Response to referee comments

Reviewer 1

I am happy to see that the authors tested a simpler model architecture, as suggested in previous reviews. The paper has become shorter and more concise, which is great. It is worth remembering that this is intended as a technical note, rather than a full paper, but there are still some open questions from my side.

Reviewer Comment 1.1 — Conceptually, I don't understand why the manuscript is framed as gap filling, when the setup of the model matches the common setup of forecasting (i.e. having historic observations + additional inputs to predict the next time step). From my point of view, gap filling is a task that is performed on historic data records where observations of both sides of the gap are available. This raises the question why you would not make use of this additional information, since interpolating between two points is most likely easier than predicting into the future. If the framing of the manuscript should indeed be focused around gap filling, I think it might be worth discussing the decision for running a gap filling model in this forecasting setup.

Response:

We thank the reviewer for the positive assessment on the work we have done to improve the manuscript in our last revision. We framed our manuscript as gap filling because that is the purpose of our work. We recognize that multiple approaches can be applied for this purpose using the information before the gaps only or involving data from both sides of the gaps, as we described in the introduction, including interpolation, extrapolation, regression, predictive models and so on. In this technical report, we explore a forward forecasting approach for filling in the data gaps for a fair comparison between the LSTM models and the ARIMA models because ARIMA does not use observations after the gap for its estimations. We explicitly state this in the introduction as suggested by a reviewer in previous revisions.

We also recognize that Bi-directional LSTMs could be used to frame the gap filling as an interpolation problem as we discussed in the conclusion as a potential research topic for the future because non-trivial effort is needed to build, train and evaluate the bi-directional LSTMs.

Reviewer Comment 1.2 — 2. Why are ARIMA and LSTM not treated equally? That is, why does the ARIMA model predict all missing values at once, while the LSTM predicts only one time step at a time (and is then re-run for the next time step with the previous prediction filled into the input sequence). I think both models could be set up similarly and I wonder if you ever tested this and if the results suggested that this is the optimal setting for both models.

Response: We have revised the ARIMA predictive settings to match those of LSTM models for both single-well and multi-well models by following reviewer's suggestions, i.e., they predict the next time step immediately after the input window and the input window is sliding hour-by-hour to fill the entire length of a gap. We also searched a wider range of input windows for the ARIMA models for optimal performance per reviewer's another comment below. Please see the details for the response to comment 1.3. These changes did lead to slightly improved performance for ARIMA in terms of reducing the relative error outliers although all the conclusions on relative performance between LSTM and ARIMA stay unchanged. The new results have been updated for Figures 7, 8, 9 and 10 in the revision.

Response to referee comments

Reviewer 2

The paper explores in relative novel tools from Deep Learning LSTM models for reconstructing time series data. The potential of the paper describes well a setup and have improved significantly from the previous version. Now, it is more clear the problem, motivation and complexity in the modelling system presented. However, I still find the formulation somehow unfair in introducing spatiotemporal information. The comparison with ARIMA is 1D, but the LSTM has been fed with Spatial information (2D) so would question the idea of fair input information in both models. Now from the perspective of time series, the work is valuable and shows clear contribution on the trade-offs in complexity and usefulness of the LSTM and ARIM for filling gap (in special cases). It is to highlight that the concept here is as a technical note, and the case study is not so common and the very complex in nature. This added to the large availability of data, makes it an important problem to solve with LSTM, and this is an important aid to the area of Deep Learning and appllication cases of LSTM. I think is worth to share in this journal.

Aside from the above, maybe I would like to comment on small issues that I hope can be commented or updated if the paper is published.

Response:

In Section 4.3, we did train a multi-well ARIMA model and a multi-well LSTM model. Both types of multi-well models used data from wells 1-1, 1-10A, and 1-16A to estimate the SpC for well 1-1. We compared the multi-well models to the their single-well counterparts in Figure 9, along with a statistical analysis in Table 3 that has been added in the revision.

Minor comments are in the abstract

Reviewer Comment 2.1 — Even the process followed an Hyperparameter optimization, it is important to describe the ranges and sequence of such pipeline of optimal steps.

Response: We used a grid-search approach to find the optimal LSTM configuration for a given gap-length at each well. This involved iterating over all combinations of 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 have updated section 3.1.1 to specifically state this.

Old statement: "We performed a hyperparameter search to explore different LSTM model configurations, including the input time window size M, the number of units (U) in the LSTM layer, and the learning rate (L) at each well."

New statement: "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 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"

Reviewer Comment 2.2 — The overall LSTM input output structure for all optimals might be better understood in a table, where the performance and AIC are shown.

Response:

We have added a table in the online supplemental material (Table S1) showing the optimal LSTM configuration for a given gap length at each well, along the the models MAPE score for the gap length,

the AIC for the model on the validation set, and range of AIC scores for all models for a given gap length and well on the validation set. The table is shown below (Table 2.1).

We have also added a similar table in the online supplemental material for the optimal multi-well LSTM configurations (Table S3). The table is also shown below (Table 2.2),

Table 2.1: The best LSTM configurations and performance for a given gap length at each well based on the validation data set (2011): the input time window size (M), the number of units (U) in the LSTM layer, the learning rate (L), the SpC MAPE score, the Akaike Information Criterion (AI) for the model on the validation set, and range of AIC scores for all models for a given gap length and well on the validation set.

Well	**Gap Length**	**M**	**U**	**L**	**MAPE**	**AIC**	**AIC Min**	**AIC Max**
1-1	1	96	128	1e-3	0.189	1.31×10^4	-1.13×10^5	2.63×10^4
	6	96	128	1e-3	0.701	3.50×10^4	-9.08×10^4	4.19×10^4
	12	72	32	1e-4	1.24	-8.11×10^4	-8.11×10^4	5.40×10^4
	24	72	32	1e-4	1.95	-7.33×10^4	-7.33×10^4	6.50×10^4
	48	144	32	1e-5	2.85	-6.77×10^4	-6.80×10^4	7.50×10^4
	72	24	64	1e-5	3.67	-3.89×10^4	-6.42×10^4	8.08×10^4
1-10A	1	120	128	1e-3	0.19	4.75×10^4	-7.82×10^4	5.58×10^4
	6	120	128	1e-3	0.685	6.24×10^4	-6.29×10^4	7.39×10^4
	12	120	128	1e-3	1.1	6.69×10^4	-5.82×10^4	8.07×10^4
	24	144	128	1e-3	1.63	7.23×10^4	-5.22×10^4	8.70×10^4
	48	168	64	1e-5	2.16	-2.35×10^4	-4.90×10^4	9.14×10^4
	72	168	64	1e-5	2.39	-2.22×10^4	-4.63×10^4	9.37×10^4
1-15	1	48	32	1e-3	0.0163	-1.31×10^5	-1.32×10^5	1.93×10^4
	6	48	32	1e-3	0.0521	-1.04×10^5	-1.10×10^5	2.68×10^4
	12	48	32	1e-3	0.109	-1.02×10^5	-1.02×10^5	4.18×10^4
	24	96	128	1e-4	0.229	3.26×10^4	-9.25×10^4	5.70×10^4
	48	144	64	1e-3	0.372	-5.27×10^4	-8.50×10^4	7.15×10^4
	72	144	64	1e-3	0.506	-4.98×10^4	-8.08×10^4	8.00×10^4
2-2	1	168	32	1e-3	0.49	-6.06×10^4	-7.54×10^4	6.52×10^4
	6	168	32	1e-3	1.62	-4.86×10^4	-5.66×10^4	7.75×10^4
	12	144	128	1e-5	3.07	8.00×10^4	-4.95×10^4	8.47×10^4
	24	144	128	1e-5	4.42	8.34×10^4	-4.54×10^4	9.15×10^4
	48	144	128	1e-5	6.61	8.69×10^4	-4.19×10^4	9.63×10^4
	72	144	128	1e-5	7.52	8.66×10^4	-4.01×10^4	9.85×10^4
2-3	1	168	64	1e-3	0.142	-8.62×10^4	-1.16×10^5	2.28×10^4
	6	168	64	1e-3	0.369	-6.95×10^4	-9.87×10^4	4.64×10^4
	12	168	64	1e-3	0.663	-5.97×10^4	-8.78×10^4	6.02×10^4
	24	24	64	1e-5	1.09	-5.59×10^4	-7.92×10^4	7.59×10^4
	48	48	64	1e-5	1.7	-4.82×10^4	-7.18×10^4	8.95×10^4
	72	48	64	1e-5	2.28	-4.35×10^4	-6.72×10^4	9.51×10^4
2-5	1	120	32	1e-3	0.109	-1.15×10^5	-1.18×10^5	2.51×10^4
	6	144	64	1e-4	0.332	-6.72×10^4	-9.56×10^4	5.49×10^4
	12	144	64	1e-4	0.586	-5.70×10^4	-8.46×10^4	6.89×10^4
	24	144	64	1e-4	0.999	-4.90×10^4	-7.66×10^4	8.08×10^4
	48	120	64	1e-5	1.39	-4.51×10^4	-7.12×10^4	8.92×10^4
	72	120	64	1e-5	1.77	-4.16×10^4	-6.76×10^4	9.19×10^4

Table 2.2: The best multi-well LSTM configurations and performance for a given gap length based on the validation data set (2011): the input time window size (M), the number of units (U) in the LSTM layer, the learning rate (L), the SpC MAPE score, the Akaike Information Criterion (AI) for the model on the validation set, and range of AIC scores for all models for a given gap length on the validation set.

Gap Length	**M**	**U**	**L**	**MAPE**	**AIC**	**AIC Min**	**AIC Max**
1	144	128	1e-3	0.154	4.56×10^4	-8.40×10^4	6.11×10^4
6	144	128	1e-3	0.63	6.27×10^4	-6.68×10^4	7.13×10^4
12	24	32	1e-4	1.14	-5.97×10^4	-5.97×10^4	8.08×10^4
24	24	32	1e-4	1.75	-5.43×10^4	-5.43×10^4	8.89×10^4
48	96	128	1e-5	2.69	7.90×10^4	-4.84×10^4	9.46×10^4
72	96	128	1e-5	3.15	8.09×10^4	-4.74×10^4	9.74×10^4

Reviewer Comment 1.3 — Looking at Figure 6: If you perform hyper parameter tuning and one model (almost) constantly picks the largest (or smallest) value, usually you should increase the search range, as this indicates that eventually even larger (or smaller) parameters would be better. Looking at e.g. the input window length of the ARIMA model.

Response:

We have expanded the search range of the ARIMA model input window from 192 to 504 hours in an increment of 24 hours for all wells to identify optimal input windows to make sure the optimal input window is not at the upper or lower bound of the search range, as shown in the updated Figure 6. In general, larger optimal windows are resulted, which has consequently led to improved performance. We also noticed that the performance improvement for input windows larger than 288 hours is marginal in terms of the MAPE for all gap lengths across the wells we tested.

Reviewer Comment 1.4 — To my understanding, part of this technical note is the benchmarking/comparison of two models, LSTMs and ARIMA. As such, I think this paper is still missing a statistical analysis of the modeling results. The entire discussion is currently focussed around a few plots and a textual description of what one can see in these figures. However, to a certain degree I would argue that this analysis/interpretation is rather subjective. I think it would benefit the paper to have a table that compares both models on a range of different metrics, including statistical tests of e.g. the robustness/significance of the results. Right now, I wonder what the takeaway message of this paper is. I would argue that it was probably known that both models, LSTMs and ARIMA, are generally capable of time series forecasting. If I would be a user with similar data or a similar problem, what is the additional knowledge that I can gain from reading this paper?

Response:

We have added two table comparing the LSTM and ARIMA models on several statistics: one comparing the performance of the LSTM and ARIMA models filling in gap lengths of 24 hours (same models in Figure 8) that is added to the main paper in section 4.1 (Table 1.1 in this response, Table 2 in the revision), and another comparing the two approaches for all gap lengths that is added to the supplemental material (Table 1.2 in this response, Table S2 in the supplemental material). In addition, we have added a table of statistics for the models in Figure 9 to section 4.3, comparing the single and multi-well ARIMA and LSTM models (Table 1.3 in this response, Table 3 in the revision)

We have also added the following statistical analysis of the two models in Section 4.1:

"In addition to the relative errors, we calculated the MAPE, Root Mean Squared Error (RMSE), Nash–Sutcliffe model efficiency coefficient (NSE) [Nash and Sutcliffe, 1970], and Kling-Gupta Efficiency (KGE) [Gupta et al., 2009] for the best LSTM and ARIMA model per gap length to compare the two approaches. Table 1.1 compares the performance of the LSTM and ARIMA models filling in gap lengths of 24 hours. The table for all gap lengths is in the online supplemental material (Table S2).

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 mean of the observations, and less than 0 indicates that the model is a worse predictor than the mean of the observations. NSE is calculated on the SpC predictions by the following equation:

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

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

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

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{2}$$

$$r = \frac{\mathrm{cov}(\mathrm{O}, \mathrm{P})}{\sigma(\mathrm{O}) * \sigma(\mathrm{P})} \tag{3}$$

$$\alpha = \frac{\sigma(\mathrm{P})}{\sigma(\mathrm{O})} \tag{4}$$

$$\beta = \frac{\mu(\mathrm{P})}{\mu(\mathrm{O})}, \tag{5}$$

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

The LSTM and ARIMA models yielded comparable average metrics at all the wells for the gap length of 24 hours, as can be seen in Table 1.1. The NSE and KGE resulted from both models are close to 1 for all the wells with negligible differences between the two models. The difference in MAPE and RMSE is also small, with relatively more notable differences for wells 2-2 and 2-3, where the ARIMA models resulted in lower MAPE and RMSE."

In addition, we updated the analysis of the multi-well models in section 4.3 of the revision as follows:

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

Minor comments are in the abstract

Reviewer Comment 1.5 — Figure 4: I think this figure is misleading to someone unfamiliar with LSTMs. You actually drew a fully connected network, rather than a recurrent (sequential) neural network. As of now, it seems like all inputs go into the LSTM at once (no time steps are visible in this figure), and the outputs of all time steps (since on the left side the timesteps are top

Table 1.1: Comparison of single-well LSTM and ARIMA models on 24-hour synthetic gap in the SpC data on the test set for each well. The models are the same ones used in Figure 8. The calculated statistics are: MAPE, Root Mean Squared Error (RMSE), Nash–Sutcliffe model efficiency coefficient (NSE), and Kling-Gupta Efficiency (KGE). The T-Score and P-Value are calculated on 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
	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
	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		

to bottom) are used to predict the output. Figure (b) is actually the more correct depiction of the LSTM and I don't understand why both visualizations are needed.

Response: We have modified Figure 4 to make it clear that the middle layer (yellow and green) are individual LSTM units and not a fully connected network. The point of Figure 4(a) is to show the overall structure of the model and Figure 4(b) shows how the input data is fed into an individual LSTM unit.

Reviewer Comment 1.6 — P 3, L 32 "at the 300 Area of the U.S. Department of Energy Hanford site". What is the 300 for?

Response:
Hanford's 300 Area is a U.S. Department of Energy (DOE) site where the fuel manufacturing operations occur. It is a numbered naming convention for the site. Details can be seen https://www.hanford.gov/page.cfm/300

Reviewer Comment 1.7 — P 8, L 25: "The well data were then pre-processed by normalizing all measurements via zero-mean and unit variance for each variable". You do not normalize "via" zero-mean and unit variance, you rather normalize to zero mean and unit variance. Normalizing via zero-mean and unit-variance sounds like you subtract a mean of zero and divide by a variance of one, which is hopefully not what you have done.

Response: We have updated the line to say "normalizing all measurements to zero mean and unit variance" to clarify.

Reviewer Comment 1.8 — P9 L 7ff You use plural for "multi-well models" throughout this passage but you only trained one multi-well model to predict at well 1-1, or?

Response: Yes, only one type of multi-well LSTM model is trained to predict at well 1-1 due to data availability. It is clarified in the text. "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"

Reviewer Comment 1.9 — P17 L 17: Which "auto.arima" function?

Response: The function we used in our study applies the Hyndman-Khandakar algorithm developed by Hyndman and Khandakar [2008] that minimizes the Akaike Information Criterion (AIC) to obtain an optimized ARIMA model. The following details have been added to section 3.2:

"The ARIMA models were built using the `auto.arima` function from the R package `forecast` [?], which applies the Hyndman-Khandakar algorithm [Hyndman and Khandakar, 2008] that minimizes the AIC to obtain the best-fit parameters of the ARIMA model."

Table 1.2: Comparison of single-well LSTM and ARIMA models for all synthetic gap lengths in the SpC data. The models are the same ones used in Figure 7. The calculated statistics are: MAPE, Root Mean Squared Error (RMSE), Nash–Sutcliffe model efficiency coefficient (NSE), and Kling-Gupta Efficiency (KGE). The T-Score and P-Value are calculated on the relative errors of the two models for each well and gap length.

Gap Length = 1 hr							
Well	**Model Type**	**MAPE**	**RMSE**	**NSE**	**KGE**	**T-Score**	**P-Value**
1-1	LSTM	0.117	7.76×10^{-4}	1.0	0.999	11.6	6.14×10^{-31}
	ARIMA	0.183	1.23×10^{-3}	1.0	1.0		
1-10A	LSTM	0.185	1.41×10^{-3}	1.0	0.997	−7.52	5.89×10^{-14}
	ARIMA	0.299	2.45×10^{-3}	0.999	0.999		
1-15	LSTM	0.0559	4.00×10^{-4}	1.0	1.0	−4.29	1.79×10^{-5}
	ARIMA	0.0548	4.34×10^{-4}	0.999	1.0		
2-2	LSTM	0.276	2.11×10^{-3}	0.999	0.997	7.13	1.01×10^{-12}
	ARIMA	0.451	3.89×10^{-3}	0.997	0.998		
2-3	LSTM	0.129	1.05×10^{-3}	1.0	0.999	2.88	3.96×10^{-3}
	ARIMA	0.2	1.97×10^{-3}	0.999	0.999		
2-5	LSTM	0.113	1.02×10^{-3}	0.999	0.998	−1.09	2.76×10^{-1}
	ARIMA	0.171	1.62×10^{-3}	0.999	0.999		
Gap Length = 6 hr							
Well	**Model Type**	**MAPE**	**RMSE**	**NSE**	**KGE**	**T-Score**	**P-Value**
1-1	LSTM	0.435	2.92×10^{-3}	0.999	0.995	13.8	4.46×10^{-43}
	ARIMA	0.461	3.29×10^{-3}	0.999	0.999		
1-10A	LSTM	0.589	4.03×10^{-3}	0.996	0.984	−15.4	2.08×10^{-53}
	ARIMA	0.653	4.86×10^{-3}	0.995	0.994		
1-15	LSTM	0.109	9.16×10^{-4}	0.997	0.999	−1.86	6.31×10^{-2}
	ARIMA	0.0747	6.43×10^{-4}	0.999	0.999		
2-2	LSTM	0.981	7.14×10^{-3}	0.989	0.985	4.58	4.79×10^{-6}
	ARIMA	1.01	8.29×10^{-3}	0.984	0.992		
2-3	LSTM	0.517	4.15×10^{-3}	0.997	0.997	4.35	1.37×10^{-5}
	ARIMA	0.521	5.75×10^{-3}	0.993	0.993		
2-5	LSTM	0.314	2.64×10^{-3}	0.996	0.998	−5.03	4.98×10^{-7}
	ARIMA	0.352	3.13×10^{-3}	0.995	0.995		
Gap Length = 12 hr							
Well	**Model Type**	**MAPE**	**RMSE**	**NSE**	**KGE**	**T-Score**	**P-Value**
1-1	LSTM	0.75	5.07×10^{-3}	0.997	0.996	16.0	2.05×10^{-57}
	ARIMA	0.781	5.47×10^{-3}	0.996	0.996		
1-10A	LSTM	0.947	5.96×10^{-3}	0.992	0.969	−19.2	1.88×10^{-81}
	ARIMA	0.947	6.32×10^{-3}	0.991	0.99		
1-15	LSTM	0.166	1.37×10^{-3}	0.994	0.996	−3.41	6.49×10^{-4}

	ARIMA	0.0893	8.16×10^{-4}	0.998	0.998		
2-2	LSTM	1.87	1.22×10^{-2}	0.967	0.981	39.4	0.00
	ARIMA	1.51	1.13×10^{-2}	0.971	0.985		
2-3	LSTM	0.98	7.78×10^{-3}	0.988	0.993	4.42	1.01×10^{-5}
	ARIMA	1.03	1.26×10^{-2}	0.966	0.974		
2-5	LSTM	0.569	4.60×10^{-3}	0.989	0.994	−6.96	3.51×10^{-12}
	ARIMA	0.566	4.86×10^{-3}	0.988	0.989		

Gap Length = 24 hr

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
	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
	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		

Gap Length = 48 hr

Well	Model Type	MAPE	RMSE	NSE	KGE	T-Score	P-Value
1-1	LSTM	2.13	1.21×10^{-2}	0.98	0.988	17.8	3.24×10^{-70}
	ARIMA	2.15	1.34×10^{-2}	0.976	0.988		
1-10A	LSTM	2.09	1.09×10^{-2}	0.974	0.911	−16.4	2.68×10^{-60}
	ARIMA	2.17	1.32×10^{-2}	0.962	0.981		
1-15	LSTM	1.0	6.51×10^{-3}	0.869	0.907	−44.2	0.00
	ARIMA	0.168	1.67×10^{-3}	0.991	0.995		
2-2	LSTM	4.64	2.80×10^{-2}	0.825	0.932	60.8	0.00
	ARIMA	2.95	2.04×10^{-2}	0.905	0.952		
2-3	LSTM	3.26	2.04×10^{-2}	0.915	0.919	29.7	4.62×10^{-189}
	ARIMA	2.89	2.38×10^{-2}	0.88	0.91		
2-5	LSTM	2.34	1.22×10^{-2}	0.925	0.928	−33.9	3.18×10^{-243}
	ARIMA	1.25	1.09×10^{-2}	0.937	0.928		

Gap Length = 72 hr

Well	Model Type	MAPE	RMSE	NSE	KGE	T-Score	P-Value
1-1	LSTM	2.56	1.40×10^{-2}	0.974	0.983	26.7	4.91×10^{-154}
	ARIMA	2.57	1.46×10^{-2}	0.971	0.985		
1-10A	LSTM	2.58	1.32×10^{-2}	0.962	0.888	−17.8	5.10×10^{-70}
	ARIMA	2.84	1.75×10^{-2}	0.931	0.957		
1-15	LSTM	1.27	8.18×10^{-3}	0.794	0.845	−42.8	0.00
	ARIMA	0.211	1.79×10^{-3}	0.99	0.988		
2-2	LSTM	5.91	3.44×10^{-2}	0.736	0.914	80.0	0.00

	ARIMA	3.66	2.46×10^{-2}	0.861	0.931		
2-3	LSTM	4.8	2.90×10^{-2}	0.829	0.864	23.5	3.91×10^{-120}
	ARIMA	3.49	2.56×10^{-2}	0.862	0.901		
2-5	LSTM	3.08	1.55×10^{-2}	0.88	0.884	-31.5	9.51×10^{-212}
	ARIMA	1.47	1.06×10^{-2}	0.941	0.96		

Table 1.3: 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 ones used in Figure 9. Calculations are performed on the test data set for well 1-1 (year 2016). The calculated statistics are: MAPE, Root Mean Squared Error (RMSE), Nash–Sutcliffe model efficiency coefficient (NSE), and Kling-Gupta Efficiency (KGE). T-Score and P-Value 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	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		

References

Hoshin V. Gupta, Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1):80–91, 2009. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2009.08.003. URL https://www.sciencedirect.com/science/article/pii/S0022169409004843.

Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for r. Journal of statistical software, 27(1):1–22, 2008.

J.E. Nash and J.V. Sutcliffe. River flow forecasting through conceptual models part i — a discussion of principles. Journal of Hydrology, 10(3):282–290, 1970. ISSN 0022-1694. doi: https://doi.org/10.1016/0022-1694(70)90255-6. URL https://www.sciencedirect.com/science/article/pii/0022169470902556.