Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data

Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data

Farshid Rahmani[1], Kathryn Lawson[1], Wenyu Ouyang[2], Alison Appling[3,*], Samantha Oliver[4], Chaopeng Shen[1,*]

[1]Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA
[2]School of Hydraulic Engineering, Dalian University of Technology, Dalian, China
[3]U.S. Geological Survey, Reston, VA, USA
[4]U.S. Geological Survey, Upper Midwest Water Science Center, Middleton, WI, USA

Abstract:
Stream water temperature (T_s) is a variable of critical importance for aquatic ecosystem health. T_s is strongly affected by groundwater-surface water interactions which can be learned from streamflow records, but previously such information was challenging to effectively absorb with process-based models due to parameter equifinality. Based on the long short-term memory (LSTM) deep learning architecture, we developed a basin-centric lumped daily mean T_s model, which was trained over 118 data-rich basins with no major dams in the conterminous United States, and showed strong results. At a national scale, we obtained a median root-mean-square error (RMSE) of 0.69°C, Nash-Sutcliffe model efficiency coefficient (NSE) of 0.985, and correlation of 0.994, which are marked improvements over previous values reported in literature. The addition of streamflow observations as a model input strongly elevated the performance of this model. In the absence of measured streamflow, we showed that a two-stage model can be used where simulated streamflow from a pre-trained LSTM model (Q_{sim}) still benefits the T_s model, even though no new information was brought directly in the inputs of the T_s model; the model indirectly used information learned from streamflow observations provided during the training of Q_{sim}, potentially to improve internal representation of physically meaningful variables. Our results indicate that strong relationships exist between basin-averaged forcing variables, catchment attributes, and T_s that can be simulated by a single model trained by data on the continental scale.

Keywords: Stream Temperature, Machine Learning, streamflow, deep learning, LSTM

1. Introduction:

Stream water temperature (T_s) is a critical, decision-relevant variable that controls numerous physical, chemical, and biological processes and properties, e.g. dissolved oxygen concentrations and nutrient transformation rates, as well as industrial processes such as cooling power plants and treating drinking water (Delpla et al. 2009; Madden, Lewis, and Davis 2013; Kaushal et al. 2010). Thermal regimes of streams directly affect aquatic species (Justice et al. 2017) and in some cases, fish mortality rate increases as T_s passes a certain threshold (Martins et al. 2012; Marcogliese 2001). These are complicated by water uses in industry, such as utilizing stream water for cooling systems, which causes thermal pollution downstream (Raptis, van Vliet,

[1] * Corresponding authors: Chaopeng Shen: cshen@engr.psu.edu, Alison Appling: aappling@usgs.gov

and Pfister 2016). Fulfilling the temperature requirements of the environment, agriculture, industries, and municipalities, and coordinating these uses requires a delicate balance. Accurate T_s models can inform the decision-making process and help lower the risks of exceeding the thermal thresholds.

Myriad basin and in-stream/near-stream processes govern T_s (Poole and Berman 2000). The heat balance of the basin, as modulated by land use types (Nelson and Palmer 2007; Borman and Larson 2003; Moore, Spittlehouse, and Story 2005) is a primary control on T_s. At the basin scale, snowmelt and groundwater baseflow contributions (Kelleher et al. 2012) are also important factors due to their sharp contrast in temperature with air. In streams, T_s is influenced by solar radiation, latent heat flux, air-water heat exchange, riparian vegetation (Theurer, Lines, and Nelson 1985; Garner et al. 2017), channel geomorphology (Hawkins et al. 1997), hyporheic exchange (Evans and Petts 1997), and reservoirs and industrial discharges (Poff et al. 2010). At any point in the channel network, T_s is the spatiotemporal integration of all of the above processes. Process-based models, while offering physical explanations of causes and effects, need to embrace substantial model complexity to represent all or even parts of these complex processes with their heterogeneity and scaling effects (Johnson et al. 2020). The requirements for input data also make scaling up such simulations challenging. Some large-scale process-based models have had root-mean-square error (RMSE) values reported of greater than 2.5 °C (Wanders et al. 2019; van Vliet et al. 2012).

A large body of literature has employed statistical models to simulate T_s, with some good summaries given by Benyahya et al. (2007) and Gallice et al. (2015). Typically, T_s was regressed to air temperature (T_a), but more recent studies regressed the parameters in T_s~T_a relationships using catchment attributes. Among these studies and most relevant to our work, Segura et al. (2015) predicted the slope and intercept of the 7-day average T_s~T_a relationship based on catchment characteristics such as watershed area and baseflow index. A Nash-Sutcliffe coefficient of 0.78 was obtained for reference sites for the 7-day average T_s, and strong hysteresis was noted in the stream-air temperature relationship (Segura et al. 2015). Stewart et al. (2015), integrated an artificial neural network with a soil water balance model and obtained an RMSE of around 1.5°C and R^2 of 0.76 for 371 sites across Wisconsin. Very recently, Johnson et al. (2020) used sine-wave linear regression and reported RMSE values of 1.41°C for an extensive regional dataset and 1.85°C for the national-scale U.S. Geological Survey (USGS) dataset. They highlighted the importance of spatial scales and heterogeneity. Graf et al. (2019) used wavelet transformations of daily average air temperature as inputs to an artificial neural network to predict water temperature at eight different sites in Poland and obtained RMSE values ranging from 0.98° to 1.43° in the test period. Additionally, there are several studies that used recent water temperature data as an input to predict stream temperature, which can significantly increase model performance as it can be interpreted as a form of data assimilation (Feng et al., 2020). Sohrabi et al. (2017) obtained RMSE of ~1.25°C when they used the previous day's temperature and streamflow as drivers. On one temperature gauge, Stajkowski et al. (2020) obtained RMSE of 0.76°C using a variant of LSTM with previous hour's stream temperature in the inputs. However, here we do not use previous day measurements in inputs because the purpose of our model is long-term projections.

Streamflow conditions are often not utilized by statistical models of T_s, partially because relationships are not clear; none of the abovementioned statistical models used streamflow. However, we do know that streamflow exerts substantial control on T_s. Rivers dominated by baseflow are typically fairly stable and cool in summer, but have relatively low thermal capacity and are rapidly heated by strong solar radiation. Peak flows are typically dominated by surface runoff, the temperature of which is strongly influenced by the fast-changing air temperature (Edinger, Duttweiler, and Geyer 1968). A sensitivity analysis on large river basins found that on average, a decrease in river flow of 50% as compared to a reference condition lowered the minimum annual river discharge temperature (in winter) by -0.4°C or raised the maximum temperature (in summer) by +1.2°C (van Vliet et al. 2012). In another study on one specific river, a linear regression model based on monthly air temperature and streamflow data revealed that air temperature rise and flow reduction were responsible for 60% and 40% of June to August temperature increases, respectively (Moatar and Gailhard 2006).

Recently, deep learning (DL) models, including those based on the long short-term memory (LSTM) algorithm, have shown promise in predicting hydrologic variables such as soil moisture and streamflow by achieving superior results with low computational and human effort (Feng, Fang, and Shen 2020; Shen 2018; Shen et al. 2018; Fang and Shen 2020). LSTM can learn long-term dependencies and gave high performance in snow-dominated regions for streamflow prediction (Feng, Fang and Shen 2020). The memory mechanisms of LSTM may be able to mimic heat units, similar to heat accumulation and release processes. Thus, it is natural to think that LSTM may also be suitable for T_s modeling. However, given the complicated and scale-dependent processes influencing T_s, it is highly uncertain if there is even a stable relationship between basin-average forcing inputs and T_s across different spatial scales, and if so, whether such a relationship can be captured by LSTM given limited observational data.

One advantage of a DL model as compared to process-based ones is that it can incorporate auxiliary information without requiring explicit understanding of relationships. In our case, not only does streamflow directly influence T_s fluctuations, it also reveals multi-faceted hydrologic dynamics in a basin regarding factors such as baseflow contributions and residence times of surface runoff, which could aid T_s modeling. Therefore, we expect adding streamflow information to improve model performance. In a process-based modeling framework, streamflow data can be used to calibrate the hydrologic components. Unfortunately due to the issue of model equifinality (Beven 2006; Beven and Freer 2001), calibration may or may not improve model internal dynamics, depending on model parameterization, structure, and data information content (Huang and Liang 2006). In utilizing the DL framework, we hypothesize that models may be able to automatically extract information from hydrographs to inform T_s, which, to our knowledge, no study has examined in the context of deep learning models.

Even if streamflow data are indeed useful, real-world use can be hampered by lack of available streamflow data. Beyond existing stations, collecting new streamflow data is more expensive than collecting new T_s data. However, given that highly accurate, LSTM-based streamflow models have been reported (Feng, Fang, and Shen 2020), we wondered if a well-trained LSTM streamflow model could serve as a surrogate for actual measurements. Deep networks are known to maximally use available information, but it is not certain whether such a streamflow model presents any benefit; the model uses identical forcing data to those already used by the temperature model and do not explicitly bring in new information.

In this work, we attempted to answer two main research questions and improve our understanding of the stream heat balance: (1) Are there reliable relationships between basin-average meteorological forcing and attributes and T_s that could be learned by deep networks to predict T_s with high accuracy? (2) Can observed or simulated streamflow be used to improve temperature predictions, especially when the simulated streamflow is predicted using the same information as the T_s model?

2. Methods:

We simulated T_s from a basin perspective, that is, as a function of basin-average climate forcings and attributes. This setup greatly simplifies the model representation compared to spatially explicit models and is supported by widely-available data, but ignores some local channel characteristics. We examined the effect of including daily streamflow in the inputs to assess its information content for T_s.

2.1. Datasets

Basin characteristics came from the Geospatial Attributes of Gages for Evaluating Streamflow dataset, version II (GAGES-II), which represent geological aspects, land cover, reservoir information, and air temperature data (Falcone 2011). Historical data for daily mean T_s was downloaded from the USGS's National Water Information System (USGS NWIS) website for all 9322 basins in GAGES-II (USGS 2016). We obtained daily meteorological forcing data (precipitation, maximum and minimum air temperature, vapor pressure, solar radiation) by interpolating a gridded meteorological dataset (Daymet) (Thornton et al. 2016) from Google Earth Engine (GEE) (Gorelick et al. 2017) for each basin for the period of 2004 to 2016. Daily streamflow observations were downloaded from USGS NWIS for the same period. Forcing and attribute inputs are summarized in Table S1 (Supplementary Information).

Many of the GAGES-II basins had T_s observations recorded for only some days of the year, with unobserved days more common during the winter (there are nonetheless many sites with winter data and our model predicts temperature for all days in a year). For this paper, we selected temperature gauges with more than 60% of daily observations available between 2010/10/01 and 2014/09/30, in basins where there were no major dams (more than 50 ft in height, or having more than 5000 acre feet storage according to the definition in GAGES-II), resulting in a dataset of 118 basins ranging in size from 2 to 14,000 km^2. We simulated T_s at the pour point of basins where the USGS streamgages were located. Limiting this analysis to sites with >60% data coverage allowed us to focus on the capabilities of LSTM for T_s modeling under relatively ideal conditions. Future research on the effect of reservoir presence and data availability could further improve stream temperature predictions.

2.2. LSTM-based models for predicting T_s

We used the long short-term memory (LSTM) algorithm which has received increasing attention in hydrologic literature. This method is designed to learn and keep information for long periods using units called memory cells and gates. Cells store the information, and gates decide which information comes in and out of the cells. Because the basic LSTM architecture has been described extensively elsewhere, we refer readers to those papers for a more detailed discussion of the equations and structure of LSTM (Hochreiter and Schmidhuber 1997; Fang et al. 2017; Fang, Pan, and Shen 2019). A sketch and equations of the model are provided in Figure S1.

We standardized values for all inputs and target values. As a preprocessing step, streamflow was first divided by basin area and mean annual precipitation to obtain a dimensionless streamflow, which was then transformed to a new, more Gaussian distribution (Feng, Fang, and Shen 2020):

$$v^* = \log_{10}(\sqrt{v} + 0.1) \quad (1)$$

where v^* and v are the variables after and before transformation, respectively. Next, the transformed streamflow data along with all other meteorological forcing data, basin characteristics, and T_s observations were standardized by the following formula (Feng, Fang, and Shen 2020):

$$x_{i,new} = \frac{(x_i - \bar{x})}{\sigma} \quad (2)$$

in which $x_{i,new}$ is the standardized value, x_i is the raw value, $\bar{x}$ is the mean value, and σ is the standard deviation for each variable. Standardization better conditions the model for gradient descent and forces the model to pay roughly equal attention to both large wet basins and small dry basins (Feng, Fang, and Shen 2019). All results in this study are shown after destandardization, or reversal of all standardization procedures, for the outputs.

Hyperparameters were chosen by running multiple tests to determine the hyperparameters as listed in Table S2. The metric the model training aimed to minimize (loss) was RMSE, and we also reported an unbiased RMSE (ubRMSE), which is the RMSE minus the mean bias. We also report bias (mean of error) and Nash-Sutcliffe efficiency coefficient (NSE, equation in the supporting information) (Nash and Sutcliffe 1970) for the test periods for comparison with other studies. Further, because a model simply copying air temperature may give acceptable metrics, we added Nash-Sutcliffe coefficient (NSE_{res}) calculated based on the difference between daily mean water and daily mean air temperatures (residual temperature): $T_{res}=T_s-T_a$. To provide a baseline for comparison, we also provide a locally-fitted autoregressive model with exogenous variables (ARX_2). The ARX_2 inputs contain current and delayed atmospheric forcings (X) and ARX_2-simulated stream temperature in the last two days: $T_s^{t,*} = \sum_{i=1}^{2} a_i T_s^{t-i,*} + \sum_{i=0}^{2}\sum_{j=1}^{p} b_{i,j} X_j^{t-i} + c$, where a, b and c were fitted coefficients, $T_s^{t,*}$ is the stream temperature simulated in this model at time step t and p is the number of forcings. All temperature models were trained on data from 2010/10/01-2014/09/30 and tested on 2014/10/01-2016/09/30.

2.3. Streamflow observations or simulations as model inputs

Across USGS gages, streamflow is a more widely-available measurement than temperature, meaning that inclusion could bring additional information. To test this, the following models were trained:

$$T_{s,obsQ} = LSTM_{obsQ}(F, A_T, Q_{obs}) \quad (3)$$

$$T_{s,noQ} = LSTM_{noQ}(F, A_T) \quad (4)$$

$$T_{s,simQ} = LSTM_{simQ}(F, A_T, Q_{sim}) \quad (5)$$

where F is the forcing data time series, A_T represents static and single-valued attributes of the basin for temperature modeling, Q_{obs} is the observed time series of daily mean streamflow, and Q_{sim} is simulated streamflow (described below). $LSTM_{obsQ}$, $LSTM_{noQ}$, and $LSTM_{simQ}$ are LSTM-based models incorporating observed streamflow, no streamflow information, and simulated streamflow, respectively. For Q_{sim}, streamflow was simulated using a LSTM-based streamflow model shown to have very good performance (Feng, Fang, and Shen 2020).

$$Q_{sim} = LSTM_Q(F, A_Q) \quad (6)$$

A_Q represents static attributes of the basins for streamflow modeling (Table S1 in Supporting Information). Meteorological forcing data used for the simulations were the same as for the temperature prediction models. Q_{sim} was trained using observations from 2397 basins, and a longer training period (from 2004/10/01 to 2014/09/30) was used than for the temperature models in this study.

3. Results and Discussion

3.1. Overall results

All LSTM-based models delivered exceptionally strong performance in the test period (Figure 1). The median test-period RMSE for the model incorporating streamflow observations ($LSTM_{obsQ}$) was 0.69°C. The RMSE for the model incorporating simulated streamflow ($LSTM_{simQ}$) was 0.81°C, which was still lower than that for the model lacking any streamflow information ($LSTM_{noQ}$), for which the RMSE was 0.86°C. The corresponding median NSE values were 0.986, 0.983, and 0.979 respectively, and all of the correlation values were above 0.992, indicating that temporal fluctuations were extremely well captured. These metrics are markedly better than those reported in the literature at this scale, which demonstrates that LSTM is particularly well-suited for T_s modeling at basin outlets.

LSTM performed much better than ARX_2, which had a median RSEM of 1.41 °C. Moreover, when we evaluated T_{res}, the locally-fitted ARX_2 model's median Nash-Sutcliffe Efficiency (NSE_{res}) worsened substantially to 0.772, indicating that a substantialf portion (although not all) of ARX_2's predictive power came from air temperature and some memory (linear regression is worse than ARX_2 and not shown here). In comparison, the LSTM-based models were much less affected; the median NSE_{res} values for the conterminous United States (CONUS) values were above 0.950 and 0.924 for $LSTM_{obsQ}$ and $LSTM_{noQ}$, respectively. LSTM models captured most fluctuations unaccounted for by seasonality and they were able to capture more complicated memory effects than the simple linear autocorrelation described in ARX_2. These temporal fluctuations could have been induced by heat storages in the basin (vegetation, snow, soil, groundwater, riparian zone, urban areas) causing delayed responses to atmospheric forcings.

$LSTM_{obsQ}$ generally performed better in the eastern CONUS than the western half, and better in the northern half than the southern half (Figure 2a-b). Most of the eastern basins had NSE values above 0.975 and RMSE values below 0.9°C. Northern basins had slightly higher NSEs, presumably because in colder basins, the minimum winter liquid T_s is confined to above 0°C, and therefore easier to predict. Existing statistical models often have difficulty for northern basins where air temperature and water temperature are decoupled. LSTM has a long memory to keep track of seasonal snow states and can learn threshold-like functions. Hence LSTM is quite useful where existing models often have deficiencies. Sites with large RMSE values were scattered across the geographic extent, but there were no clear explainable patterns. Regardless, the NSE values for most of these "difficult" basins were still quite high, with only two stations out of 118 having NSEs <0.9.

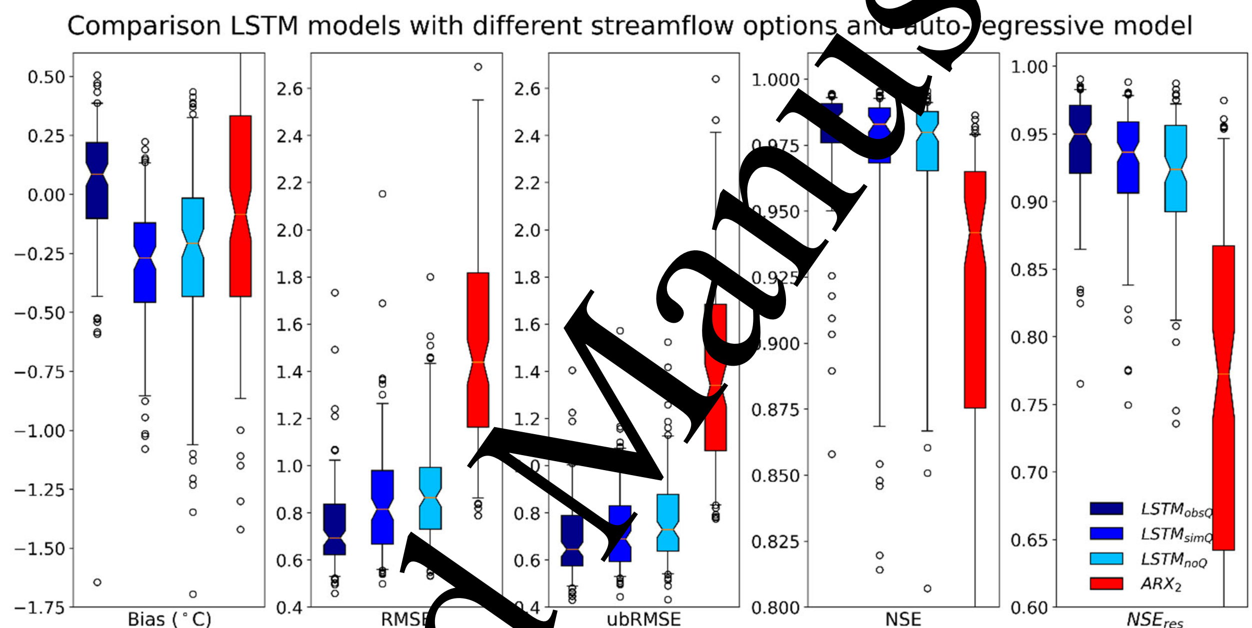

***Figure 1**. CONUS-scale aggregated metrics of stream temperature models for the test period. $LSTM_{obsQ}$ incorporated observed streamflow, $LSTM_{noQ}$ had no input streamflow information, while $LSTM_{simQ}$ incorporated simulated streamflow (Q_{sim}). ARX_2 is the locally-fitted auto-regressive model with extra inputs. The lower whisker, lower box edge, center bar, upper box edge and upper whisker represent 5%, 25%, 50%, 75% and 95% of data, respectively.*

3.2. Impacts of observed and simulated streamflow as inputs

Providing streamflow as an input to the T_s model generally improved model accuracy, but the effects were pronounced for the poorly-simulated sites. The models incorporating either observed ($LSTM_{obsQ}$) or simulated ($LSTM_{simQ}$) streamflow improved median bias (reducing the absolute median bias by 0.120°C and 0.062°C), RMSE (by 0.170°C and 0.049°C) as compared to the model lacking streamflow ($LSTM_{noQ}$) (Figure 1). Including streamflow information helped to both reduce bias and greatly improve representation of temporal fluctuations, especially for the worse-performing sites. Without the streamflow data, 10 sites had NSE values below 0.9. Additionally, $LSTM_{noQ}$ had a median bias of around -0.25°C, while the median bias of $LSTM_{obsQ}$

was much closer to 0°C. The inclusion of observed streamflow also greatly reduced overall error range, providing the largest improvements in model performance at the most troublesome sites.

The model incorporating simulated streamflow ($LSTM_{simQ}$) generally performed between $LSTM_{obsQ}$ and $LSTM_{noQ}$. Similar to $LSTM_{obsQ}$, $LSTM_{simQ}$ helped to noticeably improve the accuracy and reduced the spread of bias (decreasing error range, as shown by compressed whiskers and outliers compared to $LSTM_{noQ}$), but did not help as much to improve the median bias. Understandably, simulated streamflow had more errors compared to actual observations, and input attributes (A_Q) do not fully characterize a basin. While Q_{sim} offered state-of-the-art performance, it still encountered more errors estimating peaks (mainly due to rainfall inputs) and baseflows, especially in the western CONUS (possibly due to inadequate geological information) (Feng, Fang, and Shen 2020).

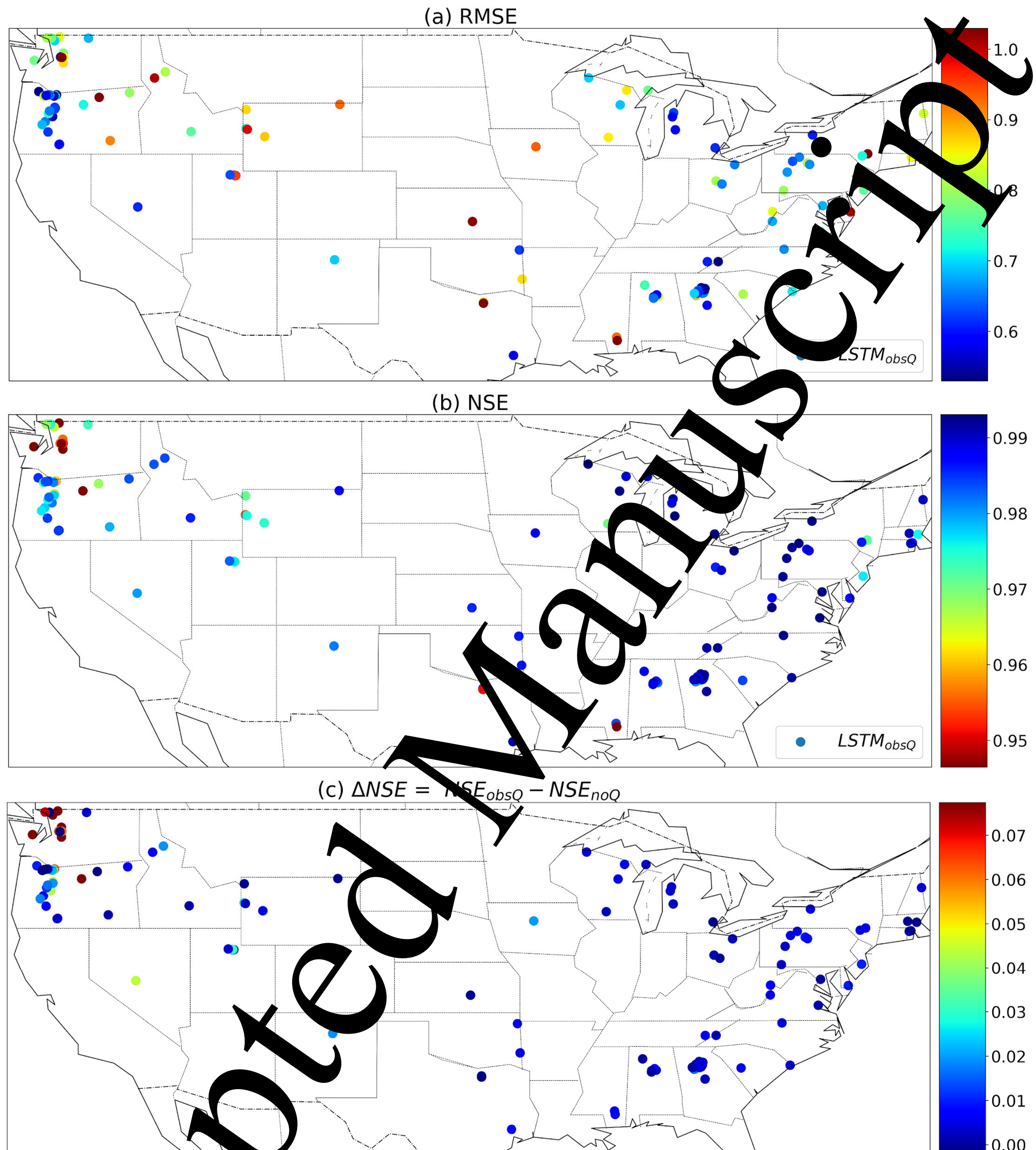

Figure 2. *Maps of the [illegible] USGS streamgage locations used in this study, where dot color represents (a) RMSE and (b) NSE values for $LSTM_{obsQ}$, and (c) ΔNSE values between $LSTM_{obsQ}$ and $LSTM_{noQ}$.*

Negative biases with $LSTM_{noQ}$ were attributable to underestimating T_s peaks in both winter and summer in some sites (e.g., Figure 3a) and a more consistent bias at other sites (e.g., Figure 3b). T_s peaks are often associated with streamflow peaks (possibly caused by warm rain) in the winter but after-storm recession limbs in the summer. For the Black River in Ohio (Figure 3a), T_s peaks were coincidental with recession periods between storms in summer 2015 (annotated points A & B). Simulated T_s by $LSTM_{noQ}$ did not rise as high as the observed T_s, possibly because $LSTM_{noQ}$ had an internal representation of baseflow that is overestimated here, while $LSTM_{obsQ}$ captured the peaks well. For the South Fork Sultan River in Washington (Figure 3b), there was a more prominent year-round bias for temperature predictions in 2015, concurrent with an overestimation of baseflow in Q_{sim}. This underestimation is potentially due to multi-year accumulation and melt of snowpacks. This basin typically has a long snow season, sometimes the whole year, but the 2015 summer saw all snow melted by June (verified via Google Earth).

Several reasons could explain why observed streamflow helps the model, but to think them through, we first need to assume that the LSTM model has internal representations of physically-relevant quantities such as water depth, snowmelt, water temperature, net heat flux, and baseflow temperature. Other studies have shown that LSTM has learned to use cell memory states to represent intermediate hydrologic variables that were not matched to observations, e.g., snow cover (Jiang, Zheng, and Solomatine 2020). With this assumption, it is possible that observed or simulated streamflow corrects the internal "water depth" variable to estimate the effect of net heat flux. From the energy balance equation, stream temperature changes are estimated by dividing the net heat flux over the flow depth. If streamflow is overestimated during summer baseflow periods, the positive heat flux is vertically diluted too much (and thus the temperature rise is underestimated). Secondly, derived from Figure 3b, the model may not be able to accurately keep track of long-term snow accumulation/melt so $LSTM_{noQ}$ misjudges the amount of cool snowmelt water. However, $LSTM_{obsQ}$ was informed by observed flow and corrected the error. In fact, the basins with the lowest NSEs were concentrated in the Rocky Mountains with long snow seasons (Figure 2c). Thirdly, LSTM-based T_s models may have learned other holistic hydrologic information from the streamflow time series. For example, they may have learned to perform baseflow separation internally, if such a feature was helpful for T_s prediction. Streamflow data may provide more clues to reduce uncertainties - for example, cool summer temperatures could be due to high baseflow or abundant riparian shade, and streamflow data may make it easier to distinguish between these causes.

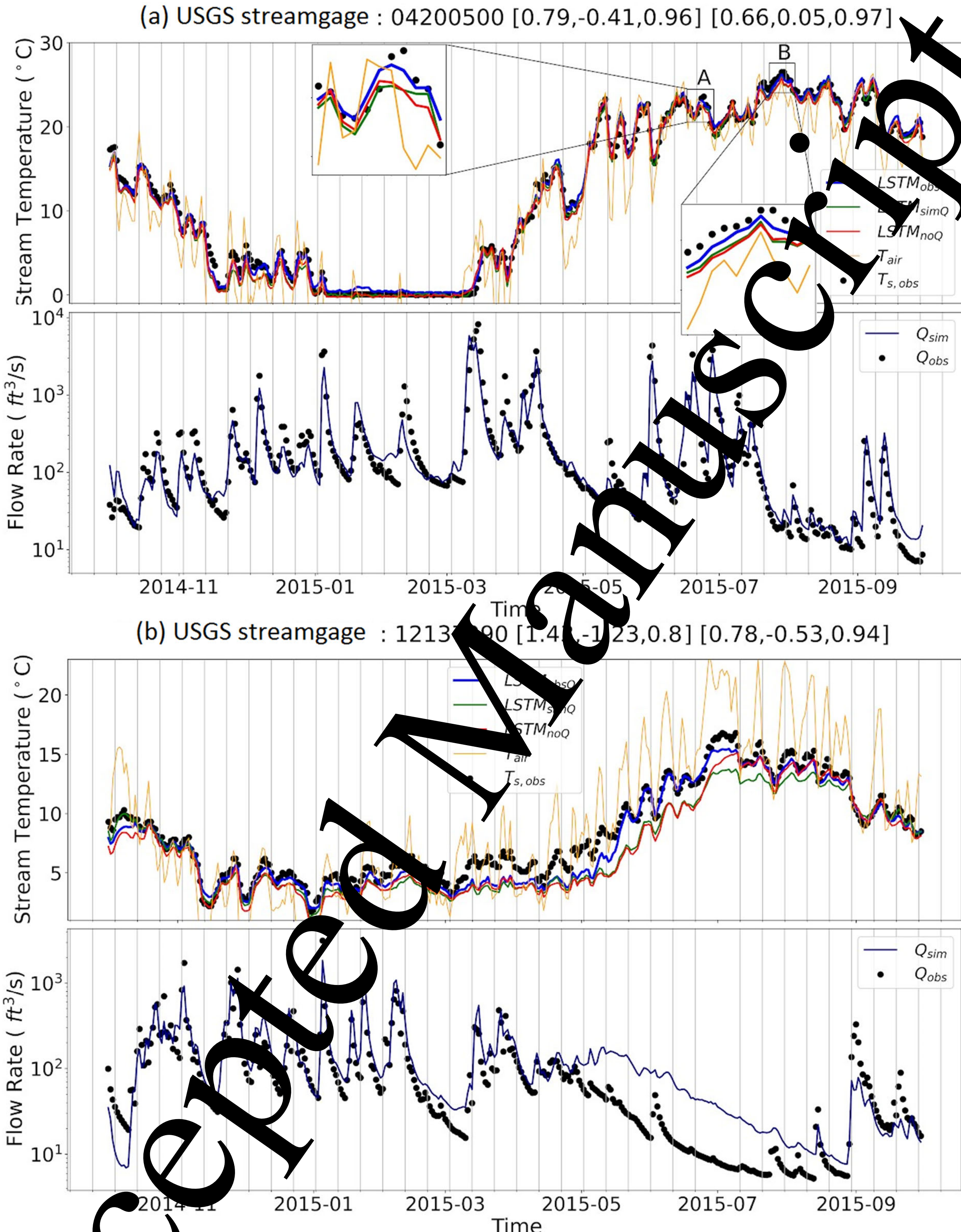

Figure [illegible]. Time series plots of observed and simulated T_s for a good performing simulation, along with the observed and simulation hydrographs shown in log scale. (a) Black River at Elyria, Ohio; (b) South Fork Sultan River, Washington. $T_{s,\,obs}$ and Q_{obs} represent observed water temperature and streamflow, respectively. The two brackets contain values for [RMSE, Bias, NSE_{res}] for $LSTM_{noQ}$ and $LSTM_{obsQ}$, respectively.

3.3. Further discussion

LSTM, with its hidden layers to store system states (100 hidden units) with different rates, is extremely well-suited to model systems with memory and hysteresis. The warming and cooling of water storage compartments (soil water, groundwater, riparian zones, etc.) are caused by different mechanisms with different rates, durations, and lags relative to drivers (accumulation, flush by storms, etc.). Such multirate exchanges, along with diffusive exchanges with soil, easily lead to hysteresis in the system (Briggs et al. 2014). We suspect that the internal states and gates of LSTM-based models mimic the effects of buffers and delays by these heat (and water) storage compartments and can be sufficiently trained by 4 years of data as was done in this study.

Streamflow data may have carried multifaceted, temperature-relevant information about stream depth, basin hydrologic properties, and the relative influence of flow versus other heat-moderating processes. Even simulated streamflow provided valuable new information to the temperature model -- despite the fact that Q_{sim} ingested identical meteorological forcing data as the T_s models. We posit that the pre-trained Q_{sim} model may have derived some of this new information from the additional catchment attributes in A_Q relative to A_T (Table S1 in Supporting Information) but that the majority of the new information came from the 10 years of streamflow observations across the 2397 stations on which Q_{sim} was trained. Q_{sim} was thus able to learn and transfer a wealth of nuanced information about each basin's hydrologic properties and responses to meteorological drivers, which in turn likely improved the implicit representation of those attributes in the T_s models.

As the first LSTM application for stream temperature, this application is focused on temporal prediction for basins with a good record of historical data, and, as such, may not generalize to ungauged basins well. It is well-known that spatial extrapolation of stream temperature models can be quite risky (Gallice et al. 2015), which could be investigated in the future. Also warranting investigation is the representation of the spatial heterogeneity at smaller scales, e.g., using a multiscale graph network or calibrating parameters of a spatially-distributed process-based model.

4. Conclusions

This is the first time a basin-centric lumped T_s model has been shown to be so effective. The results clearly indicate that robust (but complex) mapping relationships exist between basin-average attributes, climate forcings, and T_s, which can be reliably learned by a uniform, continental-scale model using a few years' worth of daily T_s observations. All models presented exceptional evaluation metrics that outperformed state-of-the-art models reported in the literature by a substantial margin. Additionally, this performance was achieved without the need for detailed representations of the subsurface or the channel network - a convenience that promises high-quality forecasts of future T_s given available climate forcings.

Our use of a basin-centric lumped model to predict T_s allowed for great simplification that potentially enabled LSTM to learn the connection between different factors influencing T_s, alongside the more obvious benefits of simplifying model assembly and training. The disadvantage of this basin-centric formulation is that it assumes each basin is homogeneous in forcings and attributes. The homogeneity assumption fundamentally limits the size of the basin that can be simulated: when predictions are needed for larger, mainstem rivers, we will need reach-centric models. Therefore, while the current model is highly capable and useful, we do not perceive the present form of the model as being complete in functionality. ●

Our results show that observed streamflow information helped to improve the T_s model, perhaps because the observations (i) allowed the model to better resolve groundwater and snowmelt contributions, and (ii) provided a more accurate water volume used to estimate the effect of net heat fluxes, especially during recession periods when T_s rapidly changed. The benefits were most substantial in basins with multiyear snow accumulations. If observed streamflow does not exist, a well-trained continental-scale streamflow model was able to indirectly bring in data from a larger training dataset, which alleviated more than half of the degradation in median NSE that would have otherwise resulted from the lack of streamflow observations.

Acknowledgments

FR was supported by the Pennsylvania Water Resources Research Center graduate internship G19AC00425, and funding for the fellowship and AA and CO was provided by the Integrated Water Prediction Program at the U.S. Geological Survey. CS was supported by National Science Foundation Award OAC #1940190. Data sources have been cited in the paper, and all model inputs, outputs, and code are archived in a data release (Rahmani et al. 2020). The LSTM code for modeling streamflow is available at https://github.com/mhpi/hydroDL. CS and KL have financial interests in HydroSapient, Inc., a company which could potentially benefit from the results of this research. This interest has been reviewed by the University in accordance with its Individual Conflict of Interest policy, for the purpose of maintaining the objectivity and the integrity of research at The Pennsylvania State University. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Benyahya, Loubna, Daniel Caissie, André St-Hilaire, Taha B.M.J Ouarda, and Bernard Bobée. 2007. "A Review of Statistical Water Temperature Models." *Canadian Water Resources Journal* 32 (3): 179–92. https://doi.org/10.4296/cwrj3203179.

Beven, Keith. 2006. "A Manifesto for the Equifinality Thesis." *Journal of Hydrology* 320 (1–2): 18–36. https://doi.org/10/ccx2ks.

Beven, Keith, and Jim Freer. 2001. "Equifinality, Data Assimilation, and Uncertainty Estimation in Mechanistic Modelling of Complex Environmental Systems Using the GLUE Methodology." *Journal of Hydrology* 249 (1): 11–29. https://doi.org/10/fgmngv.

Borman, M. M., and L. L. Larson. 2003. "A Case Study of River Temperature Response to Agricultural Land Use and Environmental Thermal Patterns." *Journal of Soil and Water Conservation* 58 (1): 8–12.

Briggs, Martin A., Frederick D. Day-Lewis, John B. Ong, Judson W. Harvey, and John W. Lane Jr. 2014. "Dual-Domain Mass-Transfer Parameters from Electrical Hysteresis: Theory and Analytical Approach Applied to Laboratory, Synthetic Streambed, and Groundwater Experiments." *Water Resources Research*. https://doi.org/10.1002/2014WR015880.

Delpla, I., A.-V. Jung, E. Baures, M. Clement, and O. Thomas. 2009. "Impacts of Climate

Change on Surface Water Quality in Relation to Drinking Water Production." *Environment International* 35 (8): 1225–33. https://doi.org/10.1016/j.envint.2009.07.001.

Edinger, John E., David W. Duttweiler, and John C. Geyer. 1968. "The Response of Water Temperatures to Meteorological Conditions." *Water Resources Research* 4 (5): 1137–43. https://doi.org/10.1029/WR004i005p01137.

Evans, E. C., and G. E. Petts. 1997. "Hyporheic Temperature Patterns within Riffles." *Hydrological Sciences Journal* 42 (2): 199–213. https://doi.org/10.1080/02626669709492020.

Falcone, James A. 2011. GAGES-II: Geospatial attributes of gages for evaluating streamflow. US Geological Survey. https://doi.org/10.3133/70046617

Fang, Kuai, Ming Pan, and Chaopeng Shen. 2019. "The Value of SMAP for Long-Term Soil Moisture Estimation with the Help of Deep Learning." *IEEE Transactions on Geoscience and Remote Sensing* 57 (4): 2221–33. https://doi.org/10/gghp8v.

Fang, Kuai, and Chaopeng Shen. 2020. "Near-Real-Time Forecast of Satellite-Based Soil Moisture Using Long Short-Term Memory with an Adaptive Data Integration Kernel." *Journal of Hydrometeorology* 21 (3): 399–413. https://doi.org/10/ggj669.

Fang, Kuai, Chaopeng Shen, Daniel Kifer, and Xiao Yang. 2017. "Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network." *Geophysical Research Letters* 44 (21): 11,030-11,039. https://doi.org/10/gcr7mq.

Feng, Dapeng, Kuai Fang, and Chaopeng Shen. 2020. "Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks with Data Integration at Continental Scales." *Water Resources Research* 56 (9): e2019WR026793. https://doi.org/10.1029/2019WR026793.

Gallice, A., B. Schaefli, M. Lehning, M. B. Parlange, and H. Huwald. 2015. "Stream Temperature Prediction in Ungauged Basins: Review of Recent Approaches and Description of a New Physics-Derived Statistical Model." *Hydrology and Earth System Sciences* 19 (9): 3727–53. https://doi.org/10.5194/hess-19-3727-2015.

Garner, Grace, Iain A. Malcolm, Jonathan P. Sadler, and David M. Hannah. 2017. "The Role of Riparian Vegetation Density, Channel Orientation and Water Velocity in Determining River Temperature Dynamics." *Journal of Hydrology* 553 (October): 471–85. https://doi.org/10.1016/j.jhydrol.2017.03.024.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment*, Big Remotely Sensed Data: tools, applications and experiences, 202 (December): 18–27. https://doi.org/10/gddm6z.

Graf, Renata, Senlin Zhu, and Bellie Sivakumar. 2019. "Forecasting River Water Temperature Time Series Using a Wavelet–Neural Network Hybrid Modelling Approach." *Journal of Hydrology* 578 (November): 124115. https://doi.org/10.1016/j.jhydrol.2019.124115.

Hawkins, Charles P., James N. Hogue, Lynn M. Decker, and Jack W. Feminella. 1997. "Channel Morphology, Water Temperature, and Assemblage Structure of Stream Insects." *Journal of the North American Benthological Society* 16 (4): 728–49. https://doi.org/10.2307/1468167.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–1780. https://doi.org/10/bxd65w.

Hu, Zhuhua, Yiran Zhang, Yaochi Zhao, Mingshan Xie, Jiezhuo Zhong, Zhigang Tu, and Juntao Liu. 2019. "A Water Quality Prediction Method Based on the Deep LSTM Network Considering Correlation in Smart Mariculture." *Sensors* (Basel, Switzerland) 19 (6). https://doi.org/10.3390/s19061420.

Huang, Maoyi, and Xu Liang. 2006. "On the Assessment of the Impact of Reducing Parameters and Identification of Parameter Uncertainties for a Hydrologic Model with Applications to Ungauged Basins." *Journal of Hydrology*, The model parameter estimation experiment, 320 (1): 37–61. https://doi.org/10.1016/j.jhydrol.2005.07.010.

Jiang, Shijie, Yi Zheng, and Dimitri Solomatine. 2020. “Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning.” *Geophysical Research Letters* 47 (13): e2020GL088229. https://doi.org/10.1029/2020GL088229.

Johnson, Zachary C., Brittany G. Johnson, Martin A. Briggs, Warren D. Devine, Craig D. Snyder, Nathaniel P. Hitt, Danielle K. Hare, and Teodora V. Minkova. 2020. “Paired Air-Water Annual Temperature Patterns Reveal Hydrogeological Controls on Stream Thermal Regimes at Watershed to Continental Scales.” *Journal of Hydrology* 587 (August): 124929. https://doi.org/10.1016/j.jhydrol.2020.124929.

Justice, Casey, Seth M. White, Dale A. McCullough, David S. Graves, and Monica R. Blanchard. 2017. “Can Stream and Riparian Restoration Offset Climate Change Impacts to Salmon Populations?” *Journal of Environmental Management* 188 (March): 212–27. https://doi.org/10.1016/j.jenvman.2016.12.005.

Kaushal, Sujay S, Gene E Likens, Norbert A Jaworski, Michael L Pace, Ashley M Sides, David Seekell, Kenneth T Belt, David H Secor, and Rebecca L Wingate. 2010. “Rising Stream and River Temperatures in the United States.” *Frontiers in Ecology and the Environment* 8 (9): 461–466. https://doi.org/10/bfxq26.

Kelleher, C., T. Wagener, M. Gooseff, B. McGlynn, K. McGuire, and L. Marshall. 2012. “Investigating Controls on the Thermal Sensitivity of Pennsylvania Streams.” *Hydrological Processes* 26 (5): 771–85. https://doi.org/10.1002/hyp.8186.

Madden, N., A. Lewis, and M. Davis. 2013. “Thermal Effluent from the Power Sector: An Analysis of Once-through Cooling System Impacts on Surface Water Temperature.” *Environmental Research Letters* 8 (3): 035006. https://doi.org/10.1088/1748-9326/8/3/035006.

Marcogliese, David J. 2001. “Implications of Climate Change for Parasitism of Animals in the Aquatic Environment.” *Canadian Journal of Zoology* 79 (8): 1331–52. https://doi.org/10.1139/z01-067.

Martins, Eduardo G., Scott G. Hinch, David A. Patterson, Merran J. Hague, Steven J. Cooke, Kristina M. Miller, David Robichaud, Karl K. English, and Anthony P. Farrell. 2012. “High River Temperature Reduces Survival of Sockeye Salmon (Oncorhynchus Nerka) Approaching Spawning Grounds and Exacerbates Female Mortality.” *Canadian Journal of Fisheries and Aquatic Sciences* 69 (2): 330-342. https://doi.org/10.1139/f2011-154.

Moatar, Florentina, and Joël Gailhard. 2006. “Water Temperature Behaviour in the River Loire since 1976 and 1881.” *Comptes Rendus Geoscience* 338 (5): 319–28. https://doi.org/10.1016/j.crte.2006.02.011.

Moore, R. Dan, D. L. Spittlehouse, and Anthony Story. 2005. “Riparian Microclimate and Stream Temperature Response to Forest Harvesting: A Review.” *JAWRA Journal of the American Water Resources Association* 41 (4): 813–34. https://doi.org/10.1111/j.1752-1688.2005.tb03772.x.

Nash, J.E., and J.V. Sutcliffe. 1970. “River Flow Forecasting through Conceptual Models Part I — A Discussion of Principles.” *Journal of Hydrology* 10 (3): 282–290. https://doi.org/10/fbg9tm.

Nelson, Kären C., and Margaret A. Palmer. 2007. “Stream Temperature Surges under Urbanization and Climate Change: Data, Models, and Responses.” *JAWRA Journal of the American Water Resources Association* 43 (2): 440–52. https://doi.org/10.1111/j.1752-1688.2007.00034.x.

Poff, N. Leroy, Brian D. Richter, Angela H. Arthington, Stuart E. Bunn, Robert J. Naiman, Eloise Kendy, Mike Acreman, et al. 2010. “The Ecological Limits of Hydrologic Alteration (ELOHA): A New Framework for Developing Regional Environmental Flow Standards.” *Freshwater Biology* 55 (1): 147–70. https://doi.org/10.1111/j.1365-2427.2009.02204.x.

Poole, Geoffrey C, and Cara H Berman. 2000. “Pathways of Human Influence on Water

Temperature Dynamics in Stream Channels." U.S. Environmental Protection Agency. https://pdfs.semanticscholar.org/49b9/149ad3c2ff489b697166ea0c7d210d93ec4[illegible].pdf.

Raptis, C. E., M. T. H. van Vliet, and S. Pfister. 2016. "Global Thermal Pollution of Rivers from Thermoelectric Power Plants." *Environmental Research Letters* 11 (10): 104011. https://doi.org/2020031315161525.

Rahmani, Farshid, Kathryn Lawson, Wenyu Ouyang, Alison Appling, Samantha Oliver, and Chaopeng Shen. 2020. Data release: Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data: U.S. Geological Survey, https://doi.org/10.5066/P97CGHZH.

Segura, Catalina, Peter Caldwell, Ge Sun, Steve McNulty, and Yang Zhang. 2015. "A Model to Predict Stream Water Temperature across the Conterminous USA." *Hydrological Processes* 29 (9): 2178–95. https://doi.org/10.1002/hyp.10357.

Shen, Chaopeng. 2018. "A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists." *Water Resources Research* 54 (11): 8558–93. https://doi.org/10/gd8cqb.

Shen, Chaopeng, Eric Laloy, Amin Elshorbagy, Adrian Albert, Jerad Bales, Fi-John Chang, Sangram Ganguly, Kuo-Lin Hsu, Daniel Kifer, Zheng Fang, Kuai Fang, Dongfeng Li, Xiaodong Li, & Wen-Ping Tsai. 2018. "HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community". *Hydrology and Earth System Sciences*, 22(11): 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

Sohrabi, Mohammad M., Rohan Benjankar, Daniele Tonina, Seth J. Wenger, and Daniel J. Isaak. 2017. "Estimation of Daily Stream Water Temperatures with a Bayesian Regression Approach." *Hydrological Processes* 31 (9): 1719–33. https://doi.org/10.1002/hyp.11139.

Stajkowski, Stephen, Deepak Kumar, Pijush Samui, Hossein Bonakdari, and Bahram Gharabaghi. 2020. "Genetic-Algorithm-Optimized Sequential Model for Water Temperature Prediction." *Sustainability* 12 (13): 5374. https://doi.org/10.3390/su12135374.

Theurer, Fred D., Ivan Lines, and Terry Nelson. 1985. "Interaction between Riparian Vegetation, Water Temperature, and Salmonid Habitat in the Tucannon River." *JAWRA Journal of the American Water Resources Association* 21 (1): 53–64. https://doi.org/10.1111/j.1752-1688.1985.tb05351.x.

Thornton, P. E., M. M. Thornton, B. W. Mayer, Y. Wei, R. Devarakonda, R. S. Vose, and R. B. Cook. 2016. "Daymet: Daily Surface Weather Data on a 1-Km Grid for North America, Version 3." *ORNL DAAC*, July. https://doi.org/10.3334/ORNLDAAC/1328.

U.S. Geological Survey. 2016. National Water Information System data available on the World Wide Web. (U.S. Geological Survey Water Data for the Nation), accessed in May 2020. http://dx.doi.org/10.5066/F7P55KJN.

Vliet, M. T. H. van, J. R. Yearsley, W. H. P. Franssen, F. Ludwig, I. Haddeland, D. P. Lettenmaier, and P. Kabat. 2012. "Coupled Daily Streamflow and Water Temperature Modelling in Large River Basins." *Hydrology and Earth System Sciences* 16 (11): 4303–21. https://doi.org/10.5194/hess-16-4303-2012.

Wanders, Niko, Michelle T. H. van Vliet, Yoshihide Wada, Marc F. P. Bierkens, and Ludovicus P. H. (Rens) van Beek. 2019. "High-Resolution Global Water Temperature Modeling." *Water Resources Research* 55 (4): 2760–78. https://doi.org/10.1029/2018WR023250.