Hydrology and
Earth System
Sciences

Technical note: Using long short-term memory models to fill data gaps in hydrological monitoring networks

Huiying Ren[1], Erol Cromwell[2], Ben Kravitz[3,4], and Xingyuan Chen[4]

[1]Earth Systems Science Division, Pacific Northwest National Laboratory, Richland, WA, USA
[2]Advanced Computing, Mathematics, and Data Division, Pacific Northwest National Laboratory, Richland, WA, USA
[3]Department of Earth and Atmospheric Sciences, Indiana University, Bloomington, IN, USA
[4]Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA, USA

Correspondence: Xingyuan Chen (xingyuan.chen@pnnl.gov)

Abstract. Quantifying the spatiotemporal dynamics in subsurface hydrological flows over a long time window usually employs a network of monitoring wells. However, such observations are often spatially sparse with potential temporal gaps due to poor quality or instrument failure. In this study, we explore the ability of recurrent neural networks to fill gaps in a spatially distributed time-series dataset. We use a well network that monitors the dynamic and heterogeneous hydrologic exchanges between the Columbia River and its adjacent groundwater aquifer at the U.S. Department of Energy's Hanford site. This 10-year-long dataset contains hourly temperature, specific conductance, and groundwater table elevation measurements from 42 wells with gaps of various lengths. We employ a long short-term memory (LSTM) model to capture the temporal variations in the observed system behaviors needed for gap filling. The performance of the LSTM-based gap-filling method was evaluated against a traditional autoregressive integrated moving average (ARIMA) method in terms of error statistics and accuracy in capturing the temporal patterns of river corridor wells with various dynamics signatures. Our study demonstrates that the ARIMA models yield better average error statistics, although they tend to have larger errors during time windows with abrupt changes or high-frequency (daily and subdaily) variations. The LSTM-based models excel in capturing both high-frequency and low-frequency (monthly and seasonal) dynamics. However, the inclusion of high-frequency fluctuations may also lead to overly dynamic predictions in time windows that lack such fluctuations. The LSTM can take advantage of the spatial information from neighboring wells to improve the gap-filling accuracy, especially for long gaps in system states that vary at subdaily scales. While LSTM models require substantial training data and have limited extrapolation power beyond the conditions represented in the training data, they afford great flexibility to account for the spatial correlations, temporal correlations, and nonlinearity in data without a priori assumptions. Thus, LSTMs provide effective alternatives to fill in data gaps in spatially distributed time-series observations characterized by multiple dominant frequencies of variability, which are essential for advancing our understanding of dynamic complex systems.

1 Introduction

Long-term hydrological monitoring using distributed well networks is of critical importance for understanding how ecosystems respond to chronic or extreme perturbations as well as for informing policies and decisions related to natural resources and environmental issues (Wett et al., 2002; Taylor and Alley, 2002; Grant and Dietrich, 2017). One of the most common methods for collecting hydrologic and chemistry data in groundwater is through wells (Güler and Thyne, 2004; Strobl and Robillard, 2008; Lin et al., 2012). However, most well data have temporal gaps due to instrument failure or poor measurement quality. These data gaps degrade the quality of the dataset and increase the uncertainty in the spatial and temporal patterns derived from the data. Gap filling is essential for developing an understanding of dynamic system behaviors and for use in creating continuous, internally

consistent boundary conditions for numerical models. One outstanding challenge in gap filling is capturing nonstationarity in data.

Various statistical methods have been developed to fill gaps in spatiotemporal datasets, with the autoregressive integrated moving average (ARIMA) method being the most commonly used (Han et al., 2010; Zhang, 2003). For any given spatial location, ARIMA uses temporal autocorrelations to predict unobserved data points in a time series. Spatiotemporal autocorrelations can be considered by using multivariate ARIMA and space–time autoregressive models (Kamarianakis and Prastacos, 2003; Wikle et al., 1998; Kamarianakis and Prastacos, 2005). However, ARIMA cannot capture nonlinear trends because it assumes a linear dependence between adjacent observations (Faruk, 2010; Valenzuela et al., 2008; Ho et al., 2002). In addition, all existing space–time ARIMA models assume fixed global autoregressive and moving average terms, which fail to capture evolving dynamics in highly dynamic systems (Pfeifer and Deutrch, 1980; Griffith, 2010; Cheng et al., 2012, 2014). Spectral-based methods, such as singular spectrum analysis, the maximum entropy method, and the Lomb–Scargle periodogram, have been used to account for nonlinear trends while filling in gaps in spatiotemporal datasets (Ghil et al., 2002; Hocke and Kämpfer, 2008; Kondrashov and Ghil, 2006). These methods use a few optimal spatial or temporal modes occurring at low frequencies to predict the missing values, with the other higher-frequency components discarded as noise. This can lead to a reduced accuracy of the statistical models when fitting observations and predicting missing values (Kondrashov et al., 2010; Wang et al., 2012). Kriging and maximum likelihood estimation used in spatial and spatiotemporal gap filling often face challenges computing the covariance matrix of the data vector, as it can be quite large (Katzfuss and Cressie, 2012; Eidsvik et al., 2014). Other nonlinear methods have been explored with some success, including expectation–maximization or Bayesian probabilistic inference such as hierarchical models, Gaussian process, and Markov chain Monte Carlo. Using models that build dependencies in different stages or hierarchies most effectively captures spatial and temporal correlations (Calculli et al., 2015; Banerjee et al., 2014; Datta et al., 2016; Finley et al., 2013; Stroud et al., 2017). In general, the expectation–maximization algorithm and Bayesian-based methods are sensitive to the choice of initial values and prior distributions in parameter space (Katzfuss and Cressie, 2011, 2012). Moreover, prior distributions and their associated parameters all need to be specified in both the spatial and temporal domains, which becomes increasingly difficult in more complex systems. Empirical orthogonal function (EOF) related interpolation methods, such as least squares EOF (LSEOF), data interpolation EOF (DINEOF), and recursively subtracted EOF (REEOF), are widely used to fill in missing data such as clouds in sea surface temperature datasets or other satellite-based images with regular gridded

domains (Beckers and Rixen, 2003; Beckers et al., 2006; Alvera-Azcárate et al., 2016). However, EOF methods require gridded data, and this limits their use in filling data gaps in irregularly spaced monitoring networks.

Deep neural networks (DNNs) (Schmidhuber, 2015) are data-driven tools that, in principle, could provide a powerful way of extracting the nonlinear spatiotemporal patterns hidden in distributed time-series data without knowing their explicit forms (Längkvist et al., 2014). They are increasingly being used in geoscience domains to extract patterns and insights from streams of geospatial data and to transform our understanding of complex systems (Reichstein et al., 2019; Shen, 2018; Sun, 2018; Sun et al., 2019; Gentine et al., 2018). The umbrella term of DNN contains numerous architecture categories which can be selected for the problem at hand. Recurrent neural networks (RNNs) are a natural choice of architecture for the analyses in this paper, which focus on filling gaps in time-series data (Jordan, 1986). RNNs take sequences (e.g., time series) as input, and they output either the single values or sequences that follow. They are designed to use information about previous events to make predictions about future events by essentially letting the model "remember." However, RNNs have been shown to lose memory from previously trained data for longer sequences of data, i.e., they "forget" (Hochreiter et al., 2001). The earlier information becomes exponentially less impactful for prediction as the sequence size increases. Long short-term memory (LSTM) networks are variations of RNNs that are explicitly designed to avoid this problem by using memory cells to retain information about relevant past events (Hochreiter and Schmidhuber, 1997). RNNs and LSTMs have been successfully applied to text prediction (Graves, 2013), text translation (Wu et al., 2016), speech recognition (Graves et al., 2013), and image captioning (You et al., 2016). Recently, hydrology applications of RNNs and LSTMs have emerged. For example, Kratzert et al. (2018) used LSTMs to predict watershed runoff from meteorological observations, Zhang et al. (2018) used LSTMs for predicting sewer overflow events from rainfall intensity and sewer water level measurements, and Fang et al. (2017) used LSTMs to predict soil moisture with high fidelity.

Our study aims to evaluate the potential of using LSTM models for filling gaps in spatiotemporal time-series data collected from a distributed network. We tested the LSTM-based gap-filling method using datasets collected to understand the interactions between a regulated river and a contaminated groundwater aquifer. We treat the gap filling as a forecasting problem, i.e., we use the historical data as input to predict the missing values in the data gaps. The future information relative to the gap is implicitly used in training the LSTM models. Framing gap filling as a predictive problem is a common practice when machine learning methods are used for filling gaps in time-series data (see examples in Kandasamy et al., 2013; Körner et al., 2018; Chen et al., 2020; Zhao et al., 2020; Sarafanov et al., 2020; and Contractor and
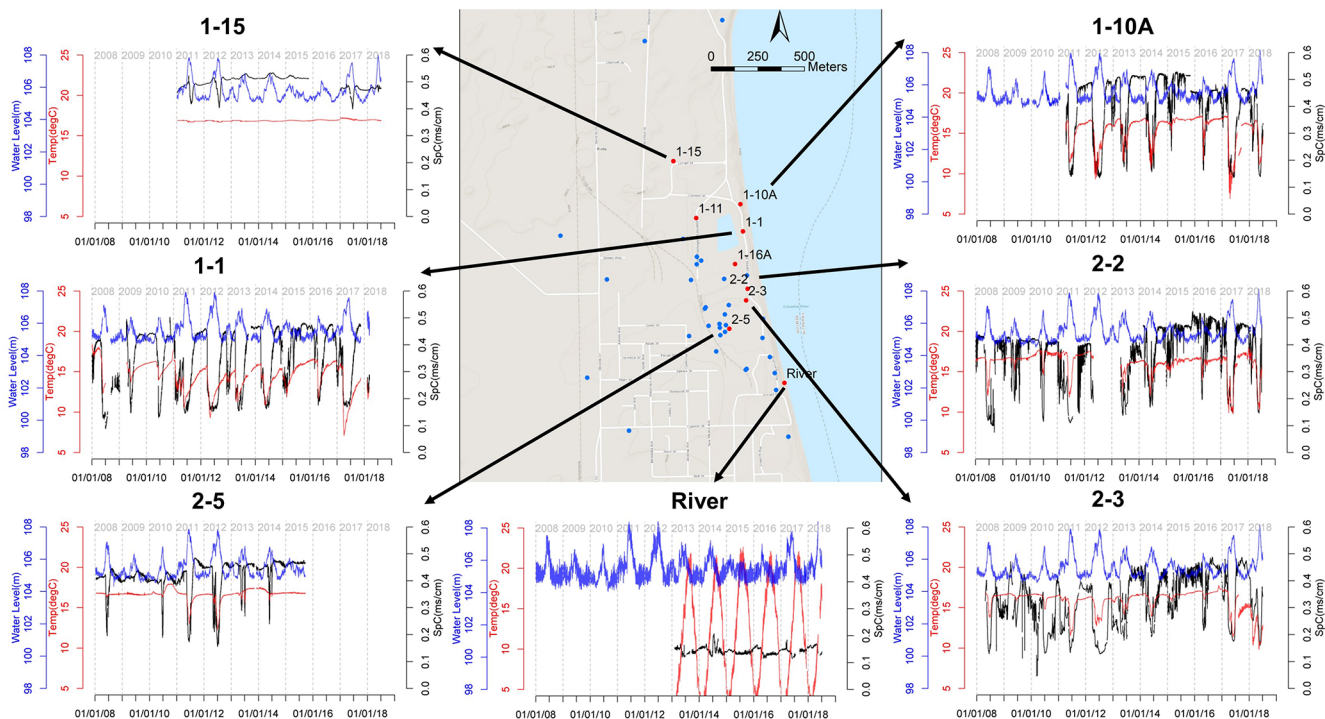
Figure 1. Groundwater monitoring well network in the 300 Area of the Hanford site and the monitoring data at select wells. Each dot represents a well instrumented to measure groundwater elevation, temperature, and specific conductance (SpC). The wells selected for this study are marked with red dots with well names. The three variables monitored are shown in time-series plots using blue (water elevation), black (SpC), and red (temperature) lines. (Base map © Google Maps.)

Roughan, 2021). The performance of the LSTM-based gap-filling method is compared with traditional time-series approaches (i.e., ARIMA) to identify situations in which LSTM models out- or underperform the ARIMA models.

2   Study site and data description

A 10-year (2008–2018) spatiotemporal dataset was collected from a network of groundwater wells that monitor temperature (CS547A water conductivity and temperature probe, Campbell Scientific), specific conductance (SpC) (CS547A water conductivity and temperature probe, Campbell Scientific), and water table elevation (CS451 stainless-steel pressure transducer, Campbell Scientific) at the 300 Area of the U.S. Department of Energy Hanford site, located in southeastern Washington State. The groundwater well network was originally built to monitor the attenuation of legacy contaminants. The groundwater aquifer at the study site is composed of two distinct geologic formations: a highly permeable formation (Hanford formation, consisting of coarse gravelly sand and sandy gravel) underlain by a much less permeable formation (the Ringold Formation, consisting of silt and fine sand). The dominant hydrogeologic features of the aquifer are defined by the interface between the Hanford

and Ringold formations as well as the heterogeneity within the Hanford formation (Chen et al., 2012, 2013).

The intrusion of river water into the adjacent groundwater aquifer causes two water bodies with distinct geochemistry to mix and stimulates biogeochemical reactions at the interface. The river water has lower SpC (0.1–$0.12\,\mathrm{m\,cm^{-1}}$) than the groundwater (averaging $\sim 0.4\,\mathrm{m\,cm^{-1}}$). Groundwater has a nearly constant temperature (16–$17\,°\mathrm{C}$) as opposed to the seasonal variations in the river temperature (3–$22\,°\mathrm{C}$). The highly heterogeneous coarse-textured aquifer (Zachara et al., 2013) interacts with dynamic river stages to create complex river intrusion and retreat pathways and dynamics. The time series of multiyear SpC and temperature observations at the selected wells show these complicated processes of river water intrusion into our study site (Fig. 1). Wells near the river shoreline (e.g., wells 1-1, 1-10A, 2-2, and 2-3) tend to be strongly affected by river water intrusion in spring and summer. As such, the dynamic patterns of SpC and temperature correspond well to river stage fluctuations, specifically that SpC decreases and temperature increases with increasing river stage. Fluctuations in SpC in well 2-2 appear to be stronger and at higher frequency than in other wells, likely indicating its higher connectivity with the river. For wells further inland (e.g., well 1-15), temperatures consistently remain within the groundwater temperature range, and SpC has three noticeable dips (dropping from

0.5 to $0.4\,\mathrm{mS\,cm^{-1}}$). The SpC dips coincide with the high river stages in the years 2011, 2012, and 2017, which featured higher peak river stages than other years that enabled river water to intrude further into the groundwater aquifer. In wells located an intermediate distance from the river, such as well 2-5, the intrusion of river water is evident in most years. It is absent in low-flow years, such as 2009 and 2015, during which both SpC and temperature remain nearly unchanged.

Earlier studies have demonstrated that physical heterogeneity contributes to the different response behaviors of different locations while the river stage dynamics lead to multifrequency dynamics in those responses. Natural climatic forcing drives seasonal and annual variations (Amaranto et al., 2019, 2018), whereas the operations of upstream hydroelectric dams to meet human societal needs primarily induce the higher-frequency (i.e., daily and subdaily) fluctuations (Song et al., 2018). Our system is representative of many dam-regulated gravel-bed rivers across the world, where anthropogenic dam operations have significantly altered the hydrologic exchanges between river water and groundwater as well as the associated thermal and biogeochemical processes (Song et al., 2018; Shuai et al., 2019; Zachara et al., 2020). Note that the multifrequency variations in data characterize the dynamic features of data, which could exist in both short-term and long-term time-series data as a result of short-term or long-term monitoring efforts.

To understand the multifrequency variations in the river water and groundwater mixing in each well at the study site, we perform spectral analysis on multiyear SpC observations at each selected well using a discrete wavelet transform (DWT). The DWT is widely used for time–frequency analysis of time series and relies on a "mother wavelet", which is chosen to be the Morlet wavelet (Grossmann and Morlet, 1984) to deal with the time-varying frequency and amplitude in this site's time-series data (Stockwell et al., 1996; Grinsted et al., 2004). We illustrate the wavelet power spectrum (WPS) in log scale and its normalized global power spectrum (average WPS over the time domain) for the multiyear SpC time series in the first two columns of Fig. 2. Data gaps are shown as blank regions in Fig. 2; examples include early 2009 at well 1-1, the beginning of 2011 at well 1-10A, and the later part of 2012 at well 2-2. The WPS amplitude represents the relative importance of variation at a given frequency compared to the variations at other frequencies across the spectrum. At wells 1-1, 1-10A, 2-3, 2-5, and 2-2, the strongest intensities of the SpC signals appear at half-yearly and yearly frequencies; however, well 1-15 has a different pattern, with most of its high intensities below the 256 h frequency. The averaged WPS further shows the behavior contrast: wells 1-1, 1-10A, 2-3, and 2-5 have a dominant frequency at half a year; well 2-2 has multiple dominant frequencies at daily, monthly, and seasonal scales; and well 1-15 has similar intensities at the half-yearly and hourly scales. Using this information to inform our approach, we hypoth-

esize that gap filling at well 2-2 could be more challenging due to the mixture of dynamics signatures.

As the system dynamics are driven by the river stage, we perform magnitude-squared wavelet coherence analysis via the Morlet wavelet to reveal the dynamic correlations between the SpC and river stage time series (Grinsted et al., 2004; Vacha and Barunik, 2012). Wavelet coherence in the time–frequency domain is plotted in the third column in Fig. 2, and the average coherence is plotted in the fourth column; red points indicate statistically significant values at the 95th percent confidence interval. A larger coherence at a given frequency indicates a stronger correlation at that frequency between the SpC at a well and the river stage. We consider these two variables to be highly correlated when the coherence is larger than 0.7 (shown using colors ranging from green to red in the "Coherence" plots). We found that such high correlations exist at multiple frequencies, from subdaily to daily to yearly, at all wells close to the river (e.g., 1-1, 1-10A, 2-2, and 2-3). Wells farther from the river (e.g., 1-15 and 2-5) have higher correlation regimes shifted towards longer periods at semiannual and annual frequencies and are less persistent in time.

As can be seen in Fig. 2, many of the wells have long data gaps which have unknown effects on our ability to estimate dynamics from the wavelet spectra. As such, gap filling is needed to infer observations and guide our modeling of the underlying system. Figure 3 provides a summary of gap lengths for the overall network of monitoring wells. The majority of the gap lengths of the three monitored variables are less than 50 h. Therefore, our investigations explore the ability of different methods to fill gaps of 1, 6, 12, 24, 48, and 72 h lengths to capture the multifrequency fluctuations using hourly data input.

3   Gap-filling methods

In this section, we describe two methods that we implemented to fill gaps of various lengths in SpC measurements at selected wells: an LSTM model and the traditional ARIMA model. We focused our analyses on filling gaps in SpC due to its importance in revealing river water and groundwater mixing. The same set of analyses can be performed on water level and temperature. An input with M time steps (input window length) is provided to both the LSTM and ARIMA models for predicting the next time step immediately following the input window. For gaps larger than 1 h, the gap-filling models are applied to fill in one missing value at a time. The entire gap is filled by sliding the input window forward hour by hour and treating the gap-filled values of the previous missing hours as observed values.

The input window may contain multiple variables relevant to the prediction from a single well or multiple wells. After experimenting with different sets of input variables (SpC only; SpC and water level; and SpC, water level, and tem-
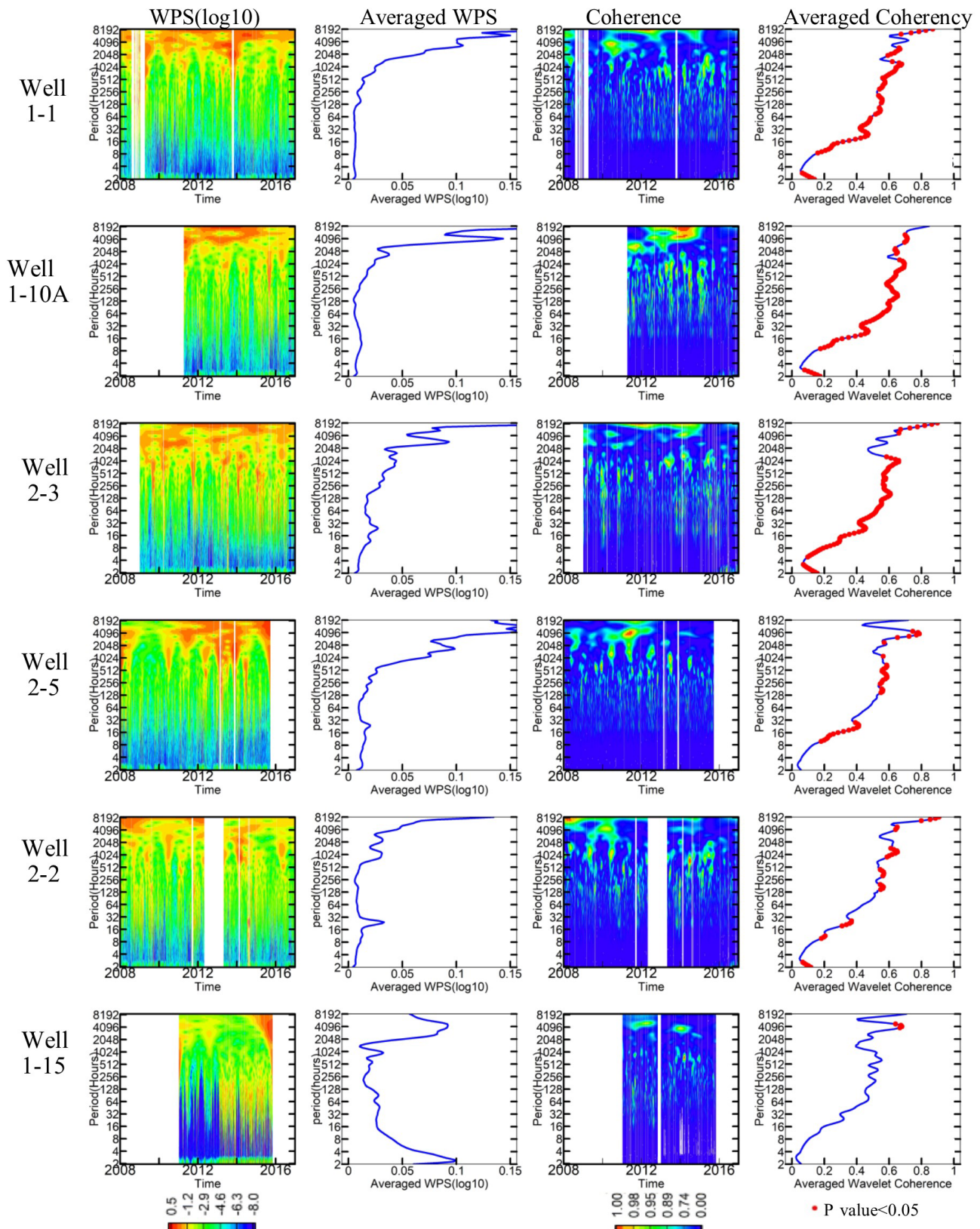
Figure 2. Wavelet power spectrum (WPS) analysis of SpC at each well from 2008 to 2018. The first column is the spectrogram (in log 10 scale) of SpC in each well; the second column is the averaged WPS; the third column is the coherence between SpC in each well and the river stage; and the fourth column is the averaged coherence, with $p < 0.05$ values indicated in red.
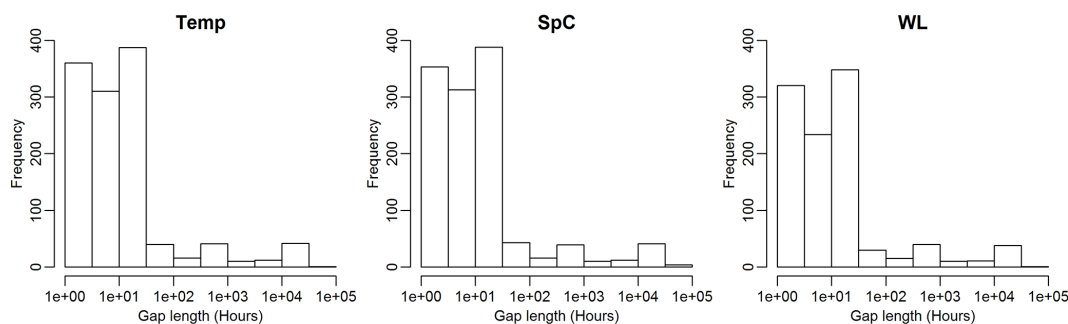
Figure 3. Histograms of gap lengths for each monitored variable, aggregated across all wells in the monitoring network from 2008 to 2018.
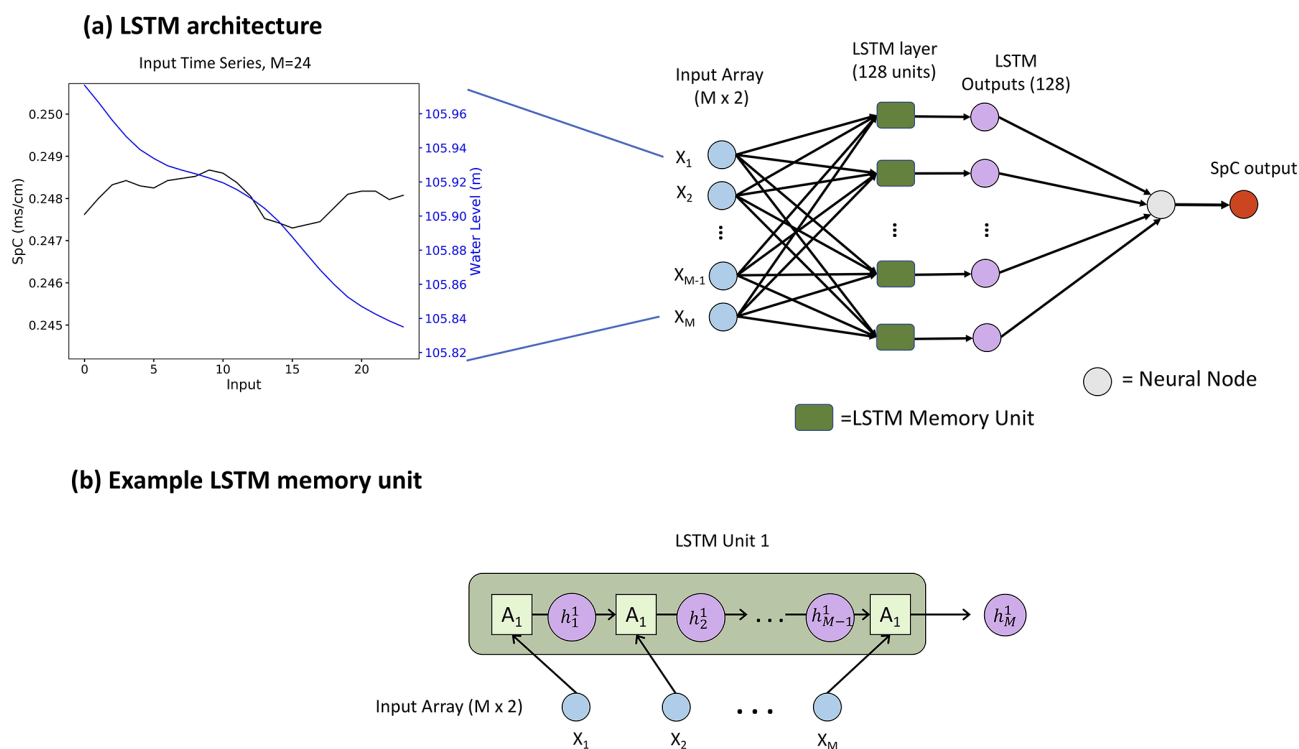
Figure 4. Illustration of LSTM models for gap filling. Panel **(a)** shows the architecture of the LSTM models, where M is the input window size; it includes example input with $M = 24$ and an example LSTM layer with 128 units. Panel **(b)** presents an example of an LSTM unit, where A is the repeating module of the LSTM unit, and h is the output.

perature), we found that including SpC and water level measurements in the input window yielded the most robust performance. Therefore, we used historic water level (m) and SpC $(\mathrm{m\,cm^{-1}})$ observations to fill gaps in the SpC time series of a single well. Using measurements from multiple wells as input allows the models to account for both the temporal and spatial correlations in the data, which impacts gap-filling performance. Wells were selected based on adequate data availability and distance from the target well. Assuming the observations from W ($W \geq 1$) wells are used to fill in data gaps, the input size of the model is then $M \times 2\,W$.

3.1 LSTM models for gap filling

We designed an LSTM architecture, shown in Fig. 4, to train models of an input size of M time steps and an output size of one time step. The LSTM model contains a single LSTM layer followed by an output dense layer. Figures S1 and S2 in the Supplement show the detailed structure of the LSTM layer.

Training data for the LSTM models were created by finding data segments of $M + 1$ h with no missing values, i.e., no gaps in the data, for both water level and SpC measurements over a specified monitoring window. The well data were then preprocessed by normalizing all measurements to

Table 1. Parameters used in training single-well LSTM models.

Parameter	Values
Training wells	1-1, 1-10A, 1-15, 2-2, 2-3, 2-5
Synthetic gap length (hours)	1, 6, 12, 24, 48, 72
Model input window (M hours)	24, 48, 72, 96, 120, 144, 168
LSTM units (U units)	32, 64, 128
Learning rate (L)	1×10^{-3}, 1×10^{-4}, 1×10^{-5}
Training period	2012–2015
Validation period[a]	2011
Testing period[b]	2008 for well 2-5; 2017 for well 1-15; 2016 for all other wells

[a] The parameter was used to select the best LSTM model configurations and hyperparameters. [b] The parameter was used to evaluate the performance of LSTM vs. ARIMA.

zero mean and unit variance for each variable, as SpC is on a scale of 10^{-1}, and water level is on a scale of 10^2. Validation datasets were used to select the best model hyperparameters and optimal input window size M (Sect. 3.1.1) for gap filling at each well. Another independent testing period was selected at each well, depending on data availability, to compare the gap-filling performance using the LSTM and ARIMA methods. The complete set of alternatives that we considered for each LSTM model configuration is shown in Table 1. We used the Adam optimizer (Kingma and Ba, 2014) for training and the mean-squared error as the loss function. The models were trained for 50 iterations (i.e., epochs) over the training data.

To evaluate the accuracy of the trained LSTM models in filling SpC data gaps during the validation and testing processes, we assumed that synthetic gaps of various lengths (e.g., 1, 6, 12, 24, 48, and 72 h, referred to as gap scenarios hereafter) exist in the validation or testing dataset of a well. We assume that only the SpC measurements are missing, whereas the water level measurements are available. An LSTM model configured with an input of M is then given M hours of data from the time series preceding the occurrence of a gap (assuming no missing values in these M hours) to fill in hour by hour. The accuracy of the gap-filling model is evaluated by calculating the mean absolute percentage error (MAPE; %) between the filled in (i.e., predicted) and observed SpC values:

$$\text{MAPE} = 100 \times \frac{1}{n} \sum_{t=1}^{n} \left| \frac{\text{Prediction} - \text{Observation}}{\text{Observation}} \right|, \quad (1)$$

where n is the total number of synthetic missing data points during the evaluation period.

In addition to MAPE, the models are evaluated using the Nash–Sutcliffe model efficiency coefficient (NSE) (Nash and Sutcliffe, 1970) and the Kling–Gupta efficiency (KGE) (Gupta et al., 2009) metric. NSE is a metric used to assess the predictive skills and accuracy of hydrological models. Values range from $-\infty$ to 1, where 1 indicates a perfect model fit, 0 indicates that the model has the same predictive power as

the observational mean, and a value less than 0 indicates that the model is a worse predictor than the observational mean. The NSE of the SpC predictions is calculated using the following equation:

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n} (P_t - O_t)^2}{\sum_{t=1}^{n} (O_t - \mu(O))^2}, \quad (2)$$

where n is the total number of synthetic missing data points during the evaluation period, P_t and O_t are the respective predicted and observed SpC values at time t, and $\mu(O)$ is the mean observed SpC value.

The KGE is another goodness-of-fit metric used to evaluate hydrological models by combining the three components of the NSE model errors (i.e., correlation, bias, and ratio of variances or coefficients of variation) in a more balanced way. It has the same range of values as the NSE, where 1 indicates a perfect model fit. The KGE is calculated on a model's SpC predictions using the following equations:

$$\text{KGE} = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \quad (3)$$

$$r = \frac{\text{cov}(O, P)}{\sigma(O) \cdot \sigma(P)}, \quad (4)$$

$$\alpha = \frac{\sigma(P)}{\sigma(O)}, \quad (5)$$

$$\beta = \frac{\mu(P)}{\mu(O)}. \quad (6)$$

Here, cov is the covariance, σ is the standard deviation, and μ is the arithmetic mean.

In addition to the LSTM models trained for the single-well setup, we also trained multi-well LSTM models that used observations from wells 1-1, 1-10A, and 1-16A to fill in data gaps for well 1-1. We explored the same set of configuration parameters in the multi-well LSTM models as shown in Table 1 for the single-well models. We then compared the gap-filling performance of the multi-well LSTM with the single-well LSTM model for well 1-1. The multi-well models were not explored for other wells due to the lack of close-proximity neighboring wells.

Optimizing LSTM model configuration

We used a grid-search approach to explore different LSTM model hyperparameter configurations to find the best model for a given gap length at each well. This involved iterating over all combinations of the input time window size (M), the number of units (U) in the LSTM layer, and the learning rate (L) listed in Table 1 for each well. We chose the optimal LSTM configuration using model performance on the validation dataset (see Table 1) based on the MAPE defined in Eq. (1). The combinations that yielded the lowest SpC MAPEs were selected as the best configuration for a given gap length at each well. These configurations are shown in Table S1 in the Supplement. The best model configurations were then used to evaluate the LSTM-based gap-filling method against the ARIMA-based method (Sect. 3.2) using relative errors (similar to MAPE by setting $n = 1$ in Eq. (1) and removing the absolute value operation) calculated for each data point in the testing period (Table 1). These varied among the wells due to the availability of continuous data required for testing.

3.2   ARIMA models for gap filling

ARIMA is one of the most general model classes for extrapolating time series to produce forecasts. We used it as a baseline to compare and assess the LSTM gap-filling method. ARIMA can be applied to nonstationary time-series data using a combination of differencing, autoregressive, and moving average components. A nonseasonal ARIMA(p, d, q) model is given by

$$Y_t = c + \phi_1 Y_{t-1}^d + \phi_p Y_{t-p}^d + \ldots + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t, \quad (7)$$

where ϕs and θs are polynomials of orders p and q, respectively, each containing no roots inside the unit circle; e represents error terms; Y^d is Y differenced d times; and c is a constant. Note that only nonseasonal terms (p, d, q) are included in Eq. (2). Seasonal structure can be added with parameters (P, D, Q)$_m$ to the base ARIMA model to convert it to ARIMA(p, d, q)(P, D, Q)$_m$, with a periodic component containing m periods. $c \neq 0$ implies a polynomial of order $d + D$ in the forecast function. The detailed mathematical equations for the seasonal ARIMA model are provided in the Supplement.

The main task in ARIMA-based forecasting is to select appropriate model orders, i.e., the values of p, q, d, P, Q, and D. If d and D are known, we can select the orders p, q, P, and Q via an information criterion such as the Akaike information criterion (AIC):

$$\text{AIC} = -2\log(L) + 2(p + q + P + Q + k), \quad (8)$$

where $k = 1$ if $c \neq 0$ and is 0 otherwise, and L is the likelihood of the model fitted to the differenced data. The ARIMA models were built using the "auto.arima" function from the "forecast" R package (Hyndman and Khandakar, 2007), which applies the Hyndman–Khandakar algorithm (Hyndman and Khandakar, 2008) that minimizes the AIC to obtain the best-fit parameters of the ARIMA model.

Similar to the LSTM-based gap filling, we explored various input window sizes, from 24 to 504 h in increments of 24 h, for the ARIMA model at each well to identify the optimal input windows within the search range. An optimal input window size is chosen for each gap length of each well using the same MAPE metric (i.e., Eq. 1) on the validation dataset.

4   Results and discussion

4.1   Performance of single-well LSTM models

We selected the best combination of LSTM units (U) and learning rate (L) for each input time window (M) under each gap length at each well using the MAPE metric. The validation MAPEs of those selected models were summarized in box plots under different groupings, examples of which are shown in Fig. 5. Each MAPE box plot was drawn from a group of models with one parameter (corresponding to each x axis) fixed at a given value while the other parameters cycle through their possible combinations.

As shown in Fig. 5a, model performance deteriorates as the gap length increases, indicating that the LSTM-based method tends to lose ground truth information from its input to inform prediction. In comparing MAPEs across the various input window sizes shown in Fig. 5b, we observe that models with all input windows have comparable MAPE summary statistics, with larger input windows (> 96 h) leading to slightly smaller MAPE quartiles. The larger input windows also yield fewer outliers on the larger MAPE end, indicating that the memory units in the LSTM layers are capturing important daily to weekly signatures for some wells (which is evident in the WPS plots in Fig. 2 for all wells except for well 1-15).

The performance of single-well LSTM models varied among the wells, as shown in Fig. 5c. The LSTM models for well 1-15 lead performance with the smallest MAPEs, whereas the model of well 2-2 yields the worst performance. The LSTM models for wells 1-1,1-10A, 2-3, and 2-5 performed comparably overall, with slightly more large MAPE outliers for well 2-3.

4.2   Single-well LSTM and ARIMA comparisons

The single-well LSTM gap-filling approach was compared to the ARIMA approach using the relative errors calculated for each data point assumed to be missing in the testing data by setting $n = 1$ in Eq. (1) for MAPE. Relative errors were used to show the overestimations or underestimations of both approaches. Their respective best model configurations, determined on the validation dataset (i.e., data from year 2011), were used to compare the two approaches. Figure 6 illus-
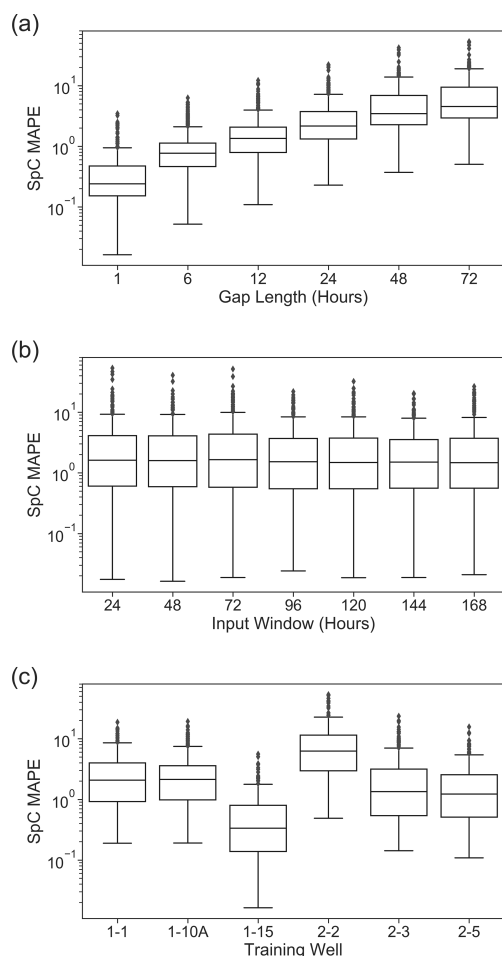
Figure 5. Gap-filling performance for SpC evaluated against the validation datasets, grouped by gap lengths **(a)**, model input window size M **(b)**, and training wells **(c)**.

trates the optimal input windows for the LSTM and ARIMA methods. We observe that the LSTM models require much less input information than the ARIMA method under all gap lengths for all of the wells.

Figure 7 shows the box plots of relative errors under different gap lengths for all testing wells. As expected, the relative errors increase as the gap length increases for both approaches. The ARIMA models tend to perform better than the LSTM models in terms of interquartile range. However, the ARIMA models produce more outliers of large positive or negative relative errors than the LSTM models in general and particularly for larger gap lengths (48 and 72 h). For well 1-15, the relative errors of both approaches are small for all gap lengths. Both approaches appear to have larger error outliers at well 2-3.

For each well, we performed a T test to calculate the T score and p value between the relative errors of the two models to determine the significance in the performance difference of the models. As seen in Table 2, each well has a

high T score and a p value significantly less than 0.05. Thus, the differences between their relative errors are significant and meaningful.

In addition to the relative errors, we calculated the MAPE, root-mean-squared error (RMSE), NSE, and KGE for the best LSTM and ARIMA model for each gap length. Table 2 compares the performance of the LSTM and ARIMA models filling in gap lengths of 24 h. The results for all gap lengths are given in the Supplement (Table S2).

The LSTM and ARIMA models yielded comparable average metrics at all wells for the 24 h gap length, as can be seen in Table 2. The NSE and KGE resulting from both models are close to 1 for all of the wells with negligible differences between the two models. The difference in the MAPE and RMSE is also small, with more notable differences for wells 2-2 and 2-3, where the ARIMA models resulted in lower MAPE and RMSE values.

In addition to the error statistics, it is also important to examine how well a gap-filling method captures the desired dynamics patterns in the gap-filled time series. Therefore, the SpC time series reproduced by the gap-filling methods during the testing period (2016 for wells 1-1, 1-10A, 2-2, 2-3; 2017 for well 1-15; 2008 for well 2-5) with 24 h synthetic gaps are evaluated against the real time series. The model configurations are the same as those used in the error statistics comparison (Fig. 6). A gap length of 24 h is the selected example, as we consider it a reasonably challenging case to fill the gaps in time-series data exhibiting significant non-stationarity, such as the SpC data at well 2-3. Moreover, the relative performance of the two approaches is similar at other gap lengths with varying error magnitudes.

As shown in the first column of Fig. 8, both approaches capture the general dynamic patterns in the data fairly well. For more detail, the time series of relative errors for both methods are provided in Fig. S3. The ARIMA approach (blue lines in column 1) missed some abrupt changes that occur over a short time window (i.e., at higher frequency), leading to more error spikes in all wells. This is consistent with the relative error outliers in Fig. 7 and is an indication that the ARIMA models lack mechanisms to represent such high-frequency changes. The LSTM approach (red lines in column 1) is able to better capture such dynamics in all of the wells. However, the inclusion of the high-frequency fluctuations may also lead to overly dynamic predictions in time windows dominated by lower-frequency fluctuations. This contributed to less desirable relative errors distributed between the first and third quartiles in some wells (i.e., wells 1-1, 2-3, 2-2, and 1-15), as shown in Fig. S4. The errors are likely caused by the variability in dynamics signatures among the training, validation, and testing periods, as well as the selection of the training loss functions and validation metric that balance between the occurrence of small and large errors to achieve optimal solutions.

To further investigate the dependence of the relative performance of the two gap-filling methods on the inherent dy-
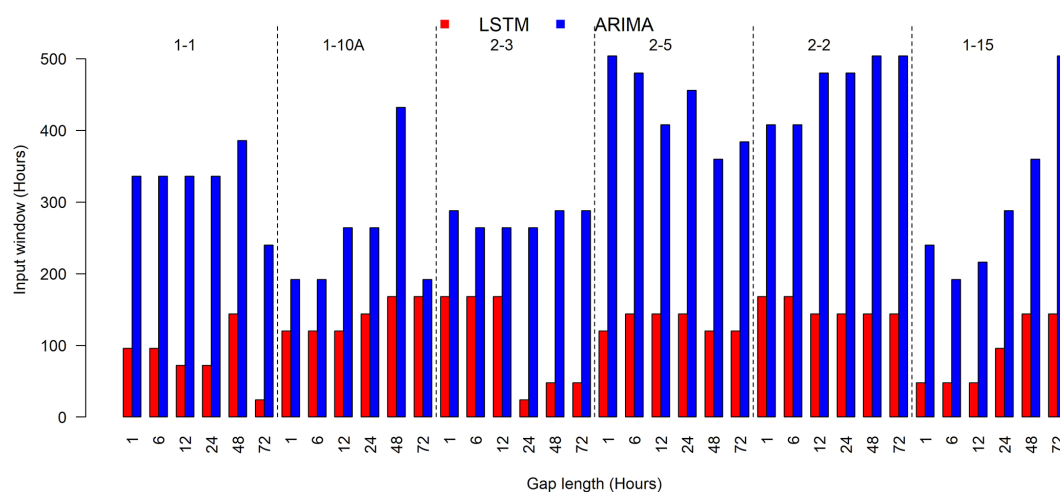
Figure 6. Optimal input windows for the LSTM and ARIMA models to fill gaps of various lengths at each well.
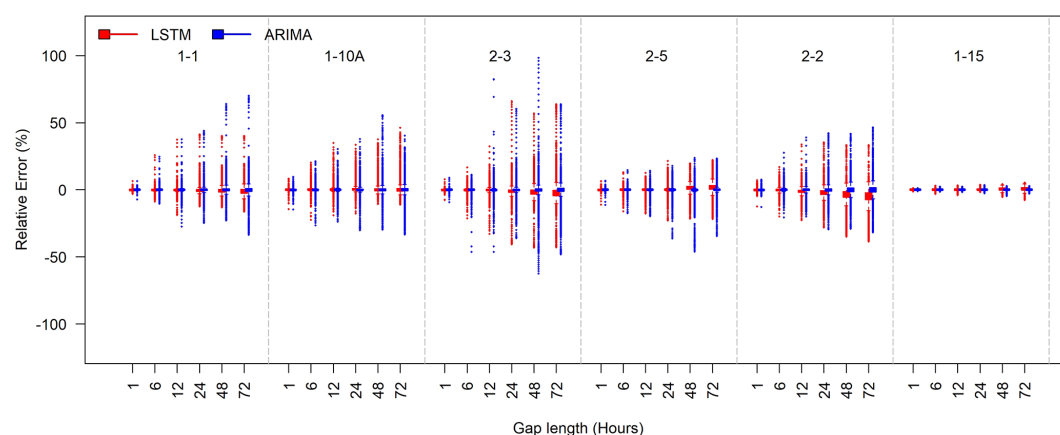

Figure 7. Box plots of the relative errors for filling SpC gaps of various lengths (1, 6, 12, 24, 48, and 72 h) at each well during the test periods. The best LSTM and ARIMA models were used for evaluation. The LSTM and ARIMA models are represented by red bars and blue bars, respectively.

namics in each time series, spectral analyses for the testing SpC datasets were performed using the same wavelet decomposition method for the multiyear analyses (shown earlier in Fig. 2). As shown in Fig. 8, the time windows of high relative errors are found to approximately co-locate with the time when the high-frequency (daily and subdaily) signals gain more power. The differences between the LSTM and ARIMA models tend to be amplified during those time windows. Wells 1-1, 1-10A, and 2-2 share similar seasonal patterns in WPS, with the highest intensity bin above 1024 h across February to July. Their average WPSs all show peaks around daily and subdaily frequencies. Well 2-3 has its greatest energy between 16 and 256 h from January to July. Well 2-5 has low variability intensities at daily and subdaily frequencies with low-frequency variations (monthly and seasonal) dominating the January to March time frame. For well 1-15, one of its strongest intensities is above 2048 h across the entire year, and the other strong intensities are narrow bands between 16 and 256 h. In general, both LSTM and ARIMA are effective at capturing low-frequency variability (monthly and seasonal). Although LSTM is more effective at capturing high-frequency (daily and subdaily) fluctuations and nonlinearities in the datasets, it may also lead to overly dynamic predictions in time windows with no considerable high-frequency fluctuations. However, the errors during these time windows are small and can be improved by smoothing if such fluctuations are undesirable.

There is also a significant difference in the computational cost for the LSTM and ARIMA methods. ARIMA requires very few computational resources: the auto.arima function in R requires approximately 40 s to fit and validate a model for each prediction segment on a personal computer with a 3.00 GHz CPU. Conversely, training and validating a single LSTM model takes approximately 20–30 min on GPUs based on dual NVIDIA P100 12GB PCI-e cards.
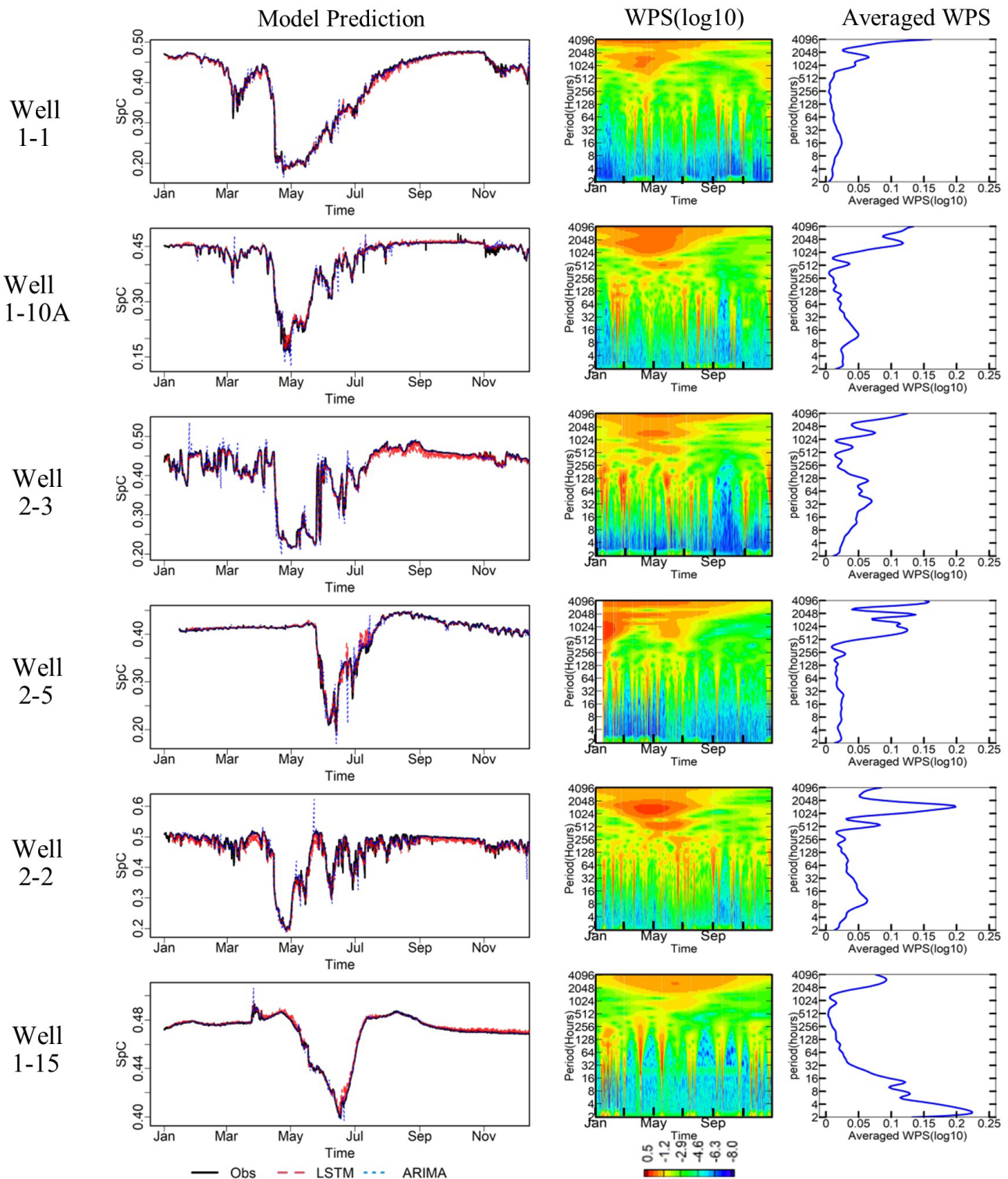
Figure 8. Column 1 shows time series of model predictions from the LSTM (in red) and ARIMA (in blue) methods, respectively, assuming a 24 h synthetic gap in the SpC data compared with observations (in black). The best model configurations were used for all models. The testing data come from the year 2016 for wells 1-1, 1-10A, 2-2, and 2-3; from the year 2017 for well 1-15; and from the year 2008 for well 2-5. Column 2 is the spectrogram of each well, and column 3 is the averaged WPS for the corresponding year.

Table 2. Comparison of single-well LSTM and ARIMA models for a 24 h synthetic gap in the SpC data of the test set for each well. The models are the same as those used in Fig. 8. The calculated statistics are as follows: the mean absolute percentage error (MAPE), the root-mean-squared error (RMSE), the Nash–Sutcliffe model efficiency coefficient (NSE), and the Kling–Gupta efficiency (KGE). The T score and p value are calculated for the relative errors of the two models per well.

Well	Model type	MAPE	RMSE	NSE	KGE	T score	P value
1-1	LSTM	1.38	8.33×10^{-3}	0.991	0.988	19.1	1.00×10^{-80}
	ARIMA	1.36	8.98×10^{-3}	0.989	0.994		
1-10A	LSTM	1.37	8.07×10^{-3}	0.986	0.968	-24.6	1.48×10^{-131}
	ARIMA	1.5	9.60×10^{-3}	0.98	0.987		
1-15	LSTM	0.259	1.88×10^{-3}	0.989	0.982	-48.9	$0.00 \times 10^{+00}$
	ARIMA	0.119	1.18×10^{-3}	0.996	0.997		
2-2	LSTM	2.97	1.87×10^{-2}	0.922	0.962	48.1	$0.00 \times 10^{+00}$
	ARIMA	2.23	1.64×10^{-2}	0.939	0.967		
2-3	LSTM	2.15	1.63×10^{-2}	0.945	0.965	21.6	4.69×10^{-102}
	ARIMA	1.72	1.48×10^{-2}	0.954	0.971		
2-5	LSTM	0.929	6.86×10^{-3}	0.976	0.988	-9.6	9.22×10^{-22}
	ARIMA	0.866	7.45×10^{-3}	0.971	0.977		

4.3 Performance of multi-well models

We evaluated the predictive ability of multi-well models using both approaches for filling gaps of various lengths in the SpC data at well 1-1 by comparing their performance against their single-well counterparts. Well 1-1 was chosen because of the data availability for nearby wells (wells 1-10A and 1-16A). Similar to the single-well ARIMA and LSTM model for well 1-1, the multi-well models also predict the SpC measurement using water level and SpC from three wells. We adopted the same LSTM architecture from the single-well LSTM model and trained the same set of alternatives considering input window sizes, LSTM units, and learning rates for various gap lengths as listed in Table 1. The same training and validation periods were adopted to select the optimal combination of M, U, and L. The optimal combinations are shown in Table S3. For the multi-well ARIMA models, we included additional variables as regression terms when building and fitting models using the auto.arima function. The optimal input window sizes of ARIMA are 216, 240, 288, 288, 288, and 192 h for gap lengths of 1, 6, 12, 24, 48, and 72 h, respectively. These are smaller than the optimal window sizes of the single-well models.

The box plots of relative errors yielded from the single-well and multi-well models using both approaches are provided in Fig. 9 for comparison. Additionally, we include performance metrics for comparing the single- and multi-well models in Table 3. Additional spatial information seems to exacerbate the relative errors of the ARIMA models, except in large gaps (e.g., 72 h). The LSTM approach, on the contrary, benefits from the information carried by the neighboring wells when filling in the larger gaps, whereas the performance for small gaps remains unchanged. The aggregated performance metrics in Table 3 show slightly improved metrics for multi-well ARIMA models for gaps smaller than 24 h compared with the single-well models, while the turning point in relative performance is at 12 h for the LSTM models. The deteriorated performance metrics of the multi-well LSTM models at the larger gap lengths are consistent with the larger interquartile ranges, as shown in the box plots of relative errors in Fig. 9. However, the multi-well LSTM and ARIMA models can reduce the occurrence of large relative errors for larger gaps and provide more robust gap filling under those circumstances.

These comparisons show that, although the information from a single well may be sufficient to fill in gaps smaller than a day, including spatial information from neighboring wells in the LSTM and ARIMA models could potentially increase the chance of successes in filling data gaps under more challenging circumstances, such as capturing more complex dynamic patterns with longer data gaps. While the aggregated metrics provide an overall assessment of model performance, examining the distribution of relative errors could provide complementary information on large error spikes while selecting optimal model configurations.

5 Conclusions

In this study, we implemented an LSTM-based gap-filling method to account for spatiotemporal correlations in monitoring data. We extensively evaluated the method on its ability to fill data gaps in the groundwater SpC measurements that are often used to indicate groundwater and river water
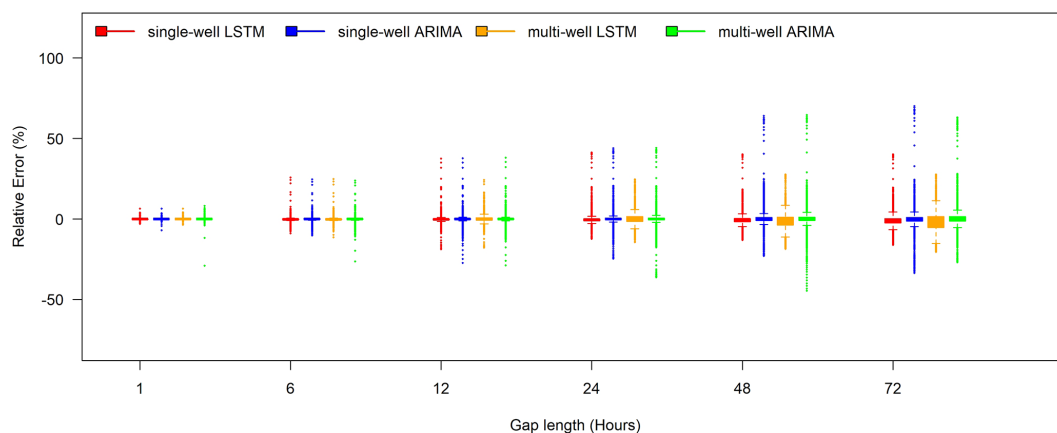
Figure 9. Comparing relative error performance between the best single-well LSTM models (well 1-1 – red), multi-well LSTM models (wells 1-1, 1-10A, and 1-16A – yellow), a single-well ARIMA model (blue), and a multi-well ARIMA model (green) for filling in various SpC gap lengths for well 1-1 during the test period (year 2016).

interactions along river corridors. We took advantage of a 10-year, spatially distributed, multivariable time-series dataset collected by a groundwater monitoring well network and optimized an LSTM architecture for filling SpC data gaps. A primary advantage of using LSTM is its ability to incorporate spatiotemporal correlations and nonlinearity in model states without a priori assuming an explicit form of correlations or nonlinear functions in advancing system states. We compared the performance of a single-well LSTM-based gap-filling method with a traditional gap-filling method, ARIMA, to evaluate how well an LSTM model can capture multifrequency dynamics. We also trained LSTM and ARIMA models that take input from multiple wells to predict responses at one well. The multi-well models were compared with single-well models to identify and assess improvements in gap-filling performance from including additional spatial correlation from neighboring wells.

In general, both LSTM and ARIMA methods were highly accurate in filling smaller data gaps (i.e., 1 and 6 h). They were reasonably effective at filling in medium gaps between 12 and 48 h, while more work is needed for gaps larger than 48 h. Both models captured the long-term trends in data (i.e., low-frequency variations at monthly or seasonal timescales). The ARIMA method was found to have difficulty capturing abrupt changes and is, thus, more suitable for time series with less dynamic behavior. Compared with the ARIMA models, the LSTM models excel in dealing with high-frequency dynamics (daily and subdaily) and nonlinearities, although they require more training data and computational resources. As a side effect of including high-frequency (daily and subdaily) fluctuations in the model, the LSTM approach may produce overly dynamic predictions in time windows that lack dynamics. As with any deep learning method, the availability of sufficient training data that cover a wide range of conditions is critical for the success of LSTM methods. Extrapolating

the LSTM models to conditions beyond those in the training data remains a major challenge.

Wavelet analysis could provide useful insights into the dynamic signatures of the data and changes in the composition of their important frequencies over time, which can serve as a basis for selecting an appropriate gap-filling method. For example, the ARIMA method would work well if the dynamics are dominated by seasonal cycles, while more sophisticated approaches like LSTM-based methods could work better if there is evidence of weekly, daily, and subdaily fluctuations. Depending on the mixture of high- and low-frequency variability inherent in the time series, different LSTM architecture and configurations can be explored and evaluated through hyperparameter searches with respect to LSTM layers, dense layers, and activation functions to achieve better performance in capturing complex dynamics. The optimal LSTM model configuration and achievable performance would vary case by case.

We also demonstrated that incorporating spatial information from neighboring stations in LSTM models could contribute to performance improvements under challenging scenarios with dynamic system behaviors and longer data gaps of up to 2 d. However, other alternatives need to be explored for gaps beyond 2 d. Bidirectional and convolutional LSTMs are two promising methods to leverage information from the future time window and spatially distributed networks, respectively. While we introduced a new method that can be broadly applied to fill in gaps in an irregularly spaced network for monitoring groundwater and surface water interactions, the transferability of this method to other monitoring systems could be more extensively evaluated by community participation. When applying LSTM or other DNN-based methods for gap filling, it is important to rigorously evaluate the model performance by asking the following questions:

Table 3. Comparison of single-well and multi-well LSTM and ARIMA models for all synthetic gap lengths in the SpC data. The models are the same as those used in Fig. 9. Calculations are performed on the test dataset for well 1-1 (year 2016). The calculated statistics are as follows: the mean absolute percentage error (MAPE), the root-mean-squared error (RMSE), the Nash–Sutcliffe model efficiency coefficient (NSE), and the Kling–Gupta efficiency (KGE). T scores and p values are calculated on the relative errors of the two single-well models for each gap length and calculated on the relative errors of the two multi-well models.

Gap length (hours)	Model type	MAPE	RMSE	NSE	KGE	T score	p value
1	Single-well LSTM	0.117	7.76×10^{-4}	1.0	0.999	11.6	6.14×10^{-31}
	Single-well ARIMA	0.183	1.23×10^{-3}	1.0	1.0		
	Multi-well LSTM	0.117	7.94×10^{-4}	1.0	1.0	9.37	8.54×10^{-21}
	Multi-well ARIMA	0.134	1.22×10^{-3}	1.0	1.0		
6	Single-well LSTM	0.435	2.92×10^{-3}	0.999	0.995	13.8	4.46×10^{-43}
	Single-well ARIMA	0.461	3.29×10^{-3}	0.999	0.999		
	Multi-well LSTM	0.435	2.98×10^{-3}	0.999	0.998	12.2	5.82×10^{-34}
	Multi-well ARIMA	0.405	3.16×10^{-3}	0.999	0.999		
12	Single-well LSTM	0.75	5.07×10^{-3}	0.997	0.996	16.0	2.05×10^{-57}
	Single-well ARIMA	0.781	5.47×10^{-3}	0.996	0.996		
	Multi-well LSTM	1.19	6.48×10^{-3}	0.994	0.985	5.25	1.55×10^{-7}
	Multi-well ARIMA	0.77	5.40×10^{-3}	0.996	0.997		
24	Single-well LSTM	1.38	8.33×10^{-3}	0.991	0.988	19.1	1.00×10^{-80}
	Single-well ARIMA	1.36	8.98×10^{-3}	0.989	0.994		
	Multi-well LSTM	2.26	1.17×10^{-2}	0.982	0.968	7.77	8.48×10^{-15}
	Multi-well ARIMA	1.47	9.55×10^{-3}	0.988	0.99		
48	Single-well LSTM	2.13	1.21×10^{-2}	0.98	0.988	17.8	3.24×10^{-70}
	Single-well ARIMA	2.15	1.34×10^{-2}	0.976	0.988		
	Multi-well LSTM	3.49	1.76×10^{-2}	0.958	0.969	28.4	7.83×10^{-174}
	Multi-well ARIMA	2.35	1.40×10^{-2}	0.974	0.981		
72	Single-well LSTM	2.56	1.40×10^{-2}	0.974	0.983	26.7	4.91×10^{-154}
	Single-well ARIMA	2.57	1.46×10^{-2}	0.971	0.985		
	Multi-well LSTM	4.41	2.19×10^{-2}	0.936	0.955	31.7	4.78×10^{-214}
	Multi-well ARIMA	3.02	1.78×10^{-2}	0.958	0.972		

– What is the dynamics signature of the data to be filled?

– How is gap-filling performance impacted by the length of gaps?

– How does the amount of training data impact the model performance?

– How does the choice of the input time window impact gap-filling performance?

– How much value can measurements at neighboring add to the performance improvement?

Better capturing spatiotemporal dynamics in system states is essential for generating the most valuable insights to advance our understanding of dynamic complex systems.

References

Alvera-Azcárate, A., Barth, A., Parard, G., and Beckers, J.-M.: Analysis of SMOS sea surface salinity data using DINEOF, Remote Sens Environ., 180, 137–145, 2016.

Amaranto, A., Munoz-Arriola, F., Corzo, G., Solomatine, D. P., and Meyer, G.: Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland, J. Hydroinform., 20, 1227–1246, 2018.

Amaranto, A., Munoz-Arriola, F., Solomatine, D., and Corzo, G.: A spatially enhanced data-driven multimodel to improve semiseasonal groundwater forecasts in the High Plains aquifer, USA, Water Resour. Res., 55, 5941–5961, 2019.

Banerjee, S., Carlin, B. P., and Gelfand, A. E.: Hierarchical modeling and analysis for spatial data, CRC Press, https://doi.org/10.1201/9780203487808, 2014.

Beckers, J.-M. and Rixen, M.: EOF calculations and data filling from incomplete oceanographic datasets, J. Atmos. Ocean. Tech., 20, 1839–1856, 2003.

Beckers, J.-M., Barth, A., and Alvera-Azcárate, A.: DINEOF reconstruction of clouded images including error maps – application to the Sea-Surface Temperature around Corsican Island, Ocean Sci., 2, 183–199, https://doi.org/10.5194/os-2-183-2006, 2006.

Calculli, C., Fassò, A., Finazzi, F., Pollice, A., and Turnone, A.: Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy, Environmetrics, 26, 406–417, 2015.

Chen, S., Wang, X., Guo, H., Xie, P., and Sirelkhatim, A. M.: Spatial and Temporal Adaptive Gap-Filling Method Producing Daily Cloud-Free NDSI Time Series, IEEE J. Sel. Top. Appl., 13, 2251–2263, 2020.

Chen, X., Murakami, H., Hahn, M. S., Hammond, G. E., Rockhold, M. L., Zachara, J. M., and Rubin, Y.: Three-dimensional Bayesian geostatistical aquifer characterization at the Hanford 300 Area using tracer test data, Water Resour. Res., 48, W06501, https://doi.org/10.1029/2011WR010675, 2012.

Chen, X., Hammond, G. E., Murray, C. J., Rockhold, M. L., Vermeul, V. R., and Zachara, J. M.: Application of ensemble-based data assimilation techniques for aquifer characterization using tracer data at Hanford 300 area, Water Resour. Res., 49, 7064–7076, https://doi.org/10.1002/2012WR013285, 2013.

Cheng, T., Haworth, J., and Wang, J.: Spatio-temporal autocorrelation of road network data, J. Geogr. Syst., 14, 389–413, https://doi.org/10.1007/s10109-011-0149-5, 2012.

Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., and Wang, J.: Spatiotemporal data mining, in: Handbook of regional science, Springer, 1173–1193, https://doi.org/10.1007/978-3-642-23430-9_68, 2014.

Contractor, S. and Roughan, M.: Efficacy of Feedforward and LSTM Neural Networks at Predicting and Gap Filling Coastal Ocean Timeseries: Oxygen, Nutrients, and Temperature, Front. Mar. Sci., 8, 637759, https://doi.org/10.3389/fmars.2021.637759, 2021.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E.: Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets, J. Am. Stat. Assoc., 111, 800–812, 2016.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J.: Estimation and prediction in spatial models with block composite likelihoods, J. Comput. Graph. Stat., 23, 295–315, 2014.

Fang, K., Shen, C., Kifer, D., and Yang, X.: Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network, Geophys. Res. Lett., 44, 11030–11039, https://doi.org/10.1002/2017GL075619, 2017.

Faruk, D. Ö.: A hybrid neural network and ARIMA model for water quality time series prediction, Eng. Appl. Artif. Intel., 23, 586–594, 2010.

Finley, A. O., Banerjee, S., and Gelfand, A. E.: spBayes for large univariate and multivariate point-referenced spatio-temporal data models, arXiv [preprint], arXiv:1310.8192, 2013.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could Machine Learning Break the Convection Parameterization Deadlock?, Geophys. Res. Lett., 45, 5742–5751, https://doi.org/10.1029/2018GL078202, 2018.

Ghil, M., Allen, M., Dettinger, M., Ide, K., Kondrashov, D., Mann, M., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., and Yiou, P.: Advanced spectral methods for climatic time series, Rev. Geophys., 40, 3.1–3.41, https://doi.org/10.1029/2000RG000092, 2002.

Grant, G. E. and Dietrich, W. E.: The frontier beneath our feet, Water Resour. Res., 53, 2605–2609, 2017.

Graves, A.: Generating sequences with recurrent neural networks, arXiv [preprint], arXiv:1308.0850, 2013.

Graves, A., Abdel-rahman, M., and Geoffrey, H.: Speech recognition with deep recurrent neural networks, in: 2013 IEEE international conference on acoustics, speech and signal process-

ing, 6645–6649, https://doi.org/10.1109/ICASSP.2013.6638947, 2013.

Griffith, D. A.: Modeling spatio-temporal relationships: retrospect and prospect, J. Geogr. Syst., 12, 111–123, 2010.

Grinsted, A., Moore, J. C., and Jevrejeva, S.: Application of the cross wavelet transform and wavelet coherence to geophysical time series, Nonlin. Processes Geophys., 11, 561–566, https://doi.org/10.5194/npg-11-561-2004, 2004.

Grossmann, A. and Morlet, J.: Decomposition of Hardy functions into square integrable wavelets of constant shape, SIAM J. Math. Anal., 15, 723–736, 1984.

Güler, C. and Thyne, G. D.: Hydrologic and geologic factors controlling surface and groundwater chemistry in Indian Wells-Owens Valley area, southeastern California, USA, J. Hydrol., 285, 177–198, 2004.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Han, P., Wang, P. X., Zhang, S. Y., and Zhu, D. H.: Drought forecasting based on the remote sensing data using ARIMA models, Math. Comput. Model., 51, 1398–1403, 2010.

Ho, S., Xie, M., and Goh, T.: A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction, Comput. Ind. Eng., 42, 371–375, 2002.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Comput., 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, http://www.bioinf.jku.at/publications/older/ch7.pdf (last access: 5 April 2022), 2001.

Hocke, K. and Kämpfer, N.: Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram, Atmos. Chem. Phys., 9, 4197–4206, https://doi.org/10.5194/acp-9-4197-2009, 2009.

Hyndman, R. J. and Khandakar, Y.: Automatic time series for forecasting: the forecast package for R, 6/07, Monash University, Department of Econometrics and Business Statistics, https://doi.org/10.18637/jss.v000.i00, 2007.

Hyndman, R. J. and Khandakar, Y.: Automatic time series forecasting: the forecast package for R, J. Stat. Softw., 27, 1–22, 2008.

Jordan, M.: Attractor dynamics and parallelism in a connectionist sequential machine, in: Proc. of the Eighth Annual Conference of the Cognitive Science Society, Erlbaum, Hillsdale, NJ, 112–127, https://ci.nii.ac.jp/naid/10018634949/en/ (last access: January 1990), 1986.

Kamarianakis, Y. and Prastacos, P.: Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches, Transp. Res. Record, 1857, 74–84, https://doi.org/10.3141/1857-09, 2003.

Kamarianakis, Y. and Prastacos, P.: Space–time modeling of traffic flow, Comput. Geosci., 31, 119–133, 2005.

Kandasamy, S., Baret, F., Verger, A., Neveux, P., and Weiss, M.: A comparison of methods for smoothing and gap filling time series of remote sensing observations – application to MODIS LAI products, Biogeosciences, 10, 4055–4071, https://doi.org/10.5194/bg-10-4055-2013, 2013.

Katzfuss, M. and Cressie, N.: Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets, J. Time Ser. Anal., 32, 430–446, 2011.

Katzfuss, M. and Cressie, N.: Bayesian hierarchical spatio-temporal smoothing for very large datasets, Environmetrics, 23, 94–107, 2012.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, CoRR, abs/1412.6980, arXiv [preprint], arXiv:1412.6980, 2014.

Kondrashov, D. and Ghil, M.: Spatio-temporal filling of missing points in geophysical data sets, Nonlin. Processes Geophys., 13, 151–159, https://doi.org/10.5194/npg-13-151-2006, 2006.

Kondrashov, D., Shprits, Y., and Ghil, M.: Gap filling of solar wind data by singular spectrum analysis, Geophys. Res. Lett., 37, L15101, https://doi.org/10.1029/2010GL044138, 2010.

Körner, P., Kronenberg, R., Genzel, S., and Bernhofer, C.: Introducing Gradient Boosting as a universal gap filling tool for meteorological time series, Meteorol. Z., 27, 369–376, 2018.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Längkvist, M., Karlsson, L., and Loutfi, A.: A review of unsupervised feature learning and deep learning for time-series modeling, Pattern Recogn. Lett., 42, 11–24, https://doi.org/10.1016/j.patrec.2014.01.008, 2014.

Lin, C. Y., Abdullah, M. H., Praveena, S. M., Yahaya, A. H. B., and Musta, B.: Delineation of temporal variability and governing factors influencing the spatial variability of shallow groundwater chemistry in a tropical sedimentary island, J. Hydrol., 432, 26–42, 2012.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Pfeifer, P. E. and Deutrch, S. J.: A three-stage iterative procedure for space-time modeling phillip, Technometrics, 22, 35–47, 1980.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Sarafanov, M., Kazakov, E., Nikitin, N. O., and Kalyuzhnaya, A. V.: A Machine Learning Approach for Remote Sensing Data Gap-Filling with Open-Source Implementation: An Example Regarding Land Surface Temperature, Surface Albedo and NDVI, Remote Sens.-Basel, 12, 3865, https://doi.org/10.3390/rs12233865, 2020.

SBRSFA: Using Deep Learning to Fill Spatio-Temporal Data Gaps in Hydrological Monitoring Networks: A Case Study at the U.S. Department of Energy's Hanford Site, SBRSFA [data set], https://sbrsfa.velo.pnnl.gov/datasets/?UUID=14febd81-05b6-47fb-be52-439c4382decd, last access: 5 April 2022.

Schmidhuber, J.: Deep learning in neural networks: An overview, Neural Networks, 61, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003, 2015.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resour. Res., 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018.

Shuai, P., Chen, X., Song, X., Hammond, G. E., Zachara, J., Royer, P., Ren, H., Perkins, W. A., Richmond, M. C., and Huang, M.: Dam Operations and Subsurface Hydrogeology Control Dynamics of Hydrologic Exchange Flows in a Regulated River Reach, Water Resour. Res., 55, 2593–2612, https://doi.org/10.1029/2018WR024193, 2019.

Song, X., Chen, X., Stegen, J., Hammond, G., Song, H.-S., Dai, H., Graham, E., and Zachara, J. M.: Drought Conditions Maximize the Impact of High-Frequency Flow Variations on Thermal Regimes and Biogeochemical Function in the Hyporheic Zone, Water Resour. Res., 54, 7361–7382, https://doi.org/10.1029/2018WR022586, 2018.

Stockwell, R. G., Mansinha, L., and Lowe, R.: Localization of the complex spectrum: the S transform, IEEE T. Signal Proces., 44, 998–1001, 1996.

Strobl, R. O. and Robillard, P. D.: Network design for water quality monitoring of surface freshwaters: A review, J. Environ. Manage., 87, 639–648, 2008.

Stroud, J. R., Stein, M. L., and Lysen, S.: Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice, J. Comput. Graph. Stat., 26, 108–120, 2017.

Sun, A. Y.: Discovering State-Parameter Mappings in Subsurface Models Using Generative Adversarial Networks, Geophys. Res. Lett., 45, 11,137–11,146, https://doi.org/10.1029/2018GL080404, 2018.

Sun, A. Y., Scanlon, B. R., Zhang, Z., Walling, D., Bhanja, S. N., Mukherjee, A., and Zhong, Z.: Combining Physically Based Modeling and Deep Learning for Fusing GRACE Satellite Data: Can We Learn From Mismatch?, Water Resour. Res., 55, 1179–1195, https://doi.org/10.1029/2018WR023333, 2019.

Taylor, C. J. and Alley, W. M.: Ground-water-level monitoring and the importance of long-term water-level data, 1217–2002, US Geological Survey, https://doi.org/10.3133/cir1217, 2002.

Vacha, L. and Barunik, J.: Co-movement of energy commodities revisited: Evidence from wavelet coherence analysis, Energ. Econ., 34, 241–247, 2012.

Valenzuela, O., Rojas, I., Rojas, F., Pomares, H., Herrera, L. J., Guillén, A., Marquez, L., and Pasadas, M.: Hybridization of intelligent techniques and ARIMA models for time series prediction, Fuzzy Set. Syst., 159, 821–845, 2008.

Wang, G., Garcia, D., Liu, Y., De Jeu, R., and Dolman, A. J.: A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations, Environ. Modell. Softw., 30, 139–142, 2012.

Wett, B., Jarosch, H., and Ingerle, K.: Flood induced infiltration affecting a bank filtrate well at the River Enns, Austria, J. Hydrol., 266, 222–234, 2002.

Wikle, C. K., Berliner, L. M., and Cressie, N.: Hierarchical Bayesian space-time models, Environ. Ecol. Stat., 5, 117–154, 1998.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, CoRR, arXiv [preprint], arXiv:1609.08144, 2016

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J.: Image Captioning with Semantic Attention, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4651–4659, https://doi.org/10.1109/CVPR.2016.503, 2016.

Zachara, J. M., Long, P. E., Bargar, J., Davis, J. A., Fox, P., Fredrickson, J. K., Freshley, M. D., Konopka, A. E., Liu, C., McKinley, J. P., Rockhold, M. L., Williams, K. H., and Yabusaki, S. B.: Persistence of uranium groundwater plumes: contrasting mechanisms at two DOE sites in the groundwater–river interaction zone, J. Contam. Hydrol., 147, 45–72, https://doi.org/10.1016/j.jconhyd.2013.02.001, 2013.

Zachara, J. M., Chen, X., Song, X., Shuai, P., Murray, C., and Resch, C. T.: Kilometer-scale hydrologic exchange flows in a gravel-bed river corridor and their implications to solute migration, Water Resour. Res., 56, e2019WR025258, https://doi.org/10.1029/2019WR025258, 2020.

Zhang, D., Lindholm, G., and Ratnaweera, H.: Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring, J. Hydrol., 556, 409–418, https://doi.org/10.1016/j.jhydrol.2017.11.018, 2018.

Zhang, G. P.: Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing, 50, 159–175, 2003.

Zhao, J., Lange, H., and Meissner, H.: Gap-filling continuously-measured soil respiration data: A highlight of time-series-based methods, Agr. Forest Meteorol., 285, 107912, https://doi.org/10.1016/j.agrformet.2020.107912, 2020.