Rahmani Farshid (Orcid ID: 0000-0001-9241-7206)
Shen Chaopeng (Orcid ID: 0000-0002-0685-1901)
Lawson Kathryn (Orcid ID: 0000-0003-0075-7911)

Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins

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

[1]Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA
[2]U.S. Geological Survey, Middleton, WI, USA
[3]HydroSapient, Inc., State College, PA, USA
[4]U.S. Geological Survey, University Park, PA, USA

Abstract

Basin-centric long short-term memory (LSTM) network models have recently been shown to be an exceptionally powerful tool for stream temperature (T_s) temporal prediction (training in one period and making predictions for another period at the same sites). However, spatial extrapolation is a well-known challenge to modeling T_s and it is uncertain how an LSTM-based daily T_s model will perform in unmonitored or dammed basins. Here we compiled a new benchmark dataset consisting of >400 basins across the contiguous United States in different data availability groups (DAG, meaning the daily sampling frequency) with or without major dams and studied how to assemble suitable training datasets for predictions in basins with or without temperature monitoring. For prediction in unmonitored basins (PUB), LSTM produced an RMSE of 1.129 °C and R^2 of 0.983. While these metrics declined from LSTM's temporal prediction performance, they far surpassed traditional models' PUB values, and were competitive with traditional models' temporal prediction on calibrated sites. Even for unmonitored basins with major reservoirs, we obtained a median RMSE of 1.202°C and an R^2 of 0.984. For temporal prediction, the most suitable training set was the matching DAG that the basin could be grouped into, e.g., the 60% DAG for a basin with 61% data availability. However, for PUB, a training dataset including all basins with data is consistently preferred. An input-selection ensemble moderately mitigated attribute overfitting. Our results indicate there are influential latent processes not sufficiently described by the inputs (e.g., geology, wetland covers), but temporal fluctuations are well predictable, and LSTM appears to be a highly accurate T_s modeling tool even for spatial extrapolation.

Highlights

1. Spatial extrapolation for stream temperature modeling is difficult, but LSTM achieved state-of-the-art prediction accuracy in unmonitored basins (PUB).
2. For temporal tests, training sets should contain basins with as much or more data as the test basins, but all the sites with more than 10 days of data can be used to train models for PUB.
3. Known input attributes do not cover all necessary features so an input-selection ensemble is useful.

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

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/hyp.14400

Keywords: Stream Temperature, Prediction in Unmonitored Basins, PUB, Reservoirs, Machine Learning, LSTM, Deep Learning

1. Introduction:

Stream temperature (T_s, temperature of water in rivers) is an important variable to both environmental health and human decisions. T_s has significant control on riverine biogeochemistry (Zhi et al., 2021), fish life cycles (Bowerman et al., 2018), invertebrate biodiversity (Hill & Hawkins, 2014), and aquatic ecosystem health (Justice et al., 2017), and thus has long been known as an environmental "master factor" (Fry, 1971). Daily T_s predictions enable informed decisions and science-driven policy, and regulations on acceptable water temperatures affect industrial processes such as power plant cooling (Gjorgiev & Sansavini, 2018; Liu et al., 2017; J. Ma et al., 2018) and reservoir operations (Tao et al., 2020; Weber et al., 2017). In line with its importance, T_s modeling has been reported in hundreds of papers, with models of different types ranging from statistical to process-based to data-driven (Marcogliese, 2001; Martins et al., 2012).

T_s is controlled by a range of climatic and hydrological processes such as snowmelt, advection by rain and streamflow, solar radiation, latent heat flux, shading from riparian vegetation, groundwater-surface water exchange, anthropogenic activities, and heat exchange with the land, streambed, and air (Essaid & Caldwell, 2017; Ficklin et al., 2012; Mohseni & Stefan, 1999; Younus et al., 2000). Physically based models seek to represent these processes as mathematical equations, while data-driven models seek to directly learn patterns from data. A small literature survey of process-based and statistical stream temperature models is provided in Appendix A1 and Table S1 in Supplementary Materials. However, these models mainly focus on providing temporal predictions in well-monitored locations, where a model is fitted or calibrated to one site and then predictions are made for new time periods only at that same site. In addition, many models focus on long-term, monthly, or seasonal mean predictions. Because ecosystem-threatening heat waves or temperature shocks can occur on a daily scale (Arambourou & Stoks, 2015), access to accurate daily-scale or even finer resolution predictions could be critical.

Rahmani, Lawson et al. (2021) showed that the long short-term memory (LSTM) algorithm, a type of recurrent neural network, provided accurate results for T_s at 118 well-monitored (>60% daily sampling frequency in training and testing period) sites. These models were tested for long-term daily temporal prediction, i.e., models were trained on a collection of well-monitored sites for one time period and tested on the same sites for another period. The inputs included daily atmospheric forcings and static characteristics of the basins. Our model obtained a median root-mean-square error (RMSE) value of 0.69°C, Nash-Sutcliffe model efficiency coefficient (NSE) of 0.985, and correlation of 0.994. Even after removing seasonality, the median NSE of the residuals was 0.95. These results echo with strong performance metrics reported for LSTM in prediction of soil moisture (Fang et al., 2017; Fang & Shen, 2020), streamflow (Feng et al., 2020; Kratzert et al., 2019; Xiang et al., 2020), and dissolved oxygen (Zhi et al., 2021), even in spatially data sparse regions (Feng et al., 2021a; K. Ma et al., 2021). However, as a data-driven model's quality largely depends on the quality and quantity of the training data, it is unclear how effective such models can be if the sampling frequency is limited, e.g., only about 10% of the days are sampled and sampling may be concentrated in time.

While temporal prediction is important, extrapolating to unmonitored sites is even more crucial because temperature in the vast majority of stream reaches remains unmonitored. For example, for the millions of river reaches in the United States, there are >5000 streamflow stations in the U.S. Geological Survey's (USGS) National Water Information System, yet only around ~820 stations had T_s measurements for >10% of the days between 2004 and 2016, and only 118 had measurement coverage >60% (USGS, 2016). It is well known in the hydrology community that spatial prediction of T_s is challenging, and that sites with little data tend to have much larger prediction errors even if there are data for nearby sites. Gallice et al. (2015) called attention to this problem of T_s prediction in ungauged basins (i.e., basins lacking observations of temperature;

we refer to such basins as “ungauged” or “unmonitored” interchangeably), and reported a mean seasonal RMSE of 1.36°C and R^2 of 0.808 at five ungauged sites using a physics-derived regression model to predict monthly mean temperature, which was calibrated separately for each season. Also, while there have been other recent studies that used newer machine learning models to estimate T_s (Graf et al., 2019; Zhu & Piotrowski, 2020), they did not consider problems with prediction in unmonitored basins (PUB) or spatial extrapolation.

This difficulty with spatial prediction may reflect the fact that there are many local and often-unmeasured mechanisms (called unknown or latent processes) that influence T_s, such as aquifer properties, travel time, snow accumulation patterns, and riparian shade, leading to fine-scale heterogeneity in T_s responses. Consequently, locally calibrated models often tend to be better than large-scale models. For example, McNyset et al. (2015) reported R^2 of 0.95 for models constructed for each individual site but only 0.87 for the model constructed for the whole group of sites in a basin in Oregon. Considering the effects of latent processes and lessons from the literature, it is unclear how to best form model training datasets for monitored sites with varied data availability and unmonitored sites. In addition, understanding the nature of the error can tell us about current weaknesses in the model-data system and opportunities for improvement; for example, bias could indicate that the model fails to capture slow-to-change latent processes such as base-flow rates or mean annual riparian shade, whereas weak correlation could indicate weakness in (potentially implicitly) representing faster-changing processes such surface runoff.

It remains an unexplored and uncertainty-laden path to make use of LSTM for modeling daily T_s in reaches with limited or no data (unmonitored). Many previous efforts of spatial generalizations relied upon the exploitation of geostatistics such as spatial autocorrelation of prediction residuals (Isaak et al., 2017). It is not clear whether LSTM, which by default does not model the spatial structures of errors, can exploit such autocorrelation. Meanwhile, given various

data limitations with T_s, it remains unclear how to best assemble training datasets for the PUB problem as there are two forces at play. On the one hand, in the past we have generally observed that for deep networks, the best approach is to compile as many data points/sites as possible into the training set given an effect we call *data synergy* (Fang et al., 2021). On the other hand, data availability sometimes poses constraints: we have many stream temperature stations with limited data availability and concentrated sampling periods; including them in the dataset may introduce noise and biases into the trained model. Hence, it is not clear how to best choose the training dataset for T_s modeling on sites with varied data availability.

Another potential limiting factor of most existing T_s models is the effects of major reservoirs. There are more than 800,000 dammed reservoirs impeding the world's rivers, including over 90,000 in the United States (International Rivers, 2007; U.S. Army Corps of Engineers, 2018). Reservoirs that are deep, with large surface area, large heat capacity, high thermal stratification, and lower albedo and vapor transfer can exert substantial control on streamflow and the water heat balance. The effect of reservoirs on downstream temperature is further complicated by variable depth release and changing human water, energy and environmental demands that affect decision making (Carron & Rajaram, 2001; Risley et al., 2010). Some of our recent work with streamflow modeling has indicated that LSTM could successfully capture some reservoir dynamics even just using generally available information about reservoirs (Ouyang et al., 2021). Hence, it is worthwhile to examine whether this approach could be extended to improve T_s modeling as well.

Here we propose LSTM for long-term daily T_s modeling at data-scarce (low sampling-frequency), unmonitored and dammed basins. The prediction is spatiotemporal (Jackson et al., 2018) in the sense that the model can output sequences of daily predictions at trained or unmonitored sites where inputs are available so the spatial coverage can be large. We show the

effect of an input-selection ensemble to prediction error, and, to the extent possible, contrast our results with the literature. The study answers several research questions: (1) How well do LSTM-based, contiguous United States (CONUS)-scale models perform for sites with low sampling frequency or no data (unmonitored sites) with generic input information? (2) How can we compile data into a model training dataset to obtain best performance for basins with different sampling frequencies? (3) How much do reservoirs affect LSTM-based temperature model performance?

2. Methods:

We explored how accurately water temperature dynamics could be captured under different data-availability and reservoir-influence scenarios, given broadly available meteorological forcing data, streamflow rate observations, and basin characteristics. Here we predicted daily mean water temperature (T_s) using an LSTM-based model with similar structure to that in recent T_s work (Rahmani, Lawson, et al., 2021). Each model was trained and tested on basins across the contiguous United States (CONUS). The overall prediction and loss equations can be written as:

$$T_{sim,b}^{t-\rho:t} = LSTM\left(F_b^{t-\rho:t}, A_b\right) \quad (1)$$

$$L = \sum_b^B \sum_{i=t-\rho}^{t} (T_{sim,b}^i - T_{obs,b}^i)^2)/n \quad (2)$$

where A_b represents all attributes (static values) in a particular basin (*b*), $F_b^{t-\rho:t}$ is the continuous forcing values from day $t-\rho$ to day t in basin *b*, and $T_{sim,b}^{t-\rho:t}$ is the simulated stream temperature from day $t-\rho$ to day t in basin *b*. L is the total loss in days that observed data are available, *B* is the number of basins in a minibatch (a random subset of the basins and time series to be aggregated for calculating the loss; for each training epoch, we will loop through many mini-batches so that on average all basins and all time series have been used once), *n* is the total

number of observed days, and $T^i_{sim,b}$ and $T^i_{obs,b}$ are the simulated and observed stream temperatures, respectively, in day *i* and basin *b*.

2.1. Datasets

We focused on stations that were included in the Geological Attributes of Gages for Evaluating Streamflow dataset, version II (GAGES-II) (Falcone, 2011). We downloaded mean daily observed streamflow and T_s data from the USGS National Water Information System (USGS, 2016). As measured streamflow was one of the desired inputs to our model, we only considered stations that had continuously observed streamflow data and at least 10 days' worth of observed T_s data from 2010 to 2016, resulting in 455 stations. Meteorological forcing data (i.e. minimum and maximum daily air temperature, precipitation, solar radiation, vapor pressure, and day length) were extracted from the Daymet dataset (Thornton et al., 2016) using Google Earth Engine (GEE) (Gorelick et al., 2017) by interpolating the stations' watersheds' shape files with the gridded Daymet dataset. Watershed attributes pertinent to climate, topography, land cover, and reservoirs were provided in GAGES-II and utilized as model inputs. Table S2 lists the forcing and attribute inputs.

Of the 455 sites that had at least 10 days of water temperature observations, 415 sites had observations for at least 10 percent of the days during both the training (2010/10/01 to 2014/09/30) and testing (2014/10/01 to 2016/09/30) time periods. For these 415 sites, the median basin area was 1,017 km^2 with the maximum and minimum basin areas being 49,264 km^2 and 2.1 km^2, respectively. The mean T_s, according to sampled data, was 12.4°C, with maximum and minimum values of 34.3°C and –2.2°C, respectively. Basins had a mean of 5.7 major dams with the median being 1 major dam (254 basins had at least one major dam). The maximum number of major dams in a basin was 154 and the minimum was 0 (Falcone, 2011; US Army Corps of Engineers, 2018). Major dams are characterized as having more than 5,000 acre-feet of normal

storage capacity, more than 50 ft in height, or a maximum storage capacity of more than 25,000 acre-feet (National Atlas of the United States, USGS, 2009).

2.2. Long short-term memory (LSTM) models

Stream temperature dynamics are affected by processes that operate from hourly to monthly or longer timescales such as rainfall-runoff processes, groundwater interaction, snow melting effects, and many more. Therefore, tracking the impacts and relationships between these processes requires methods with the capabilities of both short-term and long-term memory. The long short-term memory (LSTM) algorithm has grown immensely popular for hydrological applications in recent years, and is designed to learn and keep information for long periods using components called memory cells and gates. Memory cells store the information separately from recurrent hidden states to resolve the vanishing gradient issue faced by other recurrent neural networks, while 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 (Feng et al., 2020; Hochreiter & Schmidhuber, 1997).

We applied standardization to all inputs and target values consistently across training and test datasets. Standardization helps the model, by way of the loss function (the target metric that is minimized during training), give equal consideration to all watersheds. First, we divided streamflow by the product of basin area and annual mean precipitation to obtain dimensionless runoff. We then transformed this dimensionless runoff, along with actual daily precipitation data, to a more Gaussian distribution (Feng et al., 2020) using equation (3):

$$v^* = log_{10}(\sqrt{v} + 0.1) \quad (3)$$

where v^* is the new variable after transformation, and v represents the variable before transformation. Next, streamflow, precipitation, and all other inputs and water temperature observations were standardized using the following formula:

$$x_{i,new} = \frac{(x_i - \bar{x})}{\sigma} \quad (4)$$

in which $\underline{x}$ and σ are the mean and standard deviation, respectively, for each input variable, x_i is the raw value, and $x_{i,new}$ is the standardized value. All results in this study are reported after destandardization, which is the complete reversal of the standardization procedures.

Before starting model training and testing for specific experiments, we ran multiple tests to determine the best values for hyperparameters in our model. The hidden layer size (the intermediate layer located between input and output layers) was selected by testing with different values (50, 100, 150, and 200), with constant values of the other hyperparameters. Rho (maximum number of days used for backpropagation steps for each training sample) was tested with values of 365, 274 (9 months), and 183 (6 months) without changing other hyperparameters. Because stream temperature has an annual cycle, it was likely we would get the best results with rho equal to 365, but we wanted to test other values in case a simpler model could achieve similar performance. The number of epochs used (one epoch means using all the training data once) was selected by testing the model with 1000, 2000, and 3000 epochs where the hidden layer size, rho, and batch size (determines the number of samples given to the model for loss function calculation before updating weights in a training step) were equal to 100, 365 days, and half of the number of sites used in training section, respectively. Dropout rate (the probability that some weights will be randomly set to 0 during model training, which helps prevent model overfitting) values of 1 and 0.5 were tested. Based on the authors' experience, satisfactory results were obtained if the batch size was set to half of the number of sites in the dataset. Therefore, the final hyperparameters chosen were a hidden layer size of 100, rho of 365 days, 2000 epochs, a batch size equal to half of the sites in the training dataset, and a dropout rate of 0.5. For each model (for which hyperparameters, attributes, forcings, and training data were all fixed) we trained the model six times with different random initial weights each time; final predictions for each model were daily means of predictions from those six replicates.

The loss function was root-mean-square error (RMSE) summarized for the minibatch. We report RMSE, bias, unbiased root-mean-square error (ubRMSE), Pearson correlation, and Nash-Sutcliffe efficiency coefficient (NSE) (Nash & Sutcliffe, 1970) to enable comparison to other studies. Further, following Rahmani, Lawson et al. (2021), we computed the residual temperature (T_{res}) as the difference between daily mean water (T_s) and air temperatures (T_{air}): $T_{res}=T_s-T_{air}$. We then calculated NSE and Pearson correlation also based on this seasonality-removed T_{res} instead of T_s, because otherwise seasonality alone could explain much of the variability in T_s:

$$NSE_{res} = 1 - \frac{\sum_{i=1}^{n}(T_{i,obs} - T_{i,sim})^2}{\sum_{i=1}^{n}[(T_{i,obs} - T_{i,air}) - \bar{T}_{obs-air}]^2} \quad (5)$$

$$\bar{T}_{obs-air} = \frac{\sum_{i=1}^{n}(T_{i,obs} - T_{i,air})}{n} \quad (6)$$

$$T_{i,air} = \frac{(T_{i,max} + T_{i,min})}{2} \quad (7)$$

$$Corr_{res} = \frac{\sum_{i=1}^{n}[(T_{i,obs} - T_{i,air}) - \bar{T}_{obs-air}][(T_{i,sim} - T_{i,air}) - \bar{T}_{sim-air}]}{\sqrt{\sum_{i=1}^{n}[(T_{i,obs} - T_{i,air}) - \bar{T}_{obs-air}]^2 \sum_{i=1}^{n}[(T_{i,sim} - T_{i,air}) - \bar{T}_{sim-air}]^2}} \quad (8)$$

$$\bar{T}_{sim-air} = \frac{\sum_{i=1}^{n}(T_{i,sim} - T_{i,air})}{n} \quad (9)$$

in which $T_{i,obs}$ and $T_{i,sim}$ show the observed and simulated, respectively, daily mean stream temperatures, $T_{i,min}$ and $T_{i,max}$ represent the lowest and highest, respectively, daily air temperature values recorded in the meteorological forcing data, and $T_{i,air}$ is a representation of mean daily air temperature and calculated from $T_{i,min}$ and $T_{i,max}$. $\bar{T}_{obs-air}$ and $\bar{T}_{sim-air}$ indicate the differences between daily mean air temperature and either daily observed or simulated stream temperatures, respectively, while NSE_{res} and $Corr_{res}$ illustrate the residual NSE and residual correlation coefficient values, respectively. Index *i* indicates the *i*-th day of the testing period, and *n* is the total number of days of the testing period in which observed stream temperature data were available.

2.3. Training and testing models on different data availability groups

Many of the GAGES-II basins had water temperature observations available for only a fraction of the year. We defined three data availability groups (DAG, Figure 1), using *p* to denote the percentage of days for which temperature measurements were present. The $DAG_{p>99}$ contained basins where temperature observations were present for at least 99% of days during both training and testing periods. $DAG_{p>60\%}$ included basins in p>99% as well as basins with observations for at least 60% of days, and $DAG_{p>10\%}$ was defined similarly (Figure 1). $DAG_{60>p>10}$ has somewhat seasonally imbalanced measurements with 30.5% of available observations occurring in the summer,18.6% in the winter, and ~25% in each of spring and fall. We trained three models, each using all sites from the respective nested DAGs (p>99%, p>60%, and p>10%), which we call $LSTM_{p>99\%}$, $LSTM_{p>60\%}$, and $LSTM_{p>10\%}$. We tested each model in two different ways: temporal prediction using two different configurations, and then three experiments for spatial generalization (prediction in unmonitored basins, or PUB). We expand on the details of these tests below.

[Insert Figure 1]

For temporal prediction, the models were trained using data from 2010/10/01 to 2014/09/30 and test predictions were compared to data from 2014/10/01 to 2016/09/30 on the same basins. We would like to note that temporal prediction is still different from traditional site-by-site calibration: because a uniform model was trained for multiple sites, this model has to learn how to use different static attributes to modulate T_s fluctuations in time, and hence it is necessarily more complex than a group of many models fit to only one site apiece, but it is simultaneously constrained by all sites. On a side note, Rahmani, Lawson et al. (2021) compared temporal prediction results to locally-fitted autoregressive models with exogenous inputs (ARX), which represent natural persistence, and found that LSTM (RMSE=0.69^{o}C) significantly outperformed the ARX model (1.41^{o}C). Because the foci of this paper are on the spatial challenges of

unmonitored and dammed basins, we omitted the comparison to ARX or similar persistence models here.

In the first configuration of the temporal prediction, the model used for prediction in each basin was the model trained on basins with data availability matched to that of the test basin: if a basin had p=99%, $LSTM_{p>99\%}$ was used for prediction; if a basin had p=65%, $LSTM_{p>60\%}$ was used; and if a basin had p=11%, $LSTM_{p>10\%}$ was used. Results from this test, which we refer to as the *matching-DAG approach*, tell us how model performance varied as a function of data availability. Despite the caveat that there may be correlation between DAG and basin characteristics, we can still learn simple criteria for including sites into the training datasets. In the second configuration, the three models were tested on the same subset of basins: basins all having data availability p>99%. This comparison was to advise us whether we should include basins with lower sampling frequency in the training set if we are only interested in applying the model on a basin with a high sampling frequency. As the amount of training data available to the model increased as the p threshold lowered, we refer to this as the *maximum-sites approach*.

2.4. Testing the models on prediction for unmonitored sites (PUB)

To test model performance in cases of spatial extrapolation, also known as prediction for unmonitored basins (PUB), we ran three experiments where we tested our models with different holdout basins, which means they were not included in the training set. In the first experiment (random holdout), we randomly selected 40 basins that had at least 60% data availability to serve as the holdout basins, which constituted 10% of the total dataset. The models were trained with the remaining basins in the three DAGs. Because these basins were randomly selected, this experiment avoids the potential confounding correlation between the data availability and other physical attributes. This test was to inform us as to which training set would be best to choose for PUB scenarios.

To confirm the robustness of the conclusion, we ran another experiment (p<10% holdout) where the holdout basins were not in the three training DAGs due to having low sampling density but still had at least 10 days' worth of data. 40 basins met these criteria across CONUS to be out-of-sample basins in this test, and we tested each of the models trained on three DAGs on this group. This kind of test experiment is sometimes necessary because we may not have enough high sampling frequency sites to simultaneously train and test on, and thus we want to maximally utilize their data for training. However, the p<10% sites may have certain properties that are correlated with model performance. This experiment thus also examines the effect of this data-economic test strategy.

Finally, to be more comparable to the PUB results published in the literature, in the third experiment (northwest PUB), we held out five p>99% sites in the US Northwest (latitude > 42 and longitude < -120.2), which is closest in latitude to the Swiss basins used in the current T_s benchmark model reported in Gallice et al. (2015), to facilitate comparison. We trained a model using the rest of the p>60% basins across CONUS and, based on daily predictions, calculated mean monthly values to compare with their results. We would like to emphasize that Gallice's results are not fully comparable to the U.S. Northwest because these regions have different climate patterns, e.g., U.S. Northwest is influenced by Pacific Ocean oscillations like El Nino/La Nina while the Swiss Alps are influenced by the Atlantic Ocean. Hence, it is important not to read too much into the small differences. However, we nonetheless present this comparison because it seems more reasonable than comparing the whole CONUS to Switzerland.

2.5. Input-selection ensemble

Prediction in unmonitored basins can be an especially challenging problem because of model overfitting, a not uncommon scenario where models perform well for basins with data included in the training set but miss some of the underlying physics needed for accurate modeling (Gallice et al., 2015). This can lead to poor prediction accuracy when applied to basins outside of

the training set, especially for those with environmental conditions different from the basins that provided the training data. Here we tested an approach we call an *input-selection ensemble*, which we recently showed could mitigate some of this risk for streamflow prediction in both ungauged basins and ungauged contiguous regions (Feng et al., 2021b). The theory is that when there are not enough basins with measured data available for model training, we cannot accurately resolve the influence of each static basin attribute (e.g., land cover, or soil characteristics), and so the model will have a large variance around how the static attributes influence the model output. For new conditions where an overfitted model could perform very poorly, a minimal-attribute model (with accompanying low likelihood of overfitting) would be much more likely to have a correct understanding of the main causal factors, thus resulting in more accurate predictions. We hypothesized that ensembling (averaging) across multiple models with different amounts of input attributes could reduce the overall variance for T_s as it did for streamflow.

In the set of three DAG models $LSTM_{p>99\%}$, $LSTM_{p>60\%}$,and $LSTM_{p>10\%}$ discussed until this point, we used all available attributes as inputs (A_T in Table S2). To test our hypothesis about input selection ensemble, we needed three additional model versions, with decreasing numbers of attributes included as inputs. The attributes to exclude were conceptualized as being similar to other existing attributes. In other words, we attempted to select attributes in a way that preserved critical information while removing redundancies and lower-impact information. In the first additional set (A_{R1}), we removed some attributes from the A_T list that we thought would be redundant to those still included, e.g., attributes related to average distance of basins' outlet to all major dams. But we kept the distance of the outlet from the nearest major dam as we hypothesized the nearest major dam is the most influential one to stream temperature compared to the farther ones (Kędra & Wiejaczka, 2018) (please see Table S2). The second set (A_{R2}) excluded these along with even more attributes from the A_T list, e.g., some related to annual-

mean climate attributes. In the third set (A_{R3}), only a few critical attributes such as drainage area, stream density, reservoir storage, land cover fractions, slope, and distance of gage location to major dams, were kept as model inputs. Hidden layer sizes were reduced to 80 for A_{R1} and 70 for A_{R2} and A_{R3} to match the reduced complexity of those attribute sets. These four model versions were each applied to the PUB scenarios described previously, and results were averaged to obtain the input-selection ensemble result (Eq. (11)). These results, along with the results from the full-attribute model alone (general PUB test method, Eq. (10)), were compared to the actual T_s observations. For our experiments, separate models with each of the three additional input versions were trained for each of the three nested DAGs ($DAG_{p>99\%}$, $DAG_{p>60\%}$, $DAG_{p>10\%}$).

$$T_{sim,gen}^{i,b} = LSTM\ (F\ , A_T) \tag{10}$$

$$T_{sim,ens}^{i,b} = \frac{(LSTM\ (F\ ,A_T) + LSTM\ (F\ ,A_{R1}) + LSTM\ (F\ ,A_{R2}) + LSTM\ (F\ ,A_{R3}))}{4} \tag{11}$$

$T_{sim,gen}^{i,b}$ and $T_{sim,ens}^{i,b}$ represent the simulated stream temperatures obtained using the general PUB test method and the ensemble-selection method, respectively, for basin *b* and in day *i*.

2.6. Reservoirs: Presence or absence of major dams

To understand the effect of reservoirs on T_s modeling, we identified whether there were reservoirs in each basin from the GAGES-II dataset, and divided each DAG into two sub-groups: basins with at least one major dam upstream, and those without any. The six new dam-DAGs were therefore $DAG_{p>99+dam}$ (with dam), $DAG_{p>99-dam}$ (without dam), $DAG_{p>60+dam}$, $DAG_{p>60-dam}$, $DAG_{p>10+dam}$, and $DAG_{p>10-dam}$ (Table 1). Models were trained on each of these six new dam-DAGs, and we evaluated their performance in temporal prediction. [Insert Table 1]

3. Results and Discussion

3.1. Temporal Prediction

Overall, models trained using the matching-DAG approach (Section 2.3) performed very well across the CONUS, giving state-of-the-art performance even for basins that were not sampled frequently. The median RMSEs were 0.801°C, 0.832°C, and 0.916°C for basins with data availability p>99%, 99%>p>60%, and 60%>p>10%, respectively (Figure 2a). Lower data availability (p) in the training set led to a moderate decline in model performance, which is consistent with the data-driven nature of the model. The corresponding median NSEs were all above 0.976, which was similar to what was reported in Rahmani, Lawson et al., (2021). As the dataset expands, the median RMSE values were slightly higher in this current work, however, potentially due to inclusion of dammed basins. After removing seasonality by subtracting air temperature from both predictions and observations, the median correlations were all higher than 0.965 for all groups, indicating most of the variability beyond seasonality was captured (Figure 2a). When the three DAG models were all tested on the same sites (sites having p>99%, which were included in the training dataset for all models), both bias and correlation deteriorated as the p threshold for training data was lowered (Figure 2b). Notably, despite this decline, even the lowest DAG 60%>p>10% test group reported better metrics than had been previously reported in the literature for daily temperature prediction using process-based or statistical methods (Table S1). Rahmani, Lawson et al. (2021) concluded that LSTM is extremely well-suited to capturing T_s fluctuations with hysteresis and nonlinear behaviors. Here we show that this conclusion seems to hold true even for sites with much lower sampling frequencies, e.g., 60%>p>10%.

Each of the DAG models showed similar spatial patterns in performance. These models generally performed better in the eastern half of CONUS than the western half, with correspondingly higher NSE and lower RMSE values (e.g. Figure 3). Most of the stations in the eastern half of CONUS have NSE values above 0.975 and RMSE values below 0.9°C. We noticed

that a belt of basins in the longitudinally central CONUS, going from North Dakota to Texas, tended to have larger RMSEs (e.g., Figure 3a). This belt has traditionally been difficult to capture for process-based as well as deep learning streamflow models, as discussed in Feng et al. (2020), for possible reasons including the presence of very large basin areas with concentrated runoff production, unclear basin boundaries, and existence of cross-basin groundwater flows (O'Sullivan et al., 2020; Schaller & Fan, 2009). From our results here, it seems the difficulty of hydrologic prediction carries over to T_s prediction as well, which was expected given the strong influence of streamflow on T_s. A few other sites with larger errors occurred in the state of Washington (U.S. Northwest). LSTM is strong at modeling seasonal snowpacks, and we hypothesized that LSTM has learned to internally accrue memory that mimics snow, but perhaps there was not sufficient memory for inter-annual snowpacks, as would be found in glaciers. Additionally, the Pacific Northwest has a substantial heterogeneity in the contributions of shallow and deep groundwater, which have different signatures on streams' thermal regimes (Hare et al., 2021).

[Insert Figure 2]

[Insert Figure 3]

3.2. Prediction for unmonitored basins (PUB) experiments

Over the entire CONUS, our PUB experiment found state-of-the-art performance with the model trained on all available basins (p>10%, the maximum-site approach) and with the input selection ensemble. For the random holdout test, the CONUS-median bias was -0.21°C, RMSE was 1.129°C, ubRMSE was 0.98°C, NSE was 0.971, and r^2 was 0.983 (Figure 4a). While the RMSE was somewhat higher than for temporal prediction, all of these metrics were better than most values reported in the literature. It seems spatial extrapolation may introduce some bias, perhaps due to lack of knowledge of latent, local processes, but the seasonality and fluctuations were well captured. To put the metrics into perspective, the literature review in Gallice et al. (2015) included a list of daily RMSE values for PUB cases: 1.8°C (DeWeber & Wagner, 2014), 1.4°C

(Gardner & Sullivan, 2004), and 2.1-2.7°C (Stefan & Preud'homme, 1993). In cases where R^2 values were reported, they were 0.71 (Stewart et al., 2015), and 0.70 (Westenbroek et al., 2010), all of which were substantially lower than the CONUS-median value we reported above (0.971). Not only does LSTM exceed traditional models in terms of PUB metrics, but also its PUB metrics were even better than traditional models' temporal prediction (on calibrated or training sites) (see Table S1). For T_s modeling, where spatial extrapolation was deemed to be much more challenging than temporal prediction, this result shows that traditional models have been underutilizing the information in the inputs.

As discussed in Methods (section 2), to compare with Gallice et al. (2015), we trained a different model and calculated its monthly evaluation metrics in five p>99% holdout basins in a latitude-longitude box (see Figure 3(c)) in the U.S. Northwest that has the closest temperature regimes and data density to Switzerland. In this test set, our LSTM's RMSE was 1.07°C, NSE was 0.937, and r^2 was 0.942 for monthly mean prediction through the testing years. These metrics compare favorably to the values reported in Gallice et al. (2015) (an RMSE of 1.45°C and an R^2 of 0.93). In another study at a higher latitude, Jackson et al. (2018) reported an overall RMSE of 1.6°C for their leave-one-basin-out cross validation for 223 Scottish sites. We caution that, despite our best effort to enable comparisons, these results are still not directly comparable because different basins were tested and we have different data density and climate patterns. However, the comparisons indicate that our PUB model represents noticeable advances.

We notice that for PUB, it was the most beneficial to use the maximum-site approach, which is in sharp contrast to the temporal prediction experiments described above, where a matching-DAG approach was most effective. For random holdout sites, as the training set expanded from $DAG_{p>99}$ to $DAG_{p>10}$ in input selection models, the median RMSE improved from 1.696°C to 1.129°C (Figure 4a). In fact, for the full-attribute model, all of the metrics (RMSE,

ubRMSE, R^2, and $Corr_{res}$) improved as the training sites increased. The improvement in median RMSE was largest between $DAG_{p>99\%}$ and $DAG_{p>60\%}$, likely because the number of training basins increased substantially: a three-fold increase, from 99 to 306 basins, whereas the change from $DAG_{p>60}$ to $DAG_{p>10}$ was only a thirty percent increase, from 306 to 415 basins. We expand on this contrast in Section 3.3.

[Insert Figure 4]

The p<10% holdout experiment confirms that the maximum-site approach works the best for PUB (Figure 4b). The errors had the same pattern as Figure 4a, with more training basins giving better metrics, but for these p<10% sites, there was a significantly larger bias. This agreement indicates our conclusion is robust while the larger errors show that there is indeed some systematic correlation between data availability and test metrics. One of the reasons, as will be discussed in section 3.4, is that the models generally have better performance in fall, winter, and spring than summer. Summer data constitute more than 50 percent of observations in p<10% test sites while only 25 percent of data in the p>60% test sites (Table 2). On a side note, this result cautions us that the data-economic testing scheme may have some limitations.

The input-selection ensemble slightly outperformed the full-attribute model, across all DAG training datasets. Compared to the full-attribute model trained on each DAG, the corresponding input-selection ensemble generally had slightly less negative median bias (Figure 4). For NSE, even though the medians were similar between two kinds of models, the input-selection ensemble models were less likely to produce very poor performance as the lower whisker is shorter. This supports our theory that the relationships built on static attributes are uncertain, so ensembling in the dimension of the static attributes can reduce the variance. For models slightly overfitted to the static attributes, some basins may be modeled very well and some

may be very poor. However, utilizing the input-selection ensemble substantially reduced the prediction risks of running into major failures.

3.3. Selecting appropriate training sets for sparsely-monitored or unmonitored basins

Our results indicate that, without using the input-selection ensemble, the best modeling results are achieved by selecting the matching-DAG approach to form a training dataset for basins with extensive records, while using the maximum-site approach for PUB. With the input-selection ensemble, if the ranking became slightly nuanced but at least the maximum-site approach will not produce a noticeably inferior model. As discussed in the Introduction (section 1), when previously working with deep learning models for streamflow, soil moisture, and other environmental variables, we have repeatedly observed that deep networks benefit from the inclusion of larger quantities and larger diversity of data (Fang et al., 2021; Feng et al., 2020; Shen, 2018). This observation was consistent with our PUB results (Figure 4), but in conflict with our temporal prediction results. This conflict is more apparently seen from the performance matrix in Table 2. We see that errors increase as we go from the diagonal to right, indicating the inclusion of more sites degrade the results for temporal prediction. Meanwhile, errors generally go down as we go from the left side of the table to the diagonal, indicating expanding sites is beneficial to spatial extrapolation.

Many reasons could contribute to the apparent contradiction described above, the first being correlation between DAG and basin characteristics. The adverse effects of including more basins in the training set for temporal prediction are likely related to the seasonality of data collection in basins in lower DAGs. Compared to sites with data availability p>99%, sites with data availability 99%>p>60% and 60%>p>10% had more data present in the summer, when the temperatures are higher and model RMSEs are also larger (Table 1). In addition, the models encountered more basins with major dams as the p threshold was lowered (Table 1), which also

had a material effect on T_s prediction accuracy (more in Section 3.5) and may have caused the models trained on more sites to learn patterns that did not occur in the extensively monitored sites. In this case, bringing together data from many basins with low sampling frequency may introduce noise to the supervision of the model, thus slightly degrading the performance for extensively sampled sites.

Our study results indicate that the maximum-site approach is beneficial for PUB because the model and inputs are not sufficient to fully capture the fundamental relationships, so the model relies on having training data that are *close* to the test basin to be accurate: more basins that are spatially proximate or physiographically similar to the ungauged test basins can better represent that part of the input space. Due to the limited number of basins (a few hundred basins is not dense compared to the dimension of the static attributes), the model cannot fully resolve the effect of each input. Moreover, there are important bias-inducing latent processes so that the inputs are not a complete description of the problem, leading to spatial nonstationarity in the relationships between forcings, attributes, and stream temperature responses. It is then different to accurately infer the bias (mean difference between prediction and observations) for PUB. When we expand to p>10% (for PUB), we simply have more basins that are adjacent to the test basins so the model has a better spatial coverage. Previous research already demonstrated that the error residuals tend to be spatially autocorrelated (Isaak et al., 2017). Thus, having similar or adjacent basins to the test ones can reduce the errors. Deep networks can utilize such proximity even if it does not explicitly model error autocorrelation (Fang et al., 2020; Gal & Ghahramani, 2016). Furthermore, there are some examples (e.g., in Georgia) where there is a negative autocorrelation between two adjacent basins, which is probably because these basins are geologically and topographically distinct (O'Sullivan et al., 2019, 2020). A deep learning approach may be able to model both positive and negative relationships more flexibly than a statistical spatial covariance model.

[Insert Table 2]

3.4. Model performance in different seasons

When we break down the results of temporal prediction by season, we can see that the NSEs were better for spring and fall, the two seasons with large temperature shifts, but lower for winter and summer (Figure 5). Pooling together dammed and undammed basins, the median NSE values for temporal prediction (p>10% model) in spring, summer, fall, and winter were 0.942, 0.845, 0.947, and 0.890, respectively, and the corresponding median RMSE values were 0.912°C, 0.827°C, 0.864°C, and 0.742°C. As noted in Rahmani, Lawson et al. (2021) and other studies, previous statistical models often fail in winter in northern basins where air temperature and water temperature are decoupled. In contrast, our LSTM-based models had the smallest median RMSE (0.742 for p>10%, also due to small variation) and high median correlation (0.963 for p>10%) in winter. This is potentially enabled by LSTM's ability to model thresholded functions (relevant to freezing conditions), keep track of long-term memory, and utilize time-dependent relationships.

In contrast, LSTM has relatively lower performance in the summer, mostly caused by a larger bias and lower correlation for either temporal prediction (Figures 5, S1, and S2) or PUB. The CONUS-median bias for summer was closer to zero compared to other seasons, but the range was larger for the summer and noticeably smaller for the winter. We observe that LSTM sometimes underestimated summer flash peaks (Figure 6c), which were likely caused by surface runoff washing out the heat from the land surface which is then countered by the colder base flow. Previous results with streamflow indicate that relative to their performance in other periods, LSTM models sometimes had difficulty capturing baseflow (Feng et al., 2020). Thus, the stream temperature model may not be able to accurately account for the effect of the cooler base flow in its internal representations (our temperature model may not have the ability to track base flow

and runoff dynamics). Because base flow is often a substantial fraction of total flow during the summer, any inaccurate estimations or representations could result in larger errors in temperature for this season.

A state-of-the-art regional-scale study reporting strong metrics for August was the NorWest model (Issak et al., 2017), who reported a spatial RMSE of 1.1°C for their mean August temperature prediction (one data point per site) for the western United States. They collected >63,000 site-years worth of data from >22,000 sites. Our dataset did not allow the calculation of a comparable spatial RMSE because there are not enough basins to simultaneously train a model and have sufficient PUB sites to test it on. However, their work does indicate it will be highly beneficial to adapt the daily LSTM model to utilize a larger (though temporally less well-sampled) dataset with more diverse attributes in future modeling efforts. Pathways like loss function modification can accommodate such kinds of data.

[Insert Figure 5]

3.5. Impact of reservoirs

For temporal prediction, although we saw a strong, adverse effect of reservoirs on the accuracy and bias of our water temperature model, overall model performance for dammed basins remained quite strong. The presence of major dams led to an increase in ubRMSE by a mean of 0.15°C (Figure 7). Compared to the basins without reservoirs, there was a noticeable increase in the range of bias and RMSE in summer and fall (Figure 5). However, despite this increase in error, LSTM produced a median RMSE of <0.88°C for the dammed basins, which is still smaller than those reported in most other studies (Table S1). Even in the most adversarial situation, for the CONUS-scale PUB test sets, the median RMSE, NSE, and r^2 for dammed basins (23 basins) were 1.537°C, 0.861 and 0.964 for $PUB_{test\_p<10\%}$ (23 basins) (Figure S3); and were 1.202, 0.972, and 0.984 for $PUB_{test\_p>60\%}$ (22 basins) (Figure S4), respectively.

The lower 4 panels in Figure 6 showed some sample time series for PUB predictions, while the right 4 panels show basins with major dams. Both PUB and dams have negative effects on the simulations. The PUB panels have noticeably more continuous errors (autocorrelated in time), e.g., around 2016-07 in Figure 6g, which are rare for in-training basins (Figure 6a-d). In some basins with reservoirs, the T_s can sometimes show erratic and sudden changes (Figure 6b,f). However, in some other situations, the model captured the fluctuations quite well (Figure 6d), even in a PUB scenario (Figure 6h).

[Insert Figure 6]

[Insert Figure 7]

The LSTM model is informed on some attributes of the reservoir (degree of regulation, normal capacity, etc.). However, the LSTM model does not know other specifics such as the dynamical surface area, albedo, current water depth, or release depth, and thus cannot infer the extent of heating, heat storage, stratification of reservoir water, or propagation of reservoir

influences to the monitoring site. The larger ubRMSE may also be due to the schedule of the release of reservoir water, for those reservoirs that are actively managed, being unpredictable for the LSTM model. Nevertheless, we note that even for the reservoir group, the RMSE of the LSTM model appears to be smaller than those reported in other studies (Table S1).

3.6. Further discussion

In contrast to previous work that utilized spatial structure in error residuals, we train models with a loss function defined over an entire dataset and obtain a uniform model that resolves the effects of static basin attributes. Because these attributes are themselves autocorrelated, we expect the deep network to intrinsically exhibit characteristics of spatial autocorrelation even without any explicit supervision, which was observed previously with soil moisture (Fang et al., 2020). However, because the model does not rely on explicit autocorrelation assumptions (but depend on spatial pattern in input attributes), the model can flexibly capture highly complex, anisotropic, multidimensional and sometimes even negative autocorrelations. For example, the autocorrelation of stream temperature is predominantly positive in connected rivers while it may in fact be negative in short distances in unconnected rivers. Such spatial relationships will be exceedingly difficult to represent in spatial variograms, but can be seamlessly represented if they are caused by small-scale heterogeneity in basin attributes, e.g., land use, geology, or upland-lowland configurations.

Fortunately, bias is arguably a less severe limitation than low correlation would have been, and so in lieu of additional monitoring data through novel means allowing for bias reduction, we can provide predictions on a relative-change basis, predicting change in temperature from yesterday. A primary suspect for inducing bias is poor characterization of geologic formations, which makes it difficult to accurately model base-flow contributions (Briggs & Hare, 2018; Johnson et al., 2017; O'Sullivan et al., 2019). This is perhaps also why it was previously found that including

observed streamflow information in model inputs could somewhat improve T_s estimates (Rahmani, Lawson, et al., 2021). Secondly, the presence of water withdrawals used for irrigation and other purposes, or release of heated water from power plants, could lead to systematic errors. These are not problems unique to deep learning, however; without local observations, these errors are likely difficult to correct with process-based models as well. The errors can be alleviated in the future by increasing observational constraints, perhaps through the use of new and unconventional approaches such as satellite remote sensing of stream temperatures (Martí-Cardona et al., 2019), thermal infrared imagery (Caldwell et al., 2019; Dugdale et al., 2019), or citizen scientists.

The data used here represent the best-instrumented sites from USGS, and 415 locations are only a tiny fraction of the millions of river reaches in the United States. In the future, the combination of process-based modeling and machine learning may allow more robust predictions on a global scale which are already started by other scholars (Jia et al., 2020; Karpatne et al., 2018; Read et al., 2019; Tsai et al., 2020).

3.7. Limitations

While the simulations gave unprecedentedly strong metrics for most basins with reservoirs, there are some dammed basins with sudden T_s spikes that were missed by the model, e.g., in Figure 6b. There could be days with reservoir-controlled low flows combined with heat stress for the aquatic ecosystem and downstream heat-sensitive water users. Hence, these scenarios would benefit from further investigation and careful error quantification with respect to the extreme values, which is outside of the scope of this paper. Prediction in ungauged basins is still a difficult task, especially for the basins that have data in the outer bounds of training data.

4. Conclusions

This work expands recent research of deep-learning-based modeling of stream temperature to data-sparse, unmonitored, or dammed basins. While these challenges slightly degraded model accuracy, LSTM still presented state-of-the-art performance for daily stream temperature predictions. Even under the most adverse situations tested here, with significant amounts of missing data and the presence of major reservoirs, the model still produced strong NSE in the test period. Extrapolation of the model to basins outside of the training set tended to incur larger bias, likely due to uncaptured processes or attributes. However, the RMSE and r^2 metrics remained substantially higher than the results reported in the literature. The problem of prediction of stream temperature in basins with reservoirs has not been adequately resolved in the past and there has not been a comparable study. We showed that LSTM's performance is indeed affected by reservoirs but overall, the model was still functional. simulating this effect in a process-based manner would have required far more input information about the reservoirs and their operations.

The results of this study can help select the right training dataset to obtain the best-performing models. For a basin with observed stream temperature available, the best results were obtained by pooling data from basins with similar or more available data to include in the training set. For a basin without observed stream temperature available, the best results were obtained by including all basins with stream temperature observation records (even those with temperature observations present little more than 10% of the time) to form the training dataset, so that the model had the largest spatial coverage possible. A training dataset separation is also useful for the treatment of reservoirs -- separating out basins with and without major reservoirs and train models separately for these two groups would be beneficial. These results indicate our inputs do not fully characterize the stream temperature prediction problem and future improvement efforts could focus on collecting input and observational data. With increasing amounts of data, deep-

learning-powered models can increase accuracy and applicability, offering a plausible pathway toward reliable stream temperature predictions for a wide variety of situations around the world.

Acknowledgments

FR was supported by the Pennsylvania Water Resources Research Center graduate internship G19AC00425. Funding for the internship and AA and SO 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, Shen, et al., 2021). 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

Arambourou, H., & Stoks, R. (2015). Combined effects of larval exposure to a heat wave and chlorpyrifos in northern and southern populations of the damselfly Ischnura elegans. *Chemosphere*, *128*, 148–154. https://doi.org/10.1016/j.chemosphere.2015.01.044

Bowerman, T., Roumasset, A., Keefer, M. L., Sharpe, C. S., & Caudill, C. C. (2018). Prespawn mortality of female chinook salmon increases with water temperature and percent hatchery origin. *Transactions of the American Fisheries Society*, *147*(1), 31–42. https://doi.org/10.1002/tafs.10022

Briggs, M. A., & Hare, D. K. (2018). Explicit consideration of preferential groundwater discharges as surface water ecosystem control points. *Hydrological Processes*, *32*(15), 2435–2440. https://doi.org/10.1002/hyp.13178

Caldwell, S. H., Kelleher, C., Baker, E. A., & Lautz, L. K. (2019). Relative information from thermal infrared imagery via unoccupied aerial vehicle informs simulations and spatially-distributed assessments of stream temperature. *Science of The Total Environment*, *661*, 364–374. https://doi.org/10.1016/j.scitotenv.2018.12.457

Carron, J. C., & Rajaram, H. (2001). Impact of variable reservoir releases on management of downstream water temperatures. *Water Resources Research*, *37*(6), 1733–1743. https://doi.org/10.1029/2000wr900390

DeWeber, J. T., & Wagner, T. (2014). A regional neural network ensemble for predicting mean daily river water temperature. *Journal of Hydrology*, *517*, 187–200. https://doi.org/10.1016/j.jhydrol.2014.05.035

Dugdale, S. J., Kelleher, C. A., Malcolm, I. A., Caldwell, S., & Hannah, D. M. (2019). Assessing the potential of drone-based thermal infrared imagery for quantifying river temperature heterogeneity. *Hydrological Processes*, *33*(7), 1152–1163. https://doi.org/10.1002/hyp.13395

Essaid, H. I., & Caldwell, R. R. (2017). Evaluating the impact of irrigation on surface water – groundwater interaction and stream temperature in an agricultural watershed. *Science of The Total Environment*, *599–600*, 581–596. https://doi.org/10.1016/j.scitotenv.2017.04.205

Falcone, J. A. (2011). *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow* [Report]. USGS Publications Warehouse. https://doi.org/10.3133/70046617

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2021). The data synergy effects of time-series deep learning models in hydrology. *ArXiv:2101.01876 [Cs, Stat]*. http://arxiv.org/abs/2101.01876

Fang, K., Kifer, D., Lawson, K., & Shen, C. (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. *Water Resources Research*, *56*(12), e2020WR028095. https://doi.org/10.1029/2020wr028095

Fang, K., & Shen, C. (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, K., Shen, C., Kifer, D., & Yang, X. (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, D., Fang, K., & Shen, C. (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

Feng, D., Lawson, K., & Shen, C. (2021a). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data.

Geophysical Research Letters, *48*(14), e2021GL092999. https://doi.org/10.1029/2021GL092999

Feng, D., Lawson, K., & Shen, C. (2021b). Prediction in ungauged regions with sparse flow duration curves and input-selection ensemble modeling. *ArXiv*. http://arxiv.org/abs/2011.13380

Ficklin, D. L., Luo, Y., Stewart, I. T., & Maurer, E. P. (2012). Development and application of a hydroclimatological stream temperature model within the Soil and Water Assessment Tool. *Water Resources Research*, *48*(1), W01511. https://doi.org/10/c984xf

Fry, F. E. J. (1971). 1—The Effect of Environmental Factors on the Physiology of Fish. In W. S. Hoar & D. J. Randall (Eds.), *Fish Physiology* (Vol. 6, pp. 1–98). Academic Press. https://doi.org/10.1016/S1546-5098(08)60146-6

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 1050–1059.

Gallice, A., Schaefli, B., Lehning, M., Parlange, M. B., & Huwald, H. (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), 3727–3753. https://doi.org/10.5194/hess-19-3727-2015

Gardner, B., & Sullivan, P. J. (2004). Spatial and temporal stream temperature prediction: Modeling nonstationary temporal covariance structures. *Water Resources Research*, *40*(1). https://doi.org/10.1029/2003wr002511

Gjorgiev, B., & Sansavini, G. (2018). Electrical power generation under policy constrained water-energy nexus. *Applied Energy*, *210*, 568–579. https://doi.org/10.1016/j.apenergy.2017.09.011

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*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Graf, R., Zhu, S., & Sivakumar, B. (2019). Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach. *Journal of Hydrology*, *578*, 124115. https://doi.org/10.1016/j.jhydrol.2019.124115

Hare, D. K., Helton, A. M., Johnson, Z. C., Lane, J. W., & Briggs, M. A. (2021). Continental-scale analysis of shallow and deep groundwater contributions to streams. *Nature Communications*, *12*(1), 1450. https://doi.org/10.1038/s41467-021-21651-0

Hill, R. A., & Hawkins, C. P. (2014). Using modelled stream temperatures to predict macro-spatial patterns of stream invertebrate biodiversity. *Freshwater Biology*, *59*(12), 2632–2644. https://doi.org/10.1111/fwb.12459

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10/bxd65w

International Rivers. (2007). *Damming Statistics*. International Rivers. https://archive.internationalrivers.org/damming-statistics

Isaak, D. J., Wenger, S. J., Peterson, E. E., Hoef, J. M. V., Nagel, D. E., Luce, C. H., Hostetler, S. W., Dunham, J. B., Roper, B. B., Wollrab, S. P., Chandler, G. L., Horan, D. L., & Parkes-Payne, S. (2017). The NorWeST summer stream temperature model and scenarios for the Western U.S.: A crowd-sourced database and new geospatial tools foster a user community and predict broad climate warming of rivers and streams. *Water Resources Research*, *53*(11), 9181–9205. https://doi.org/10.1002/2017wr020969

Jackson, F. L., Fryer, R. J., Hannah, D. M., Millar, C. P., & Malcolm, I. A. (2018). A spatio-temporal statistical model of maximum daily river temperatures to inform the management of Scotland's Atlantic salmon rivers under climate change. *Science of The Total Environment*, *612*, 1543–1558. https://doi.org/10.1016/j.scitotenv.2017.09.010

Jia, X., Zwart, J., Sadler, J., Appling, A., Oliver, S., Markstrom, S., Willard, J., Xu, S., Steinbach, M., Read, J., & Kumar, V. (2020). Physics-Guided Recurrent Graph Networks for Predicting Flow and Temperature in River Networks. *ArXiv:2009.12575 [Physics]*. http://arxiv.org/abs/2009.12575

Johnson, Z. C., Snyder, C. D., & Hitt, N. P. (2017). Landform features and seasonal precipitation predict shallow groundwater influence on temperature in headwater streams. *Water Resources Research*, *53*(7), 5788–5812. https://doi.org/10.1002/2017WR020455

Justice, C., White, S. M., McCullough, D. A., Graves, D. S., & Blanchard, M. R. (2017). Can stream and riparian restoration offset climate change impacts to salmon populations? *Journal of Environmental Management*, *188*, 212–227. https://doi.org/10.1016/j.jenvman.2016.12.005

Karpatne, A., Watkins, W., Read, J., & Kumar, V. (2018). Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. *ArXiv:1710.11431 [Physics, Stat]*. http://arxiv.org/abs/1710.11431

Kędra, M., & Wiejaczka, Ł. (2018). Climatic and dam-induced impacts on river water temperature: Assessment and management implications. *Science of The Total Environment*, *626*, 1474–1483. https://doi.org/10.1016/j.scitotenv.2017.10.044

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Benchmarking a catchment-aware Long Short-Term Memory network (LSTM) for large-scale hydrological modeling. *Hydrology and Earth System Sciences Discussions*, 1–32. https://doi.org/10/ggj67p

Liu, L., Hejazi, M., Li, H., Forman, B., & Zhang, X. (2017). Vulnerability of US thermoelectric power generation to climate change when incorporating state-level environmental regulations. *Nature Energy*, *2*(8), 17109. https://doi.org/10.1038/nenergy.2017.109

Ma, J., Li, C., Liu, F., Wang, Y., Liu, T., & Feng, X. (2018). Optimization of circulating cooling water networks considering the constraint of return water temperature. *Journal of Cleaner Production*, *199*, 916–922. https://doi.org/10.1016/j.jclepro.2018.07.239

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., & Shen, C. (2021). Transferring hydrologic data across continents – Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5), e2020WR028600. https://doi.org/10.1029/2020wr028600

Marcogliese, D. J. (2001). Implications of climate change for parasitism of animals in the aquatic environment. *Canadian Journal of Zoology*, *79*(8), 1331–1352. https://doi.org/10.1139/z01-067

Martí-Cardona, B., Prats, J., & Niclòs, R. (2019). Enhancing the retrieval of stream surface temperature from Landsat data. *Remote Sensing of Environment*, *224*, 182–191. https://doi.org/10.1016/j.rse.2019.02.007

Martins, E. G., Hinch, S. G., Patterson, D. A., Hague, M. J., Cooke, S. J., Miller, K. M., Robichaud, D., English, K. K., & Farrell, A. P. (2012). *High river temperature reduces survival of sockeye salmon (Oncorhynchus nerka) approaching spawning grounds and exacerbates female mortality*. https://doi.org/10.1139/f2011-154

McNyset, K. M., Volk, C. J., & Jordan, C. E. (2015). Developing an effective model for predicting spatially and temporally continuous stream temperatures from remotely sensed land surface temperatures. *Water*, *7*(12), 6827–6846. https://doi.org/10.3390/w7126660

Mohseni, O., & Stefan, H. G. (1999). Stream temperature/air temperature relationship: A physical interpretation. *Journal of Hydrology*, *218*(3), 128–141. https://doi.org/10.1016/S0022-1694(99)00034-7

Nash, J. E., & Sutcliffe, J. V. (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

National Atlas of the United States, USGS. (2009). *Major Dams of the United States* [Data set]. https://web.archive.org/web/20090814080910/http://nationalatlas.gov/mld/dams00x.html

O'Sullivan, A. M., Devito, K. J., & Curry, R. A. (2019). The influence of landscape characteristics on the spatial variability of river temperatures. *CATENA*, *177*, 70–83. https://doi.org/10.1016/j.catena.2019.02.006

O'Sullivan, A. M., Devito, K. J., Ogilvie, J., Linnansaari, T., Pronk, T., Allard, S., & Curry, R. A. (2020). Effects of topographic resolution and geologic setting on spatial statistical river temperature models. *Water Resources Research*, *56*(12), e2020WR028122. https://doi.org/10.1029/2020WR028122

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., & Shen, C. (2021). Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *Journal of Hydrology*, *599*, 126455. https://doi.org/10.1016/j.jhydrol.2021.126455

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*, *16*(2), 024025. https://doi.org/10.1088/1748-9326/abd501

Rahmani, F., Shen, C., Oliver, S. K., Lawson, K., Watkins, W. D., & Appling, A. P. (2021). *Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins: U.S. Geological Survey data release*. U.S. Geological Survey. https://doi.org/10.5066/P9VHMO56

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., Karpatne, A., Hansen, G. J., Hanson, P. C., Watkins, W., & others. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, *55*(11), 9173–9190. https://doi.org/10.1029/2019wr024922

Risley, J. C., Constantz, J., Essaid, H., & Rounds, S. (2010). Effects of upstream dams versus groundwater pumping on stream temperature under varying climate conditions. *Water Resources Research*, *46*(6). https://doi.org/10.1029/2009wr008587

Schaller, M. F., & Fan, Y. (2009). River basins as groundwater exporters and importers: Implications for water cycle and climate modeling. *Journal of Geophysical Research*, *114*(D4). https://doi.org/10/bg3nv2

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10/gd8cqb

Stefan, H. G., & Preud'homme, E. B. (1993). Stream temperature estimation from air temperature. *JAWRA Journal of the American Water Resources Association*, *29*(1), 27–45. https://doi.org/10.1111/j.1752-1688.1993.tb01502.x

Stewart, J. S., Westenbroek, S. M., Mitro, M. G., Lyons, J. D., Kammel, L. E., & Buchwald, C. A. (2015). *A model for evaluating stream temperature response to climate change in Wisconsin* (Scientific Investigations Report 2014–5186; p. 64). U.S. Geological Survey. http://dx.doi.org/10.3133/sir20145186

Tao, Y., Wang, Y., Rhoads, B., Wang, D., Ni, L., & Wu, J. (2020). Quantifying the impacts of the Three Gorges Reservoir on water temperature in the middle reach of the Yangtze River. *Journal of Hydrology*, *582*, 124476. https://doi.org/10.1016/j.jhydrol.2019.124476

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wei, Y., Devarakonda, R., Vose, R. S., & Cook, R. B. (2016). *Daymet: Daily surface weather data on a 1-km grid for North America, version 3*. ORNL Distributed Active Archive Center. https://doi.org/10.3334/ORNLDAAC/1328

Tsai, W.-P., Pan, M., Lawson, K., Liu, J., Feng, D., & Shen, C. (2020). From parameter calibration to parameter learning: Revolutionizing large-scale geoscientific modeling with big data. *ArXiv:2007.15751 [Preprint]*. http://arxiv.org/abs/2007.15751

US Army Corps of Engineers. (2018). *National Inventory of Dams (NID)* [Data set]. https://nid.sec.usace.army.mil/

USGS. (2016). *USGS Surface-Water Data for the Nation*. http://waterdata.usgs.gov/nwis/sw

Weber, M., Rinke, K., Hipsey, M. R., & Boehrer, B. (2017). Optimizing withdrawal from drinking water reservoirs to reduce downstream temperature pollution and reservoir hypoxia. *Journal of Environmental Management*, *197*, 96–105. https://doi.org/10.1016/j.jenvman.2017.03.020

Westenbroek, S., Stewart, J. S., Buchwald, C. A., Mitro, M., Lyons, J. D., & Greb, S. (2010). A model for evaluating stream temperature response to climate change scenarios in wisconsin. *Watershed Management 2010*, 1–12. https://doi.org/10.1061/41143(394)1

Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resources Research*, *56*(1), e2019WR025326. https://doi.org/10.1029/2019WR025326

Younus, M., Hondzo, M., & Engel, B. A. (2000). Stream temperature dynamics in upland agricultural watersheds. *Journal of Environmental Engineering*, *126*(6), 518–526. https://doi.org/10.1061/(ASCE)0733-9372(2000)126:6(518)

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., & Li, L. (2021). From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environmental Science & Technology*, *55*(4), 2357–2368. https://doi.org/10.1021/acs.est.0c06783

Zhu, S., & Piotrowski, A. P. (2020). River/stream water temperature forecasting using artificial intelligence models: A systematic review. *Acta Geophysica*, *68*(5), 1433–1442. https://doi.org/10.1007/s11600-020-00480-7

Table 1. Data Availability Groups.

Percentage of data availability	Least number of days observed T_s available (train)	Least number of days observed T_s available (test)	Number of Sites	Sites without major dams	Sites with major dams	No. observed data (thousand) [percentage of total sample]	Spring data (%)	Summer data (%)	Fall data (%)	Winter data (%)
p>99%	1445	723	99	34	65	216 [28]	25.2	25.1	24.9	24.7
99%>p>60%	876	438	207	84	123	413 [54]	25.5	25.6	24.8	23.9
p>60%	876	438	306	118	188	630 [82]	25.4	25.5	24.9	24.2
60%>p>10%	146	73	109	43	66	131 [17]	25.7	30.5	25.1	18.6
p>10%	146	73	415	161	254	761 [99]	25.5	26.4	24.9	23.2
p>10% (PUB)	0	10	40	17	23	5.8 [0.7]	26.3	50.6	19.9	3.1
p>60% (PUB)	0	438	40	18	22	81.2 [10]	25.6	25.5	24.4	24.5

Table 2. Median RMSE in different training and testing sets from the input-selection ensemble model for ungauged basins and from regular full-attribute model for gauged basins. Going to the right side of the table, the training set becomes broader. Going down the table, the test set becomes larger. To the right of the diagonal, we are training on a larger set than the test set. To the left of the diagonal, we are training on a small set and extrapolate the model to test basins. The underlined cells are input-selection ensemble results. Bold numbers are the best results achieved in the testing experiments.

	train p>99%	train p>60%	train p>10%
test p>99%	**0.801**	0.804	0.878
test 99%>p>60%	1.887	**0.830**	0.877
test 60%>p>10%	2.053	1.559	**0.916**
test p<10% (PUB)	2.911	1.556	**1.536**
Test p>60% (PUB)	1.696	1.162	**1.129**

Figure 1. Data availability groups (DAGs) and other basin categories. (a) DAGs are nested (i.e. all sites in $DAG_{p>99}$ are also contained in $DAG_{p>60}$ and $DAG_{p>10}$). These DAGs should not be confused with the separate descriptors of (b) basins with data availability between 60% and 10% and (c) basins with data availability between 99% and 60%, which are used to discuss model results.

Figure 2. CONUS-scale aggregated metrics of stream temperature models individually trained on each data availability group. Each boxplot shows the distribution of that metric over all sites in the relevant test set using (a) the matching-DAG approach; and (b) the maximum-site approach. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker represent 5%, 25%, 50%, 75% and 95% of data, respectively.

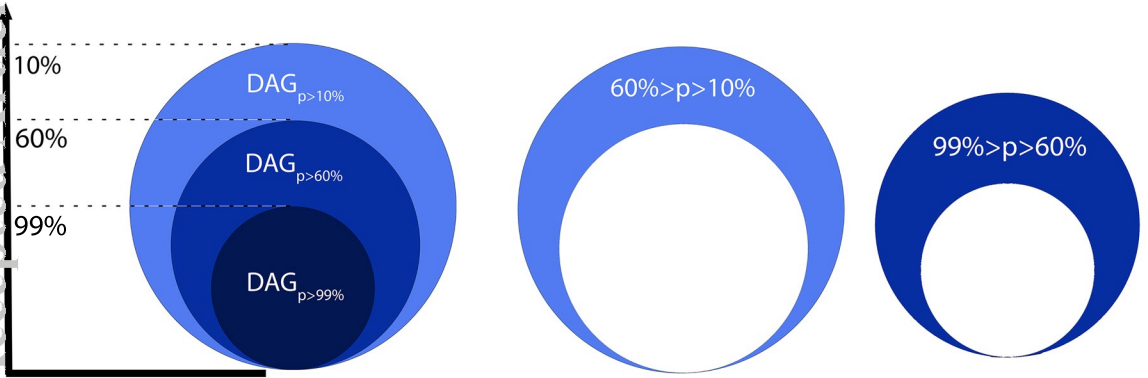
Figure 3. Map of (a) RMSE, (b) NSE, and (c) R^2 values for $LSTM_{p>10}$, which is the model trained on all sites with p>10%. The size of the symbol represents data availability, while the shape (square or circle) indicates with or without major dams, respectively. The blue box in the northwest of the map shown in (c) is the latitude-longitude box used to compare with Gallice et al.(2015).

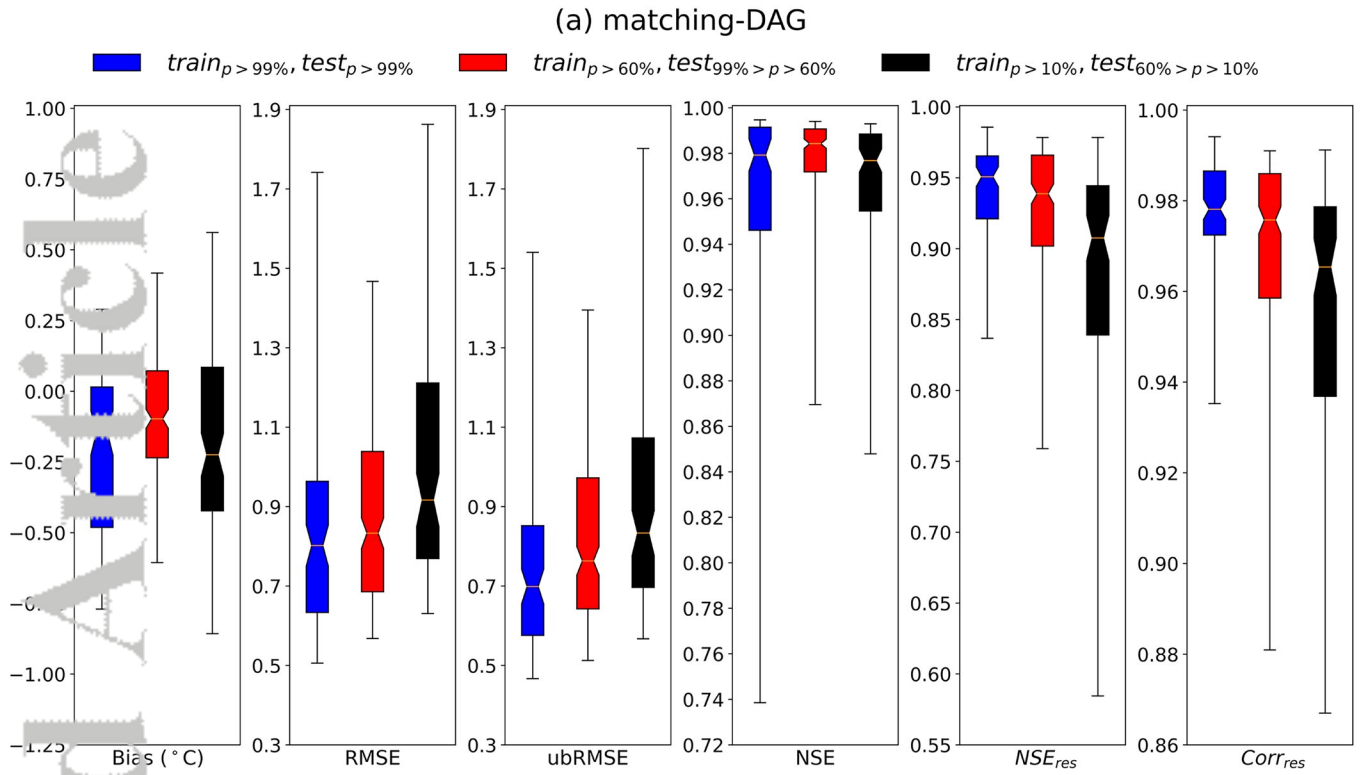
Figure 4. Results from the prediction in unmonitored basins (PUB) tests from different training data (different data availability groups (DAGs)) and different input attributes (full-attribute vs. the input-selection ensemble). (a) 40 random holdout basins with p>60%. Note that here, the number of basins in $DAG_{p>10\%}$, $DAG_{p>60\%}$, *and* $DAG_{p>99\%}$ *are 375, 266, and 85, respectively, which are different from the number of basins in three DAGs in experiments in Figures 2 and 4b. (b) 40 unmonitored basins with p<10%. The lower box edge, center bar, and upper box edge represent 25%, 50%, and 75% of data, respectively. However, lower and upper whiskers' lengths are not greater than 1.5 times of the interquartile range.*

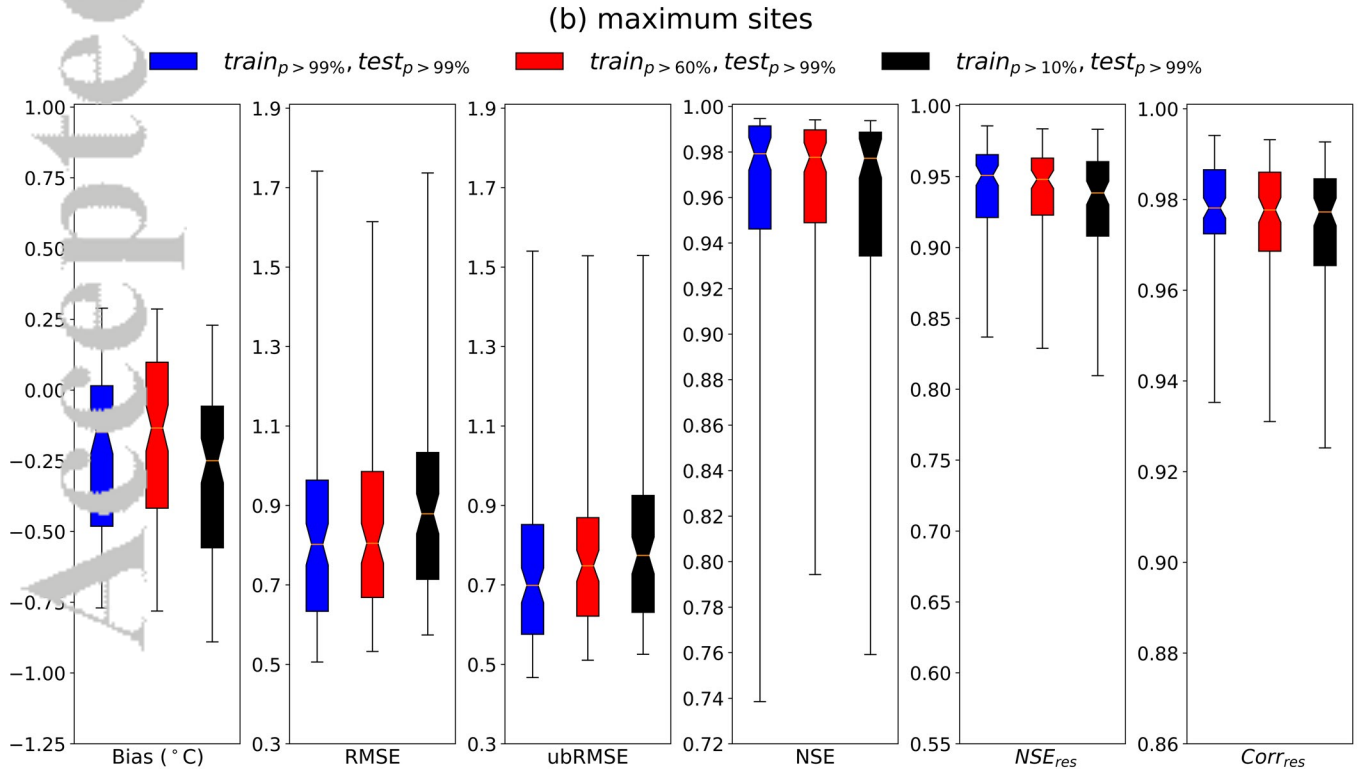
Figure 5. Seasonality plot in temporal prediction for dammed and undammed basins. The model is $LSTM_{p>10\%}$ and is trained with both dammed and undammed basins. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker represent 5%, 25%, 50%, 75% and 95% of data, respectively.

Figure 6. Time series plots of observed and simulated T in the test period for temporal prediction (trained with $DAG_{p>10\%}$) (a-d) and for spatial generalization ($PUB_{test\_p>60\%}$ testing) (e-f). a, e, and f show a positive bias, while the rest show the more common negative bias. Observed (obs) stream temperature data from USGS (USGS, 2016)

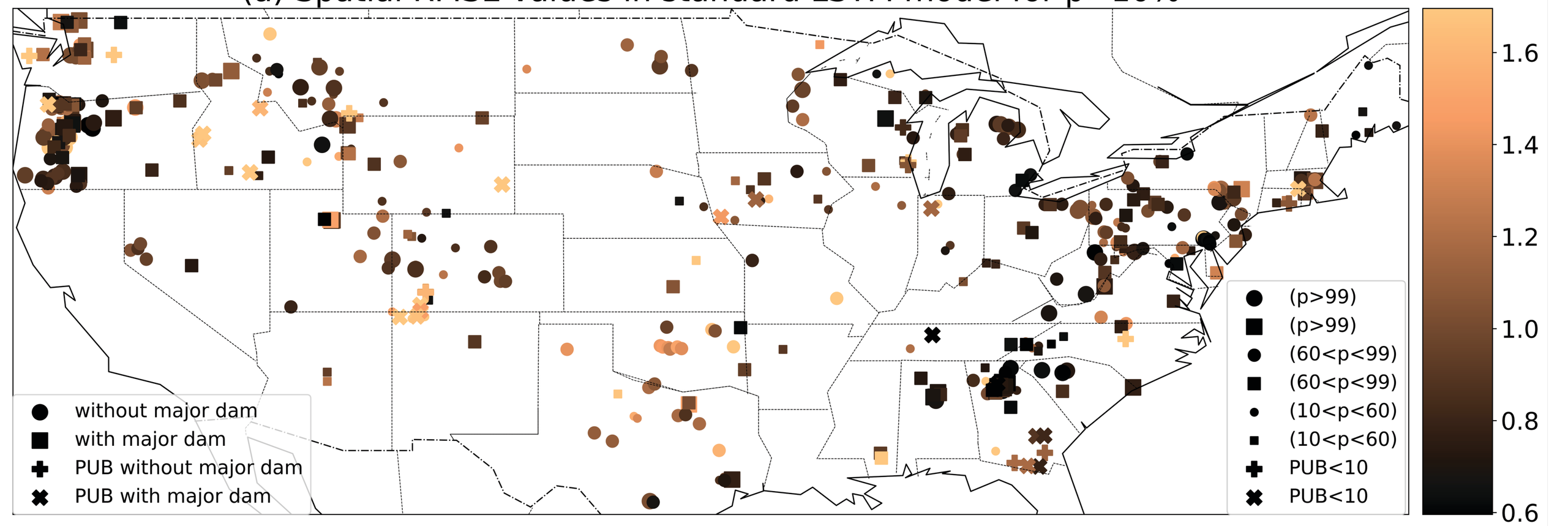
Figure 7. CONUS-scale aggregated metrics of the stream temperature models individually trained on data availability groups also split into natural and unnatural basins. This is essentially a repetition of the first temporal prediction (Figure 2a), except that here, models were trained and tested on basins either with or without major dams present, not both. For example, ">99%, with dam" means that both the training and testing sets only contained basins with observations available more than 99% of the time, and also had at least one major dam. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker represent 5%, 25%, 50%, 75% and 95% of data, respectively.



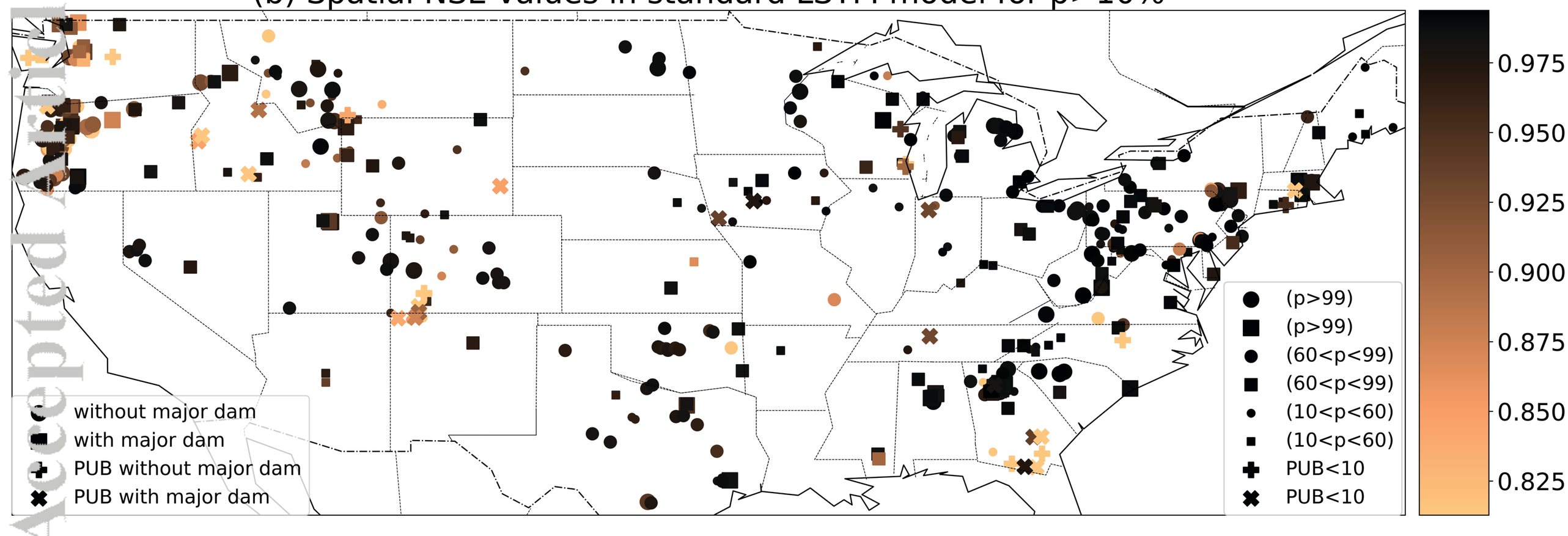




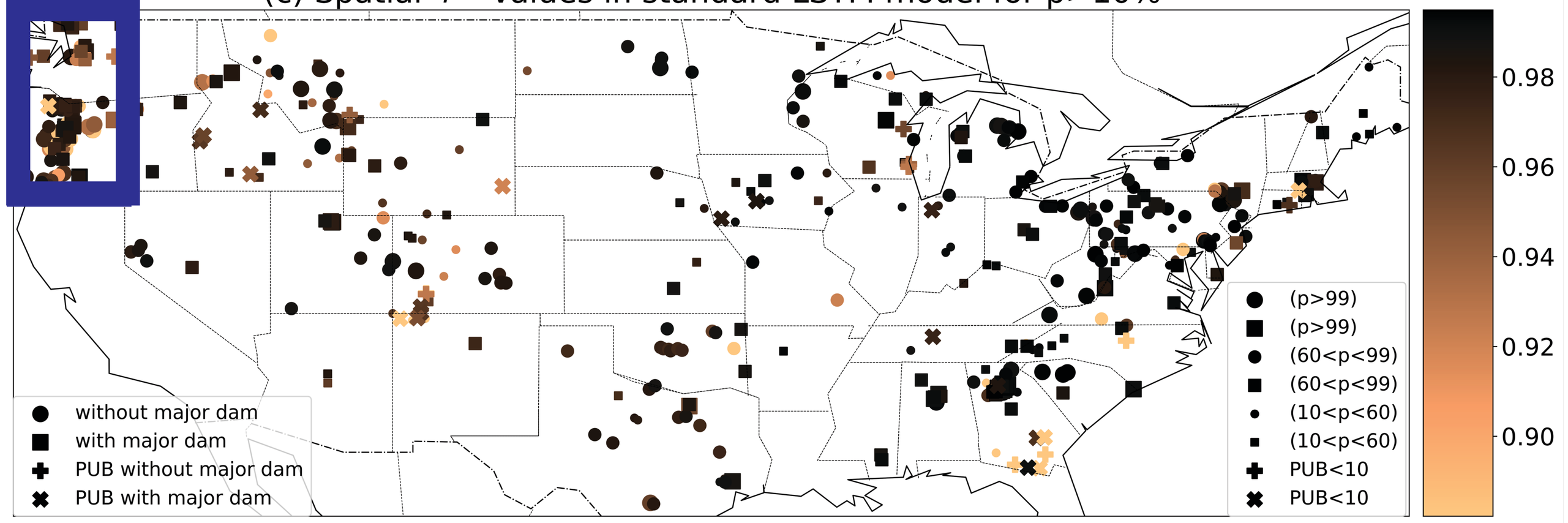
(a) Spatial RMSE values in standard LSTM model for p>10%



(b) Spatial NSE values in standard LSTM model for p>10%



(c) Spatial r^2 values in standard LSTM model for p>10%



PUB tests for Full Attributes versus Input selection ensemble

(a)Random holdout

> 99%-Full Attr, $test_{PUB>60\%}$

Models> 99%-Ensemble, $test_{PUB>60\%}$

> 60%-Full Attr, $test_{PUB>60\%}$

Models> 60%-Ensemble, $test_{PUB>60\%}$

> 10%-Full Attr, $test_{PUB>60\%}$

Models> 10%-Ensemble, $test_{PUB>60\%}$

Bias (° C) RMSE ubRMSE NSE NSE_{res} $Corr_{res}$ r^2

(b)p<10% holdout

> 99%-Full Attr, $test_{PUB<10\%}$

Models> 99%-Ensemble, $test_{PUB<10\%}$

> 60%-Full Attr, $test_{PUB<10\%}$

Models> 60%-Ensemble, $test_{PUB<10\%}$

> 10%-Full Attr, $test_{PUB<10\%}$

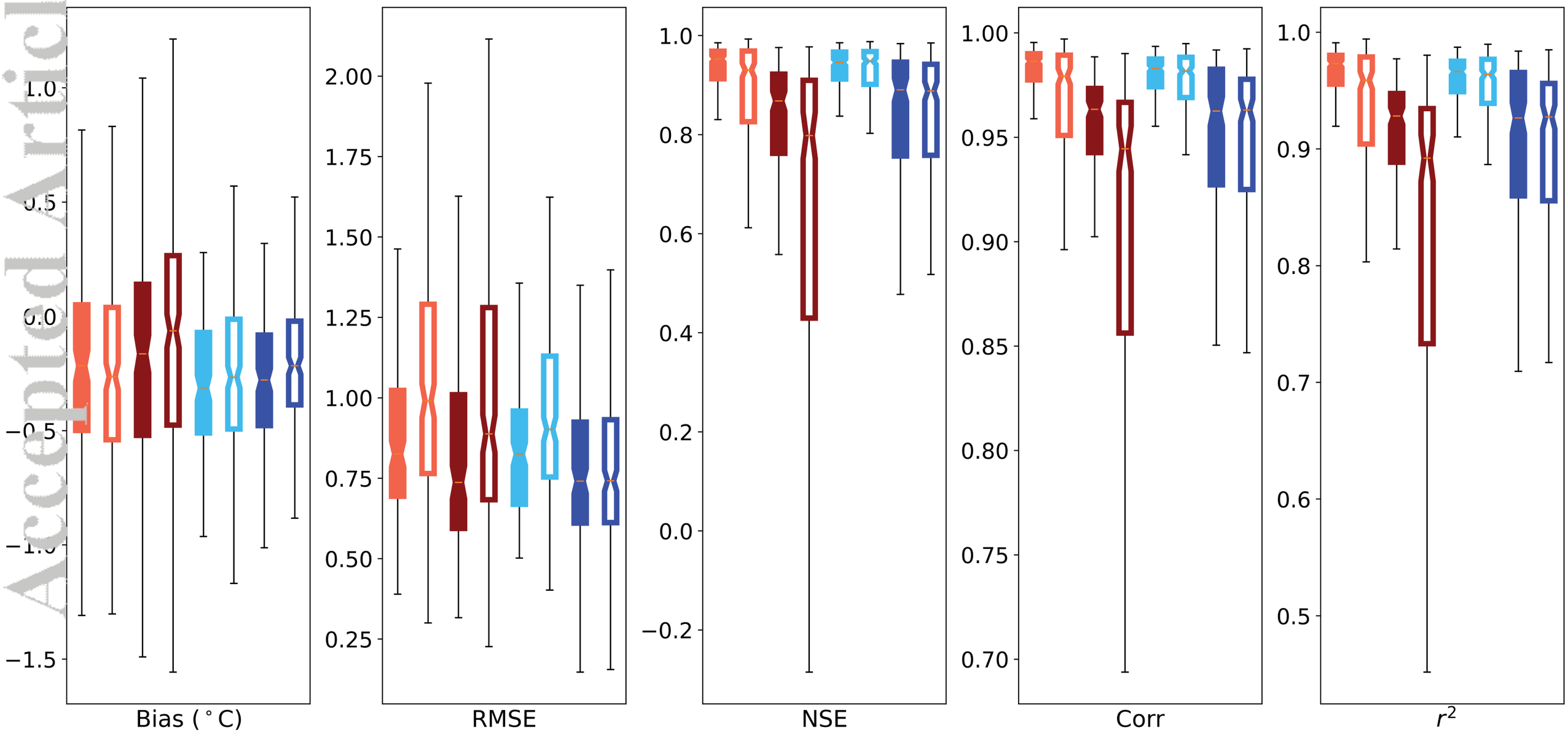
Models> 10%-Ensemble, $test_{PUB<10\%}$

Bias (° C) RMSE ubRMSE NSE NSE_{res} $Corr_{res}$ r^2

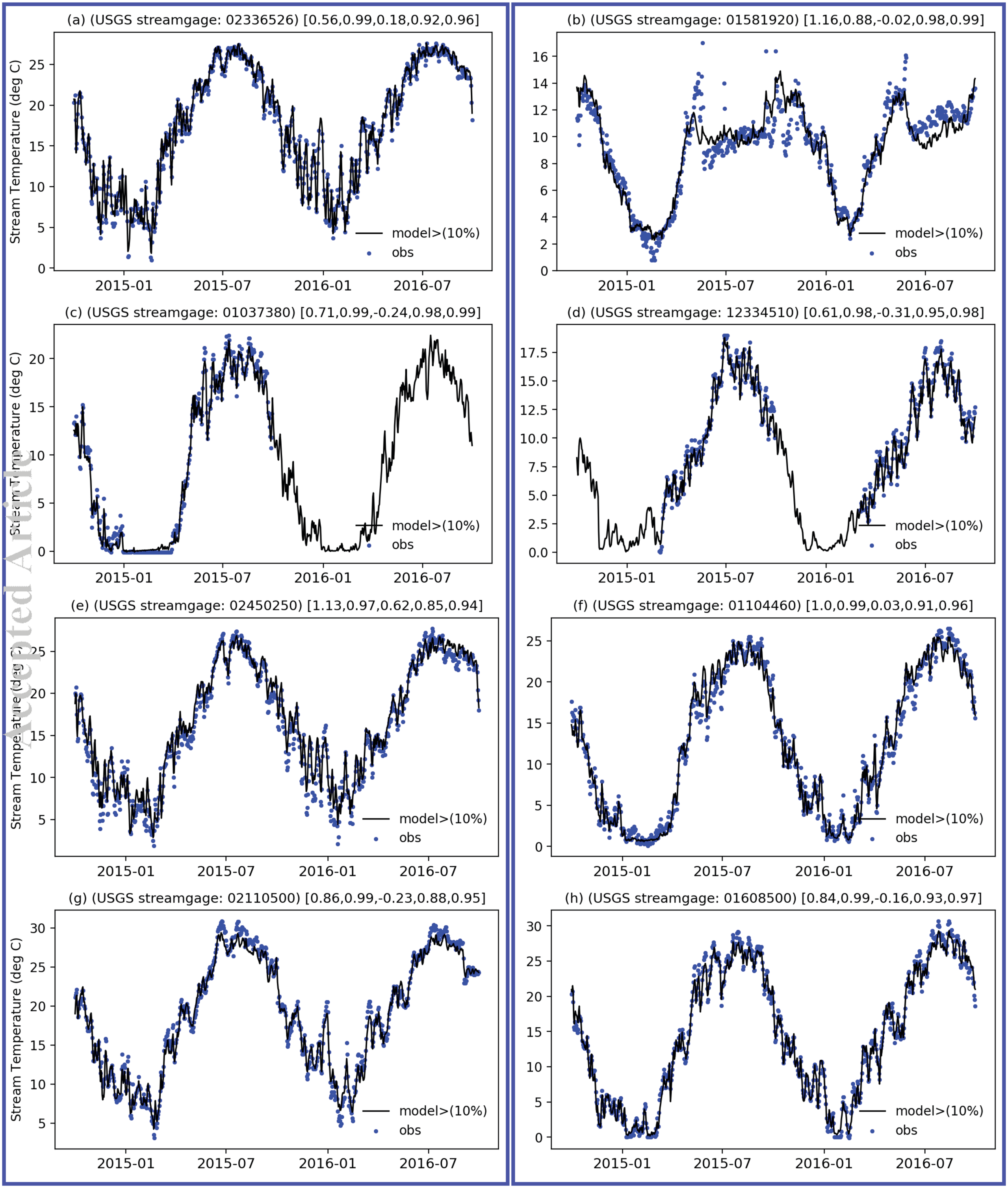
$Spring_{NoDam,\,p>10\%}$ $Summer_{NoDam,\,p>10\%}$ $Fall_{NoDam,\,p>10\%}$ $Winter_{NoDam,\,p>10\%}$

$Spring_{Dammed,\,p>10\%}$ $Summer_{Dammed,\,p>10\%}$ $Fall_{Dammed,\,p>10\%}$ $Winter_{Dammed,\,p>10\%}$

Seasonality comparison between undammed and dammed basins in $LSTM_{p>10\%}$

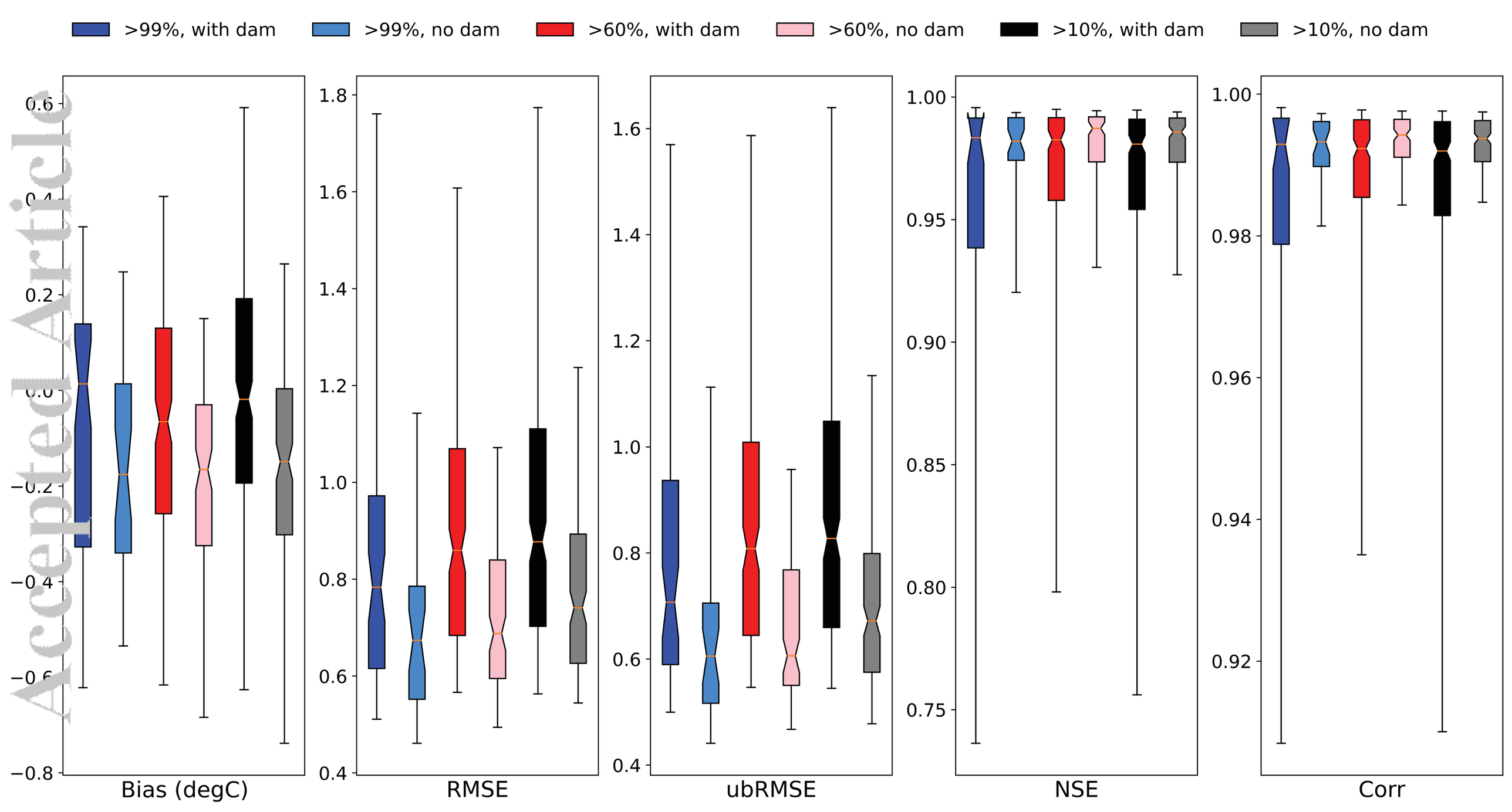


Time Series simulated and observed data in testing period- values in brackets are [RMSE, NSE, Bias, NSE_res, Corr_res]

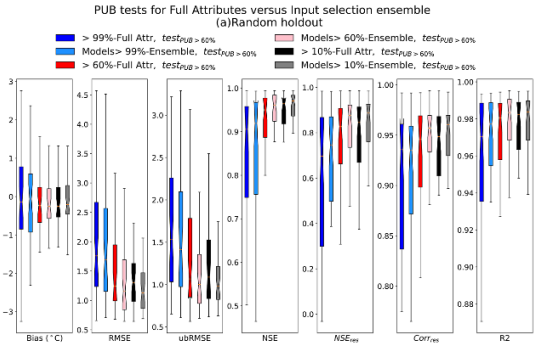


without major dam

with major dam



LSTM presented state-of-the-art stream temperature prediction performance in both dammed and unmonitored basins. Known input attributes do not cover all necessary features so an input-selection ensemble is useful. For temporal prediction, the most suitable training set was the matching data availability group (DAG) that the basin could be grouped into, However, for spatial extrapolation (unmonitored basins), a training dataset including all basins with data is consistently preferred.



Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins

Farshid Rahmani, Chaopeng Shen*, Samantha Oliver, Kathryn Lawson and Alison Appling*