris, C., Hatzigiannakis, E., Arampatzis, G., 2022. Evaluation of discharge measurements methods in a natural river of low or middle flow using an electromagnetic flow meter. River Research and Applications 38, 1003–1013.

Philip, E., McLaughlin, J., 2018. Evaluation of stream gauge density and implementing the concept of virtual gauges in Northern Ontario for watershed modeling. Journal of Water Management Modeling.

Pool, S., Seibert, J., 2021. Gauging ungauged catchments–Active learning for the timing of point discharge observations in combination with continuous water level measurements. Journal of Hydrology 598, 126448.

Razavi, T., Coulibaly, P., 2013. Streamflow prediction in ungauged basins: review of regionalization methods. Journal of hydrologic engineering 18, 958–975.

Read, J.S., Jia, X., Willard, J., Appling, A.P., Zwart, J.A., Oliver, S.K., Karpatne, A., Hansen, G.J.A., Hanson, P.C., Watkins, W., Steinbach, M., Kumar, V., 2019. Process-Guided Deep Learning

Predictions of Lake Water Temperature. Water Resources Research 55, 9173–9190. https://doi.org/10.1029/2019WR024922

Regan, R.S., Juracek, K.E., Hay, L.E., Markstrom, S., Viger, R.J., Driscoll, J.M., LaFontaine, J., Norton, P.A., 2019. The US Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States. Environmental Modelling & Software 111, 192–203.

Rhea, S., 2023. NEON Continuous Discharge Evaluation. HydroShare, http://www.hydroshare.org/resource/1a388391632f4277992889e2de152163. Accessed 2023-04-14.

Rhea, S., Gubbins, N., DelVecchia, A.G., Ross, M.R., Bernhardt, E.S., 2023a. User-focused evaluation of National Ecological Observatory Network streamflow estimates. Scientific Data 10, 89.

Rhea, S., Vlah, M., Slaughter, W., Gubbins, N., 2023b. macrosheds: Tools for interfacing with the MacroSheds dataset.

Sadler, J.M., Appling, A.P., Read, J.S., Oliver, S.K., Jia, X., Zwart, J.A., Kumar, V., 2022. Multi-task deep learning of daily streamflow and water temperature. Water Resources Research 58, e2021WR030138.

Sauer, V.B., 2002. Standards for the analysis and processing of surface-water data and information using electronic methods. US Geological Survey.

Seibert, J., Beven, K.J., 2009. Gauging the ungauged basin: how many discharge measurements are needed? Hydrology and Earth System Sciences 13, 883–892.

Seibert, J., Strobl, B., Etter, S., Hummer, P., van Meerveld, H.J. (Ilja), 2019. Virtual Staff Gauges for Crowd-Based Stream Level Observations. Frontiers in Earth Science 7. https://doi.org/10.3389/feart.2019.00070

Seibert, J., Vis, M.J.P., Lewis, E., van Meerveld, H.J. 2018. Upper and lower benchmarks in hydrological modelling. Hydrological Processes 32, 1120–1125. https://doi.org/10.1002/hyp.11476

Shen, H., Tolson, B.A., Mai, J., 2022. Time to update the split-sample approach in hydrological model calibration. Water Resources Research 58, e2021WR031523.

Shen, J., 1981. Discharge characteristics of triangular-notch thin-plate weirs. United States Department of the Interior, Geological Survey.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

Tazioli, A., 2011. Experimental methods for river discharge measurements: comparison among tracers and current meter. Hydrological Sciences Journal 56, 1314–1324.

Thornton, M.M., Shrestha, R., Wei, Y., Thornton, P.E., Kao, S.-C., Wilson, B.E., 2022. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1. https://doi.org/10.3334/ORNLDAAC/2129

Turnipseed, D.P., Sauer, V.B., 2010. Discharge measurements at gaging stations. US Geological Survey.

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vlah, M.J., Rhea, S., Bernhardt, E.S., Slaughter, W., Gubbins, N., DelVecchia, A.G., Thellman, A., Ross, M.R., 2023. MacroSheds: A synthesis of long-term biogeochemical, hydroclimatic, and geospatial data from small watershed ecosystem studies. Limnology and Oceanography Letters.

Wang, C.P., 1988. Laser doppler velocimetry. Journal of Quantitative Spectroscopy and Radiative Transfer 40, 309–319.

White, A.F., Blum, A.E., 1995. Effects of climate on chemical_ weathering in watersheds. Geochimica et Cosmochimica Acta 59, 1729–1747.

Whittaker, J., Whitehead, C., Somers, M., 2005. The neglog transformation and quantile regression for the analysis of a large credit scoring database. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54, 863–878.

Zakwan, M., Muzzammil, M., Alam, J., 2017. Developing stage-discharge relations using optimization techniques. Aquademia: Water, Environment and Technology 1, 05.

Supplemental Tables

Table S1: Methods from this study used in the construction of composite discharge series. Composite series also incorporate NEON continuous discharge product DP4.001~~2~~30.001 (NEON 2023a~~; 2023 release accessed 2023-05-01~~). “Linreg” = linear regression; “glmnet” = r~~R~~idge regression; “lm” = OLS regression; “segmented” = segmented regression; “abs” = absolute discharge; “spec” = specific discharge; “pgdl” = process-guided deep learning.

Site	KGE linreg	NSE linreg	Method linreg	KGE LSTM	NSE LSTM	Method LSTM
FLNT	0.989	0.980	glmnet_spec	0.664	0.507	generalist
TOMB	0.970	0.993	glmnet_abs			
HOPB	0.966	0.937	lm_abs	0.852	0.704	generalist
BLUE	0.962	0.932	lm_spec	0.746	0.567	specialist
REDB	0.946	0.973	lm_abs	0.511	0.551	generalist_pgdl
KING	0.935	0.888	glmnet_abs			
LEWI	0.929	0.875	glmnet_abs	0.848	0.724	specialist
SYCA	0.919	0.938	segmented_spec			
MCDI	0.912	0.897	glmnet_spec			
LECO	0.877	0.833	lm_spec			

MCRA	0.868	0.866	glmnet_spec	0.723	0.531	generalist
MART	0.811	0.706	glmnet_spec	0.779	0.566	generalist
POSE	0.803	0.648	glmnet_spec			
MAYF	0.787	0.806	glmnet_abs	0.586	0.666	generalist
BLWA	0.779	0.892	glmnet_abs			
COMO	0.771	0.806	glmnet_composite_spec			
BLDE	0.744	0.863	glmnet_abs	0.744	0.687	generalist
CARI	0.721	0.637	glmnet_abs			
GUIL	0.692	0.653	glmnet_abs			
ARIK	0.674	0.596	glmnet_abs			
CUPE	0.663	0.728	glmnet_spec			
WALK	0.607	0.532	glmnet_spec			
BIGC				0.895	0.827	specialist
WLOU				0.778	0.596	generalist_pgdl
TECR				0.711	0.904	generalist
PRIN						
OKSR						

Table S2. Model input data used in this study.

Resource	Description	Source/Link
NEON discharge field collection	Discharge measurements from field-based surveys	NEON 2023b, NEON 2023c
NEON continuous discharge	Discharge calculated from a rating curve and sensor measurements of water level	NEON 2023a
User-focused evaluation of NEON streamflow estimates	3-tier classification of the reliability of NEON continuous discharge by site-month	https://www.nature.com/articles/s41597-023-01983-w
CAMELS dataset	Catchment Attributes, Meteorology, (and streamflow) for Large-sample Studies	https://ral.ucar.edu/solutions/products/camels

National Hydrologic Model (NHM)	USGS infrastructure that, when coupled with the Precipitation-Runoff Modeling System, can produce streamflow simulations at local to national scale	https://www.usgs.gov/mission-areas/water-resources/science/national-hydrologic-model-infrastructure
MacroSheds	A synthesis of long-term biogeochemical, hydroclimatic, and geospatial data from small watershed ecosystem studies	https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=1262
Daymet	Gridded estimates of daily weather parameters	https://developers.google.com/earth-engine/datasets/catalog/NASA_ORNL_DAYMET_V4
HJ Andrews Experimental Forest stream discharge	Stream discharge in gaged watersheds, 1949 to present	https://portal.edirepository.org/nis/mapbrowse?packageid=knb-lter-and.4341.33
USGS National Water Information System	Streamflow and associated data for thousands of gauged streams and rivers within the USA	https://waterdata.usgs.gov/nwis, e.g. https://waterdata.usgs.gov/monitoring-location/06879100/

Table S3. Products of this study.

Product	Description	Link
Data archive landing page	Figshare page linking to each of four archives described below	https://doi.org/10.6084/m9.figshare.c.6488065
Composite discharge timeseries	Analysis-ready CSVs combining the best available discharge estimates across linear regression and LSTM approaches from this study, and NEON's published data	https://doi.org/10.6084/m9.figshare.23206592.v1
Composite discharge plots	Interactive plots of our composite discharge product	https://macrosheds.org/data/vlah_etal_2023_composites

All model outputs and results	Complete predictions from all linear regression and LSTM models, run results, and diagnostics	https://doi.org/10.6084/m9.figshare.22344589.v1
All model input data	Donor gauge streamflow, training data for LSTMs, model configurations, etc.	https://doi.org/10.6084/m9.figshare.22349377.v1
All code associated with this paper	Zenodo archive of GitHub repository	https://doi.org/10.5281/zenodo.7976251
All figures associated with this paper	High-resolution images of all figures from the main body and appendix	https://doi.org/10.6084/m9.figshare.23169362.v1

Supplemental Figures

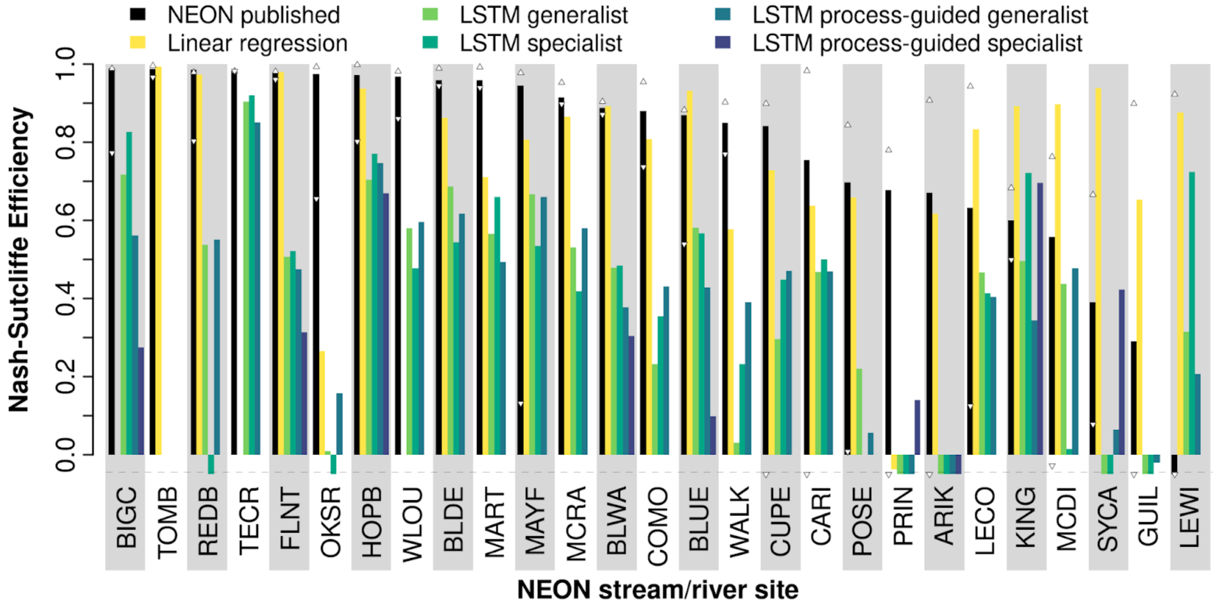
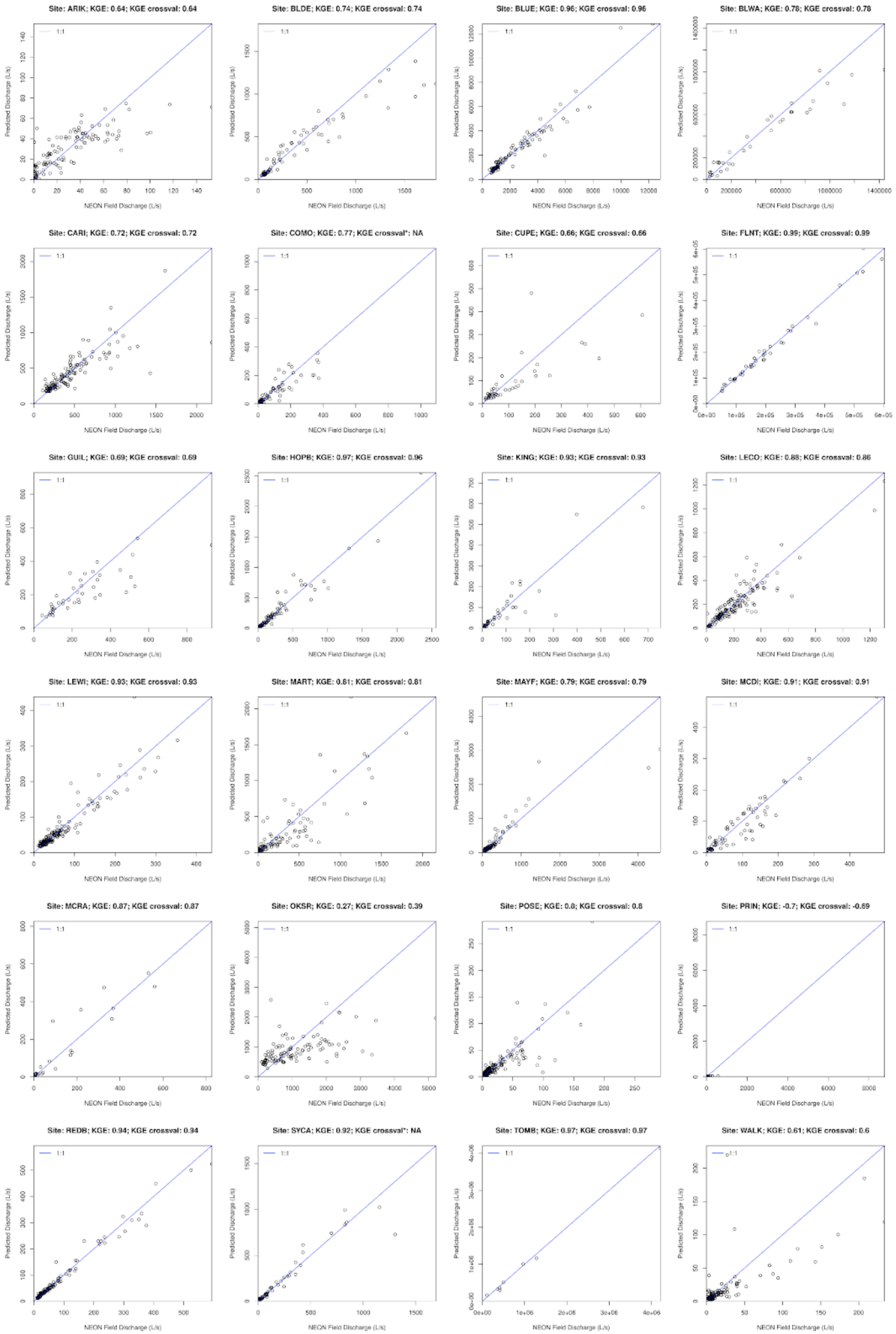


Figure S1. Efficiency of five stream discharge prediction methods and NEON's published continuous discharge product at 27 NEON gauge locations, versus field-measured discharge. Small, white triangles represent max/min NSE of published discharge by water year (Oct 1 through Sept 30) with at least 5 field measurements (or 2 for site OKSR). NSE was computed on all available observation-estimate pairs except those with quality flags (dischargeFinalQF or dischargeFinalQFSciRvw of 1).~~NEON estimates with quality flags (dischargeFinalQF or dischargeFinalQFSciRvw of 1) were not included in NSE calculation~~. For the best performing LSTM method, at all sites except TECR, FLNT, REDB, WALK, POSE, and KING, displayed NSE is averaged over 30 ensemble runs with identical hyperparameters. For the sites just named, performance of a chosen method, after ensembling, dropped below that of at least one other

method’s optimal NSE from parameter search. For all other LSTM site-method pairs, which were not ensembled, displayed performance is that of the best model trained during the parameter search phase. Sites are ordered by the NSE of NEON continuous discharge. See Table 3 for LSTM model definitions. NSE of 1 is a perfect prediction, while NSE of 0 is equivalent in skill to prediction from the mean. Negative values are truncated at -0.05 in this plot to improve visualization.


Site: ARIK; KGE: 0.64; KGE crossval: 0.64

Site: BLDE; KGE: 0.74; KGE crossval: 0.74

Site: BLUE; KGE: 0.96; KGE crossval: 0.96

Site: BLWA; KGE: 0.78; KGE crossval: 0.78

Site: CARI; KGE: 0.72; KGE crossval: 0.72

Site: COMO; KGE: 0.77; KGE crossval*: NA

Site: CUPE; KGE: 0.66; KGE crossval: 0.66

Site: FLNT; KGE: 0.99; KGE crossval: 0.99

Site: GUIL; KGE: 0.69; KGE crossval: 0.69

Site: HOPB; KGE: 0.97; KGE crossval: 0.96

Site: KING; KGE: 0.93; KGE crossval: 0.93

Site: LECO; KGE: 0.88; KGE crossval: 0.86

Site: LEWI; KGE: 0.93; KGE crossval: 0.93

Site: MART; KGE: 0.81; KGE crossval: 0.81

Site: MAYF; KGE: 0.79; KGE crossval: 0.79

Site: MCDI; KGE: 0.91; KGE crossval: 0.91

Site: MCRA; KGE: 0.87; KGE crossval: 0.87

Site: OKSR; KGE: 0.27; KGE crossval: 0.39

Site: POSE; KGE: 0.8; KGE crossval: 0.8

Site: PRIN; KGE: -0.7; KGE crossval: -0.69

Site: REDB; KGE: 0.94; KGE crossval: 0.94

Site: SYCA; KGE: 0.92; KGE crossval*: NA

Site: TOMB; KGE: 0.97; KGE crossval: 0.97

Site: WALK; KGE: 0.61; KGE crossval: 0.6

Figure S2. Observed (field) discharge vs. predictions from linear regression on specific discharge (i.e. scaled by watershed area).

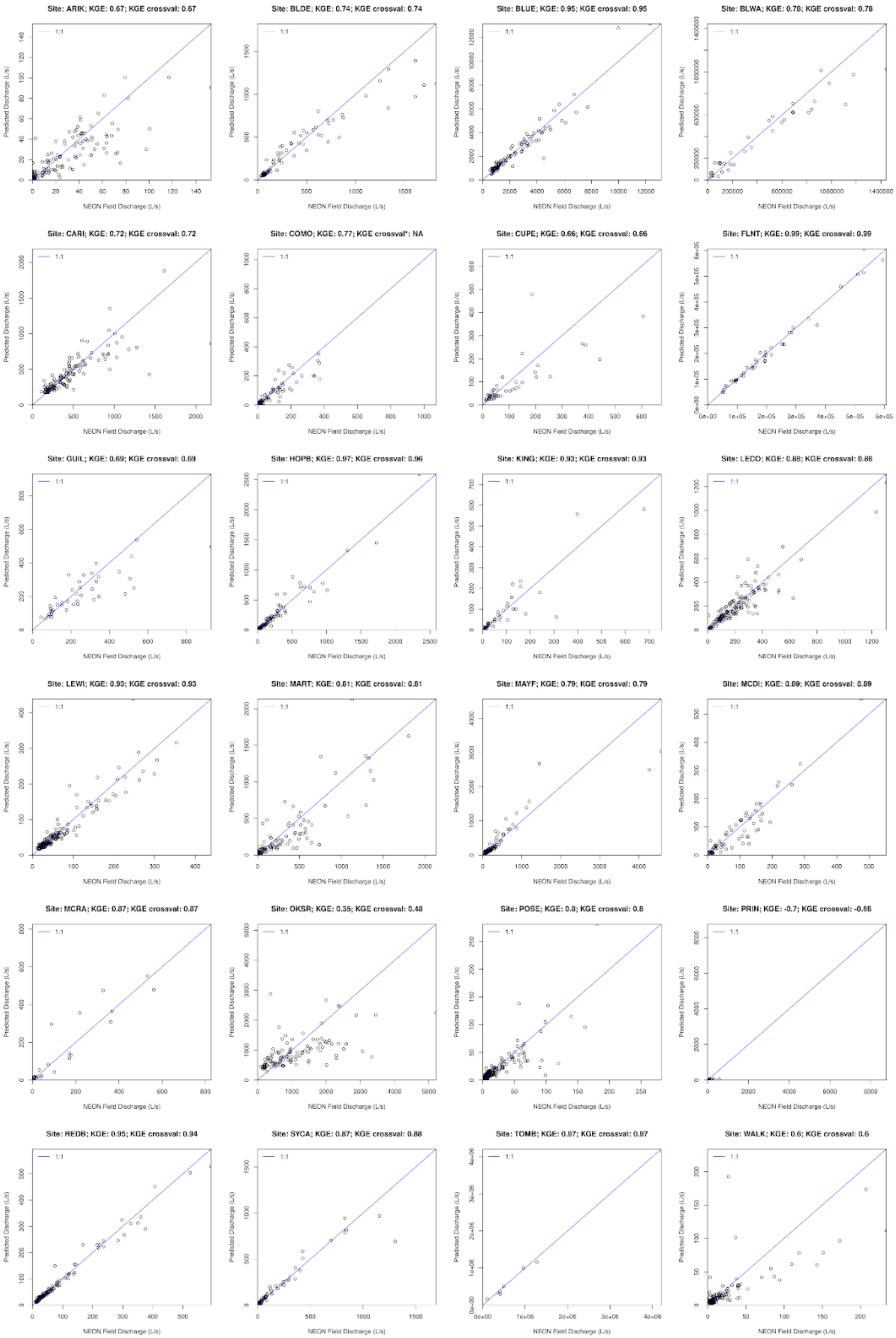

Site: ARIK; KGE: 0.67; KGE crossval: 0.67
Site: BLDE; KGE: 0.74; KGE crossval: 0.74
Site: BLUE; KGE: 0.95; KGE crossval: 0.95
Site: BLWA; KGE: 0.78; KGE crossval: 0.78
Site: CARI; KGE: 0.72; KGE crossval: 0.72
Site: COMO; KGE: 0.77; KGE crossval*: NA
Site: CUPE; KGE: 0.66; KGE crossval: 0.66
Site: FLNT; KGE: 0.99; KGE crossval: 0.99
Site: GUIL; KGE: 0.69; KGE crossval: 0.69
Site: HOPB; KGE: 0.97; KGE crossval: 0.96
Site: KING; KGE: 0.93; KGE crossval: 0.93
Site: LECO; KGE: 0.88; KGE crossval: 0.86
Site: LEWI; KGE: 0.93; KGE crossval: 0.93
Site: MART; KGE: 0.81; KGE crossval: 0.81
Site: MAYF; KGE: 0.79; KGE crossval: 0.79
Site: MCDI; KGE: 0.89; KGE crossval: 0.89
Site: MCRA; KGE: 0.87; KGE crossval: 0.87
Site: OKSR; KGE: 0.35; KGE crossval: 0.48
Site: POSE; KGE: 0.8; KGE crossval: 0.8
Site: PRIN; KGE: -0.7; KGE crossval: -0.68
Site: REDB; KGE: 0.95; KGE crossval: 0.94
Site: SYCA; KGE: 0.87; KGE crossval: 0.88
Site: TOMB; KGE: 0.97; KGE crossval: 0.97
Site: WALK; KGE: 0.6; KGE crossval: 0.6
1:1
Predicted Discharge (L/s)
NEON Field Discharge (L/s)

Figure S3. Observed (field) discharge vs. predictions from linear regression on absolute discharge (i.e. not scaled by watershed area).

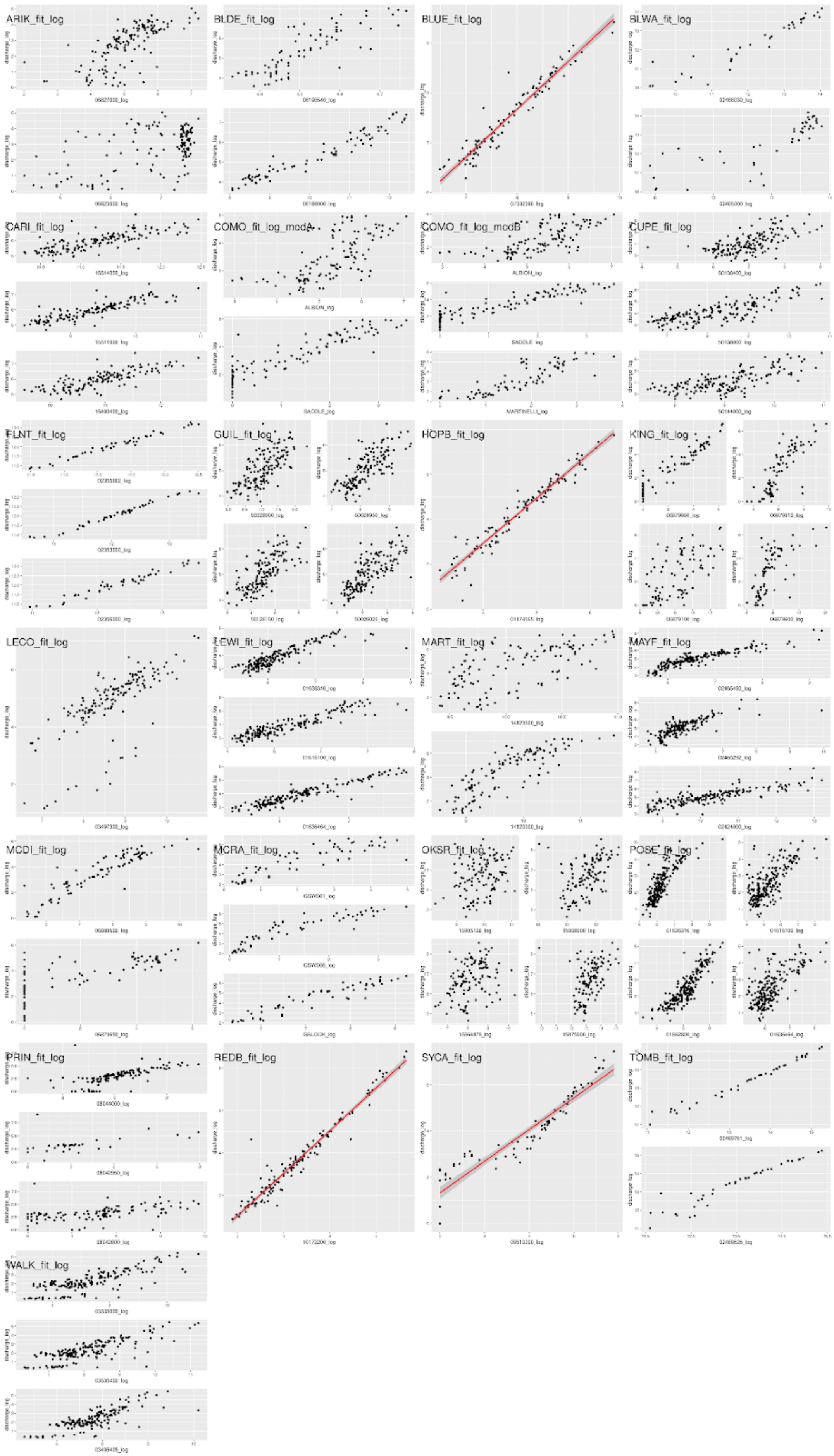
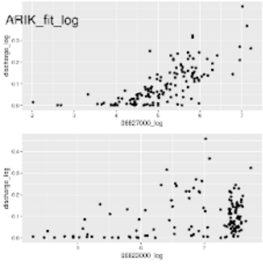
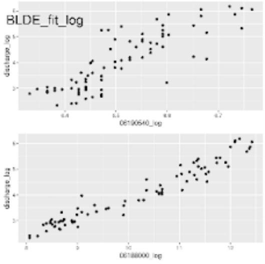
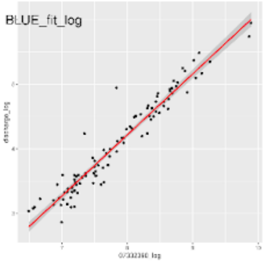
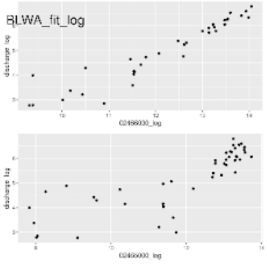

ARIK_fit_log
06827000_log
06823000_log
BLDE_fit_log
06190540_log
06188000_log
BLUE_fit_log
07332390_log
BLWA_fit_log
02466030_log
02465000_log
CARI_fit_log
15514000_log
15511000_log
15493400_log
COMO_fit_log_modA
ALBION_log
SADDLE_log
COMO_fit_log_modB
ALBION_log
SADDLE_log
MARTINELLI_log
CUPE_fit_log
50136400_log
50138000_log
50144000_log
FLNT_fit_log
02355662_log
02353000_log
02356000_log
GUIL_fit_log
50028000_log
50024950_log
50126150_log
50026025_log
HOPB_fit_log
01174565_log
KING_fit_log
06879650_log
06879810_log
06879100_log
06878600_log
LECO_fit_log
03497300_log
LEWI_fit_log
01636316_log
01616100_log
01636464_log
MART_fit_log
14123500_log
14120000_log
MAYF_fit_log
02465493_log
02465292_log
02424000_log
MCDI_fit_log
06888500_log
06879650_log
MCRA_fit_log
GSWS01_log
GSWS08_log
GSLOOK_log
OKSR_fit_log
15905100_log
15908000_log
15904870_log
15875000_log
POSE_fit_log
01636316_log
01616100_log
01632900_log
01636464_log
PRIN_fit_log
08044000_log
08045850_log
08042800_log
REDB_fit_log
10172200_log
SYCA_fit_log
09513200_log
TOMB_fit_log
02469761_log
02469525_log
WALK_fit_log
03535000_log
03535400_log
03495405_log
discharge_log

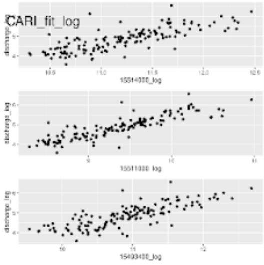
Figure S4: Marginal relationships between donor and target gauges for regression on specific discharge. Regression lines are shown only for single-donor regressions, fitted via OLS. Site SYCA, here exhibiting a breakpoint, was modeled with segmented regression, and thus the regression line shown has no relevance.

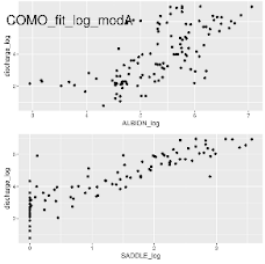

ARIK_fit_log
discharge_log
06827000_log
06823000_log

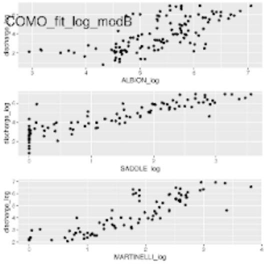

BLDE_fit_log
discharge_log
06190540_log
06188000_log

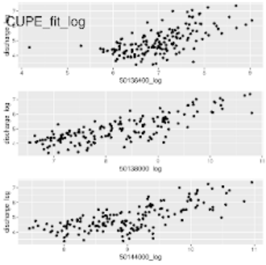

BLUE_fit_log
discharge_log
07332390_log

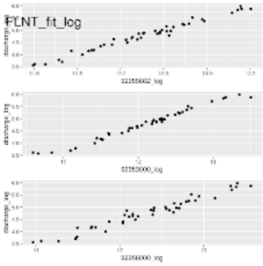

BLWA_fit_log
discharge_log
02466030_log
02465000_log

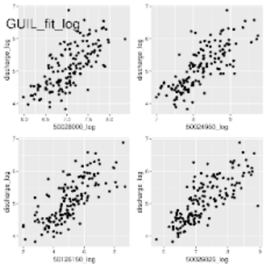

CARI_fit_log
discharge_log
15514000_log
15511000_log
15493400_log

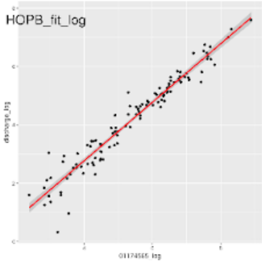

COMO_fit_log_modA
discharge_log
ALBION_log
SADDLE_log

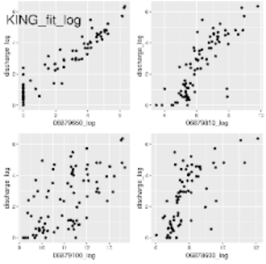

COMO_fit_log_modB
discharge_log
ALBION_log
SADDLE_log
MARTINELLI_log

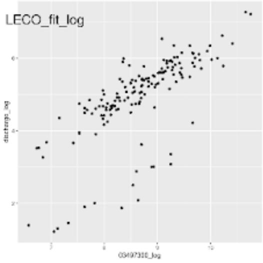

CUPE_fit_log
discharge_log
50136400_log
50138000_log
50144000_log

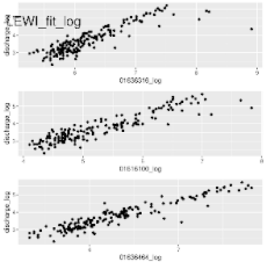

FLNT_fit_log
discharge_log
02355662_log
02353000_log
02356000_log

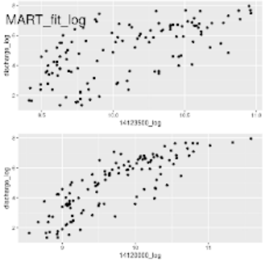

GUIL_fit_log
discharge_log
50028000_log
50024950_log
50126150_log
50026025_log

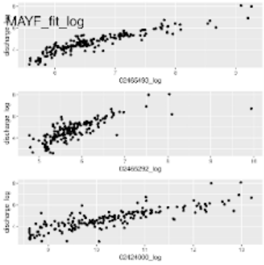

HOPB_fit_log
discharge_log
01174565_log

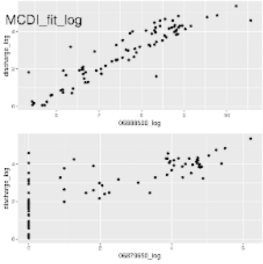

KING_fit_log
discharge_log
06879650_log
06879810_log
06879100_log
06878600_log

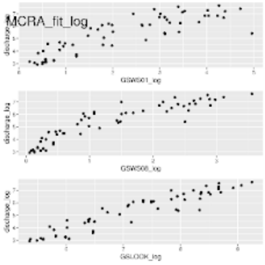

LECO_fit_log
discharge_log
03497300_log

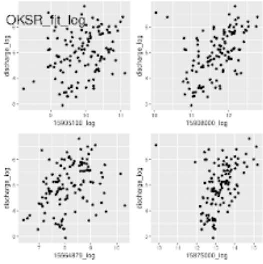

LEWI_fit_log
discharge_log
01636316_log
01616100_log
01636464_log

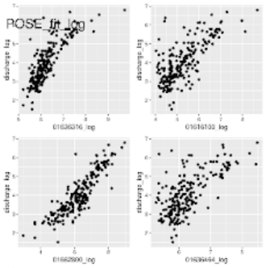

MART_fit_log
discharge_log
14123500_log
14120000_log

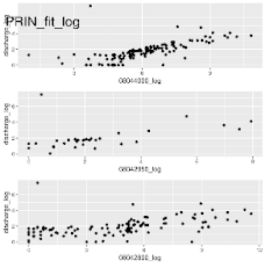

MAYF_fit_log
discharge_log
02465493_log
02465292_log
02424000_log

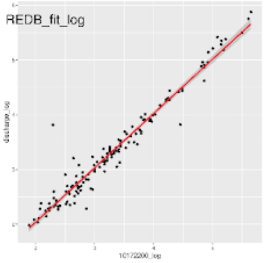

MCDI_fit_log
discharge_log
06888500_log
06879650_log

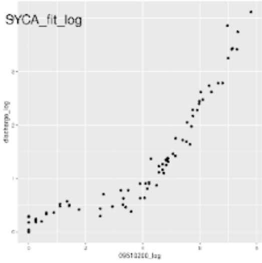

MCRA_fit_log
discharge_log
GSWS01_log
GSWS08_log
GSLOOK_log

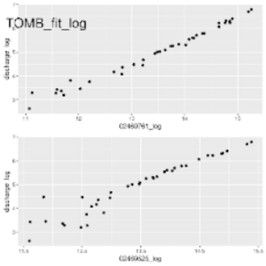

OKSR_fit_log
discharge_log
15905100_log
15908000_log
15564879_log
15875000_log


ROSE_fit_log
discharge_log
01636316_log
01616100_log
01662800_log
01636464_log


PRIN_fit_log
discharge_log
08044000_log
08042950_log
08042800_log


REDB_fit_log
discharge_log
10172200_log


SYCA_fit_log
discharge_log
09510200_log


TOMB_fit_log
discharge_log
02469761_log
02469525_log

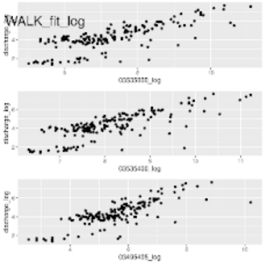

WALK_fit_log
discharge_log
03535000_log
03535400_log
03495405_log

Figure S5: Marginal relationships between donor and target gauges for regression on absolute discharge. Regression lines are shown only for single-donor regressions, fitted via OLS. Site SYCA, here exhibiting a breakpoint, could not be fitted via segmented regression in the context of absolute discharge.

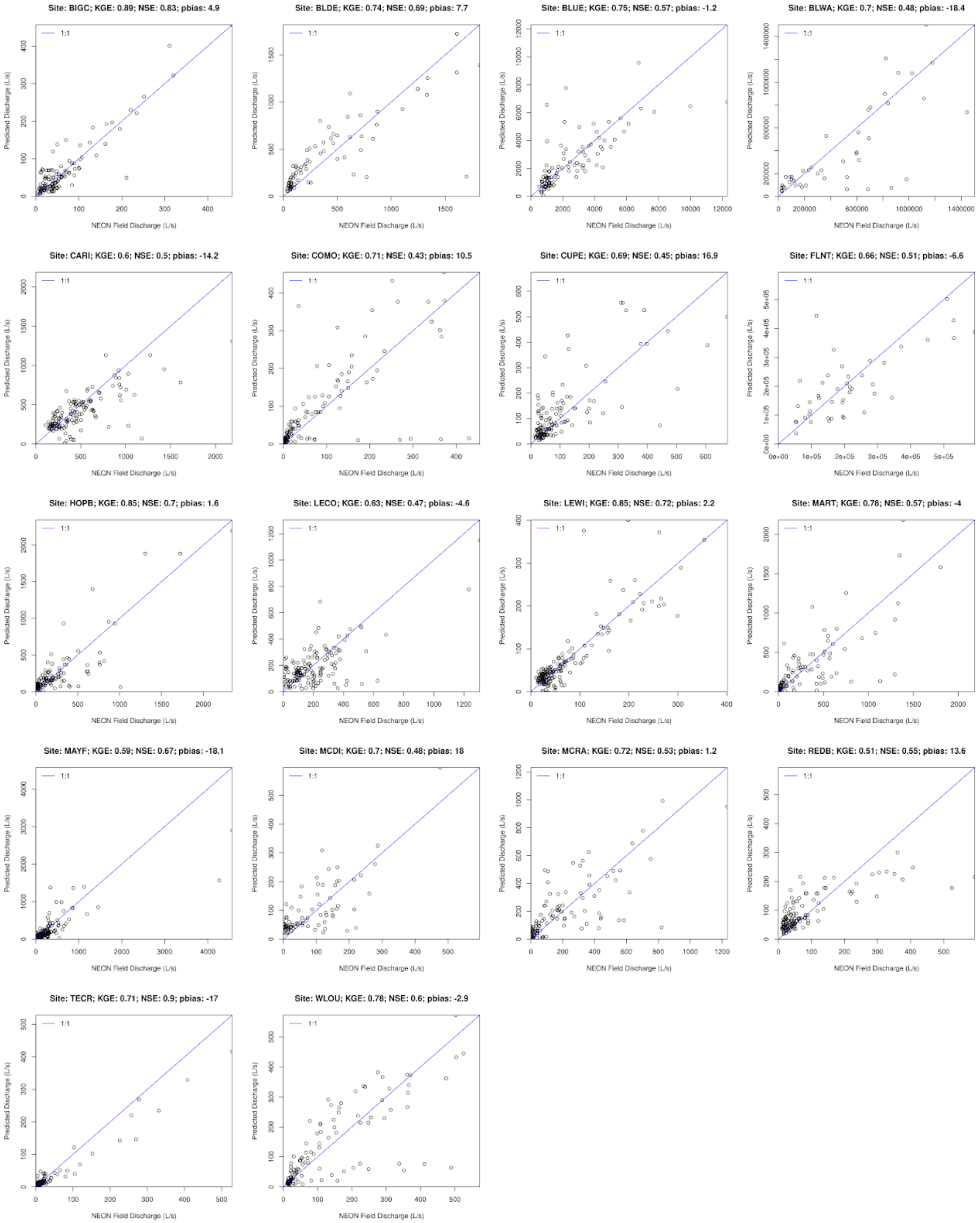


Figure S6. Observed (field) discharge vs. ensembled LSTM predictions.

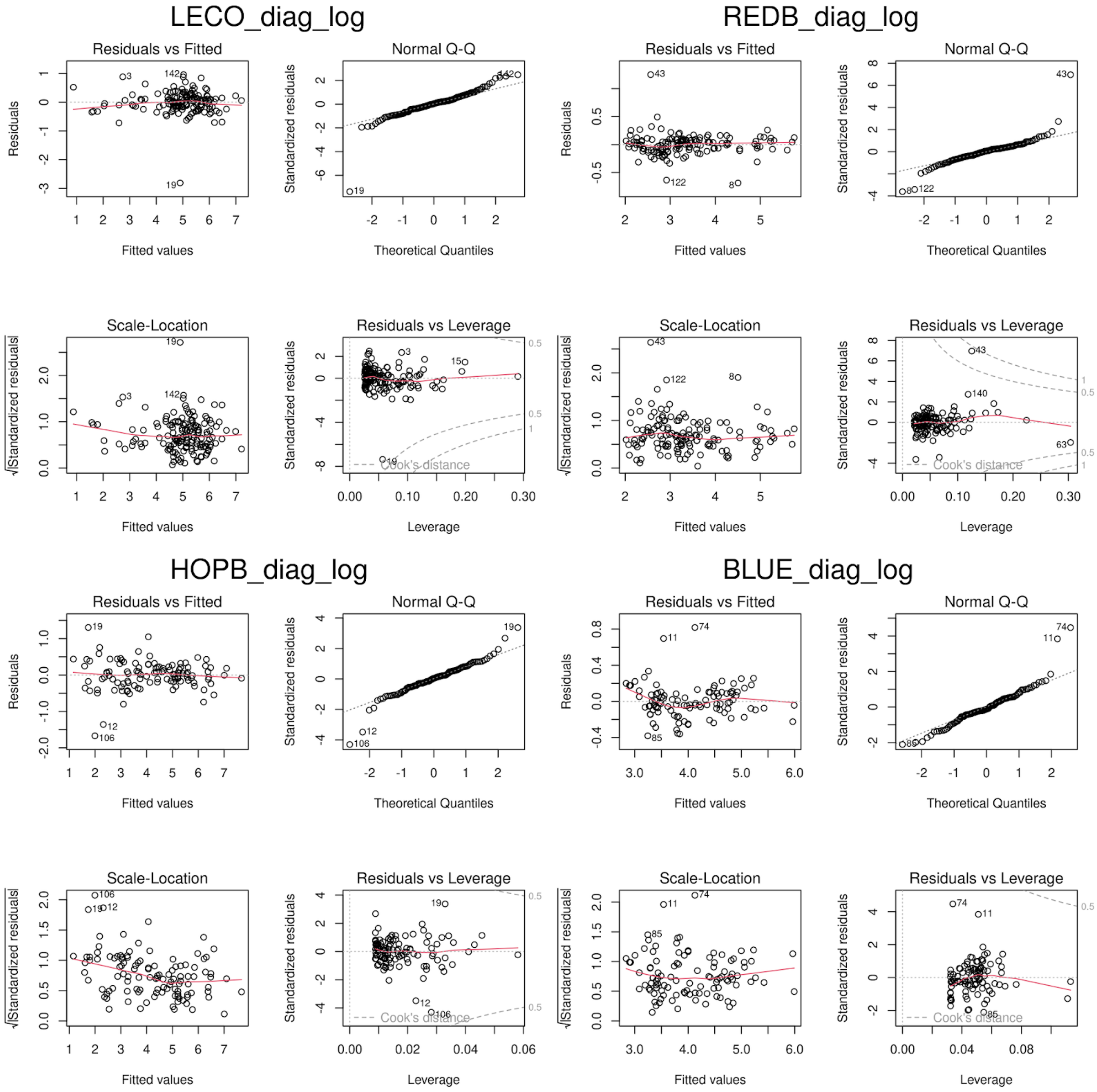


Figure S7. Diagnostic plots for the four sites modeled by OLS regression on specific discharge.

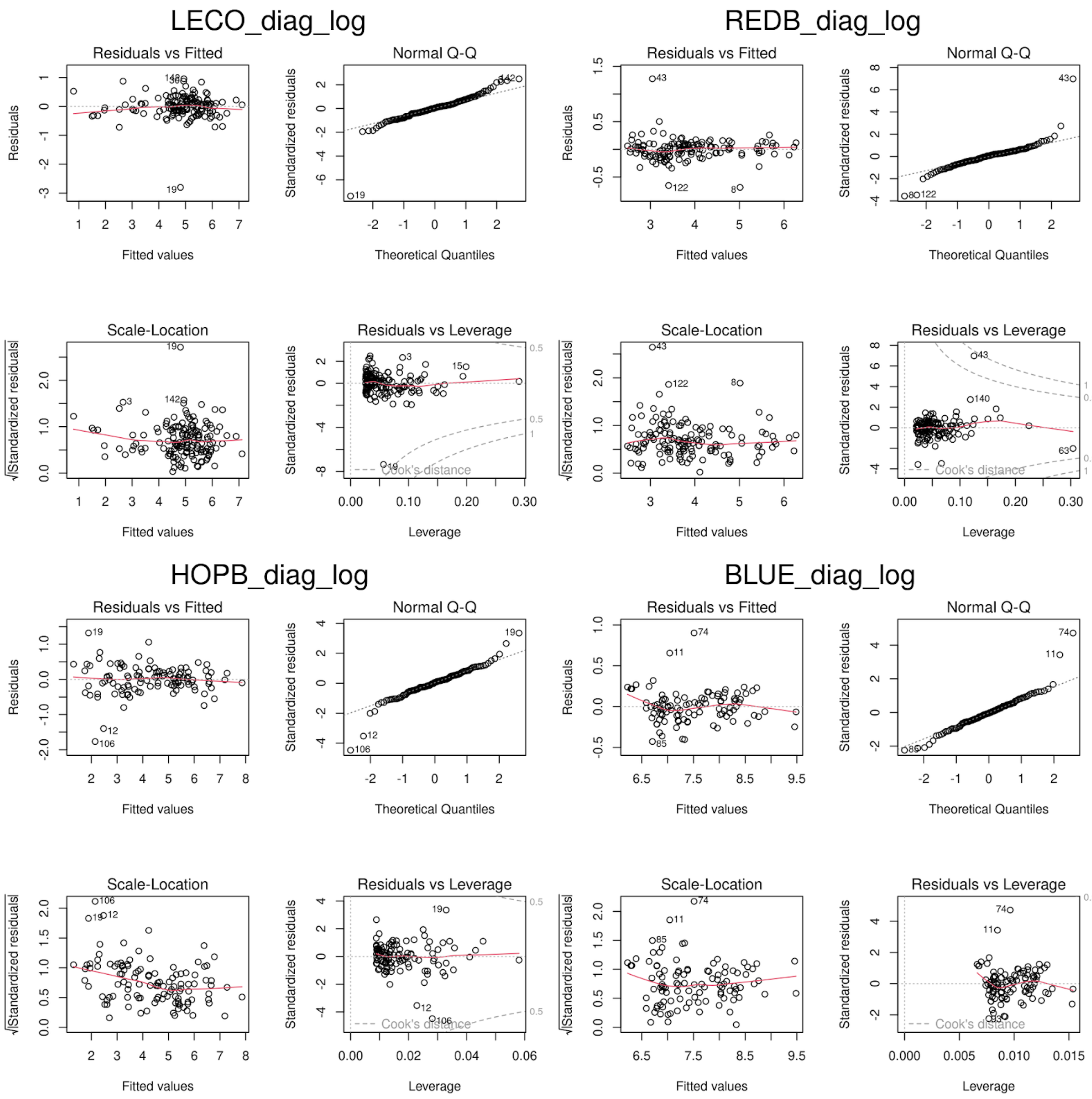


Figure S8. Diagnostic plots for the four sites modeled by OLS regression on absolute discharge.

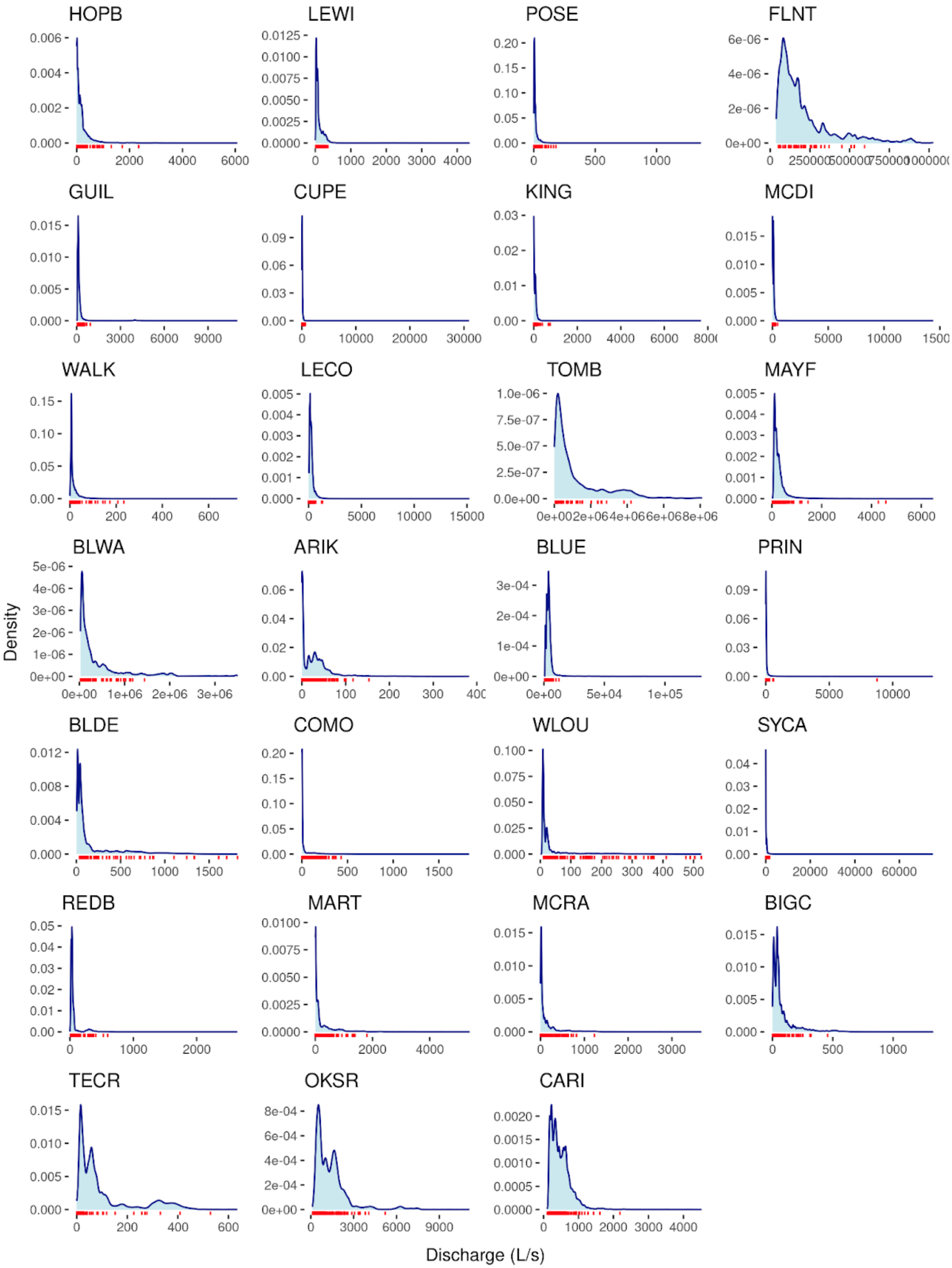

HOPB
0.006
0.004
0.002
0.000
0
2000
4000
6000
LEWI
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
1000
2000
3000
4000
POSE
0.20
0.15
0.10
0.05
0.00
0
500
1000
FLNT
6e-06
4e-06
2e-06
0e+00
0
250000500000750000100000(
GUIL
0.015
0.010
0.005
0.000
0
3000
6000
9000
CUPE
0.09
0.06
0.03
0.00
0
10000
20000
30000
KING
0.03
0.02
0.01
0.00
0
2000
4000
6000
800(
MCDI
0.015
0.010
0.005
0.000
0
5000
10000
1500
WALK
0.15
0.10
0.05
0.00
0
200
400
600
LECO
0.005
0.004
0.003
0.002
0.001
0.000
0
5000
10000
15000
TOMB
1.0e-06
7.5e-07
5.0e-07
2.5e-07
0.0e+00
0e+002e+064e+066e+068e+06
MAYF
0.005
0.004
0.003
0.002
0.001
0.000
0
2000
4000
6000
BLWA
5e-06
4e-06
3e-06
2e-06
1e-06
0e+00
0e+00
1e+06
2e+06
3e+06
ARIK
0.06
0.04
0.02
0.00
0
100
200
300
40(
BLUE
3e-04
2e-04
1e-04
0e+00
0e+00
5e+04
1e+05
PRIN
0.09
0.06
0.03
0.00
0
5000
10000
BLDE
0.012
0.008
0.004
0.000
0
500
1000
1500
COMO
0.20
0.15
0.10
0.05
0.00
0
500
1000
1500
WLOU
0.100
0.075
0.050
0.025
0.000
0
100
200
300
400
500
SYCA
0.04
0.03
0.02
0.01
0.00
0
20000
40000
60000
REDB
0.05
0.04
0.03
0.02
0.01
0.00
0
1000
2000
MART
0.0100
0.0075
0.0050
0.0025
0.0000
0
2000
4000
MCRA
0.015
0.010
0.005
0.000
0
1000
2000
3000
BIGC
0.015
0.010
0.005
0.000
0
500
1000
TECR
0.015
0.010
0.005
0.000
0
200
400
600
OKSR
8e-04
6e-04
4e-04
2e-04
0e+00
0
3000
6000
9000
CARI
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
4000
Density
Discharge (L/s)

Figure S9. Density of NEON-estimated discharge (blue polygon) relative to field-measured discharge observations (red marks).

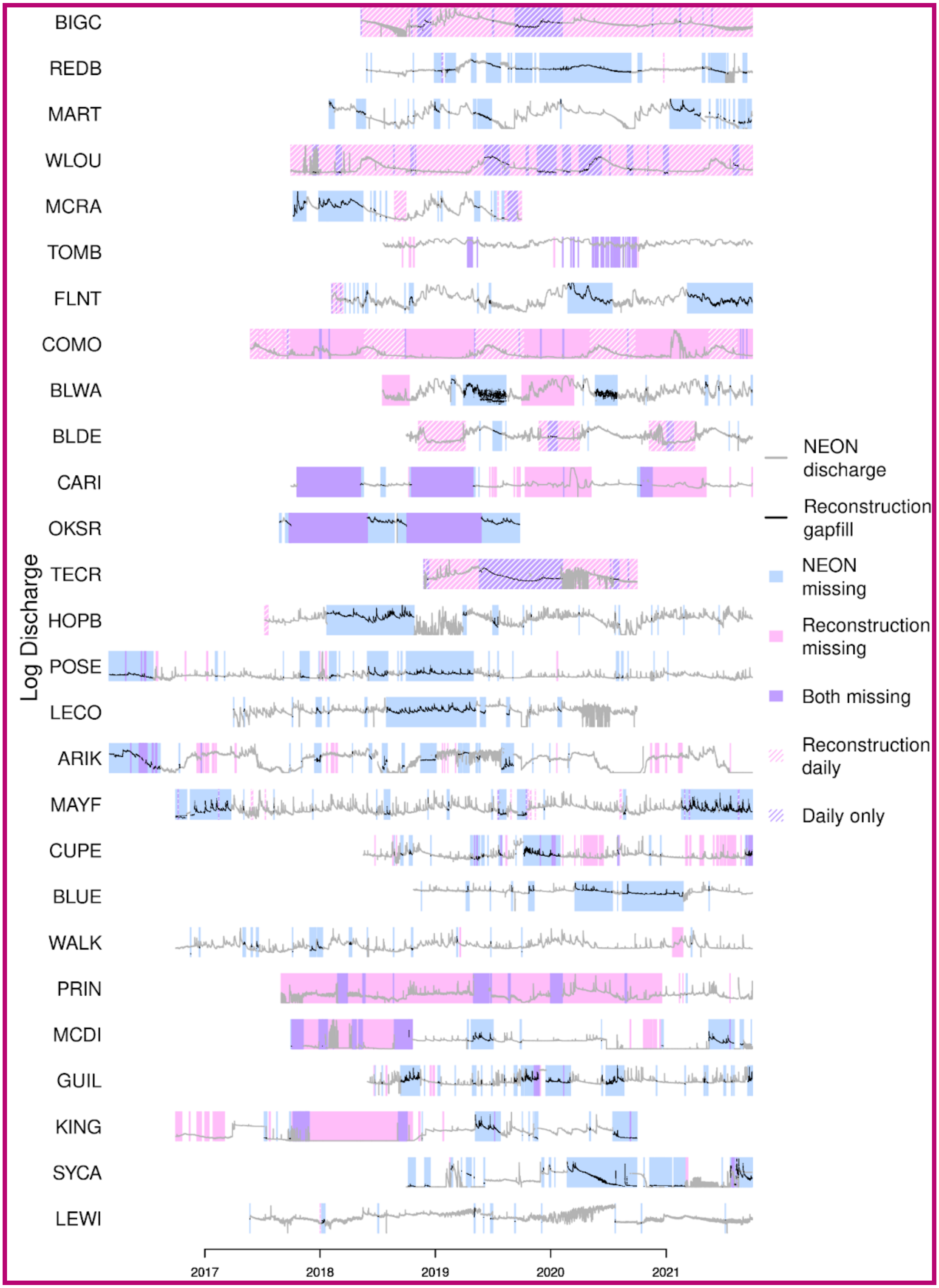

BIGC
REDB
MART
WLOU
MCRA
TOMB
FLNT
COMO
BLWA
BLDE
CARI
OKSR
TECR
HOPB
POSE
LECO
ARIK
MAYF
CUPE
BLUE
WALK
PRIN
MCDI
GUIL
KING
SYCA
LEWI
Log Discharge
2017
2018
2019
2020
2021
NEON discharge
Reconstruction gapfill
NEON missing
Reconstruction missing
Both missing
Reconstruction daily
Daily only

Figure S10. Durations of missing values (gaps) in NEON’s 2023 release of continuous discharge time series, illustrating gaps filled or informed by estimates from this analysis. All officially published values are shown, including those with quality control flags. Sites are ordered as in Figure 2. Gaps smaller than six hours are not indicated.