Response to referee comments

Reviewer 1

General note: The updated paper is has improved, and I happy with the answers to the comments and most of the changes performed.

However, on the abstract mainly, I still suggest some minor updates to reflect clearly and better the work. In general, there are so many places where the author misleads readers with terms like dynamic data, dynamic parameter, dynamic high-frequency data, complex dynamic, highly challenging situations, the great advantages, multiple dominant modes and others. It is challenging to follow with all these overused terms.

The source of the dynamics that lead to the high variability in the groundwater levels is not known, and therefore, trying to fill in data is indeed a challenge for research. There is not needed to oversell the complexity, and instead of helping the paper, it makes it cumbersome to read and very conflictive for people in the area. Terminology should be correctly used in a scientific publication.

Response: We appreciate the reviewer's overall positive assessment of our revised manuscript and the suggestion to improve the abstract. We have edited the abstract to make the important messages clear without overselling the complexity as the reviewer suggested. Please also see below our one-to-one responses to minor comments in the abstract for details.

Minor comments are in the abstract

Reviewer Comment 1.1 — What are data set with high-frequency dynamics? I believe you want to highlight the dynamics of a phenomenon reflected in data? I am not sure to understand the high-frequency term here?

Response: We reworded this to be more specific: This 10-year-long dataset contains hourly temperature, specific conductance, and groundwater table elevation measurements from 42 wells with various lengths of gaps.

Reviewer Comment 1.2 — The sentence: "We selected a location at the U.S. Department of Energy's Hanford site to demonstrate and evaluate the new method" not sure which new method?

Response: The new method here refers to the LSTM-based gap filling method. The abstract has been extensively revised to address various comments, and the phrase "new method" has been removed.

Reviewer Comment 1.3 — It is well known that wells data usually are spatially distributed (no two wells in the same place), so in many studies, the research address them as in a more precise way just saying temporal data from 42 wells.

Response: We have revised this sentence to be "This 10-year-long dataset contains hourly temperature, specific conductance, and groundwater table elevation measurements from 42 wells with various lengths of gaps. "

Reviewer Comment 1.4 — The sentence: "that monitor the dynamic and heterogeneous hydrologic exchanges between ". I would suggest making it more clear. Same as in point "a" of this comments.

Response: The long sentence has been rewritten to make it clearer. New Sentences:"In this study, we explore the ability of deep neural networks to fill in gaps in a spatially distributed time-series dataset from a well network deployed at the U.S. Department of Energy's Hanford site that monitors the dynamic and heterogeneous hydrologic exchanges between the Columbia River and its adjacent groundwater aquifer. This 10-year-long dataset contains hourly temperature, specific conductance, and groundwater table elevation measurements from 42 wells with various lengths of gaps"

Reviewer Comment 1.5 — The sentence: "capturing nonlinear, dynamic patterns in wells that exhibit various dynamics signatures", it seems trying to oversell the work. Dynamics everywhere?

Response: We added "river corridor" before the "wells" to refine the scope in river corridor where the dynamics are present in all wells. We also changed "nonlinear, dynamic patterns" to "temporal patterns".

Reviewer Comment 1.6 — The sentence: "Although the ARIMA models yield better error statistics, they fail to capture abrupt changes or high-frequency (daily and subdaily) variations in system states that are typical characteristics of a complex dynamic system." I guess there was a mathematical way to verify this statement? Maybe would be nice to elaborate in this line if this was just visual, or what was used to assess the so-called high-frequency variations?

Response: The high-frequency (hourly and subdaily) variations are assessed from the WPS spectrum plots in the second column of Figure 7. We have explained that concept in multiple places in the manuscript, and the sentence has been revised to "Our study demonstrates that the ARIMA models yield better average error statistics, yet they tend to have larger errors during time windows with abrupt changes or high-frequency (daily and subdaily) variations.".

Response to referee comments

Reviewer 2

General note: As in my previous review, I will mostly concentrate on the methodology of this manuscript, than on the data or context of this study in groundwater literature. I think that the revised manuscript overall improved by a lot over the original submission in terms of clarity and structure. The authors included information on the hyperparameter search (and made the search space larger than before) and chose the best model based on independent validation data. Although I am surprised by the result of the hyperparameter search, I guess this is something we have to accept. However, there is one point around the model architecture that I still do not see satisfiably answered (and largely ignored in the manuscript), see Comment 1 + Comment 2.

Response: We thank the reviewer for the constructive comments to improve our manuscript and for the overall positive assessment on our revision. We have taken extensive work to address the comments on the convolutional layer and the overall LSTM architecture. We have achieved improved performance with a much simpler architecture: a single-layer LSTM model without the convolutional layer. We provide the details in the one-to-one responses to your detailed comments below.

Reviewer Comment 2.1 — In my last review (Reviewer Comment 3.1d), I questioned the use of the convolutional layer. As the layer is used currently, the most recent time steps are ignored for doing the gap filling. To be more concrete: Given an input sequence of length M with N consecutive time steps, which should be gap filled. The first of the N time steps is immediately following the last time step of M. In most autoregressive tasks, the immediate preceding time steps are the most important features, especially with time series of high temporal frequencies. However, due to the choice of the convolutional layer and the filter size, the first day of N is only predicted by the M-N+1 first time steps of M, ignoring probably the most important information. The ARIMA model however, does see these time steps (and performs much better than the LSTM). From the hyperparameter search as described in Section 3.1.1. It does not seem as if the convolution layer at all was optimized. And even if the authors decide to keep this architecture, I think this is a critical point to include into the paper and to explain their decision. I could imagine that people who see this (that the most recent days are ignored in an autoregressive task) will ask why. The answer of the authors in their rebuttal (" Furthermore, the time steps immediately preceding the current time are not necessarily the most informative information in the presence of dynamical behavior ") might be true, but should definitely be tested as well as discussed in the manuscript.

Response:

We agree with the reviewer that the application of the convolutional layer limits the use of most recent time steps in predicting the future states. As a result and based on the Reviewer comment 2.2 (and previous comments), we have updated the model to a single LSTM layer and a dense layer to forecast N=1 time step immediately following the input time window. This removes the convolutional layer and ensures the model is taking advantage of the most recent time steps. With the simplified model, we have updated the hyperparameter search to include the number of LSTM units and the learning rate. In addition, we drop the temperature measurement as input. Please refer to the new Figure 4 in the manuscript for the updated architecture.

We compared the performance between the best models using the original and the new simplified architectures. Figure 2.1 and Figure 2.2 show the model performance comparisons in different metrics between the two LSTM architectures on the validation dataset from year 2011 for single-well models by setting N=1. As seen in Figure 2.1, the single-layer LSTM models outperform the original architecture on MAPE for all wells except 1-15, where both perform comparably on gaps of 24, 48, and 72. We see a similar improvement in relative error distributions (Figure 2.2).

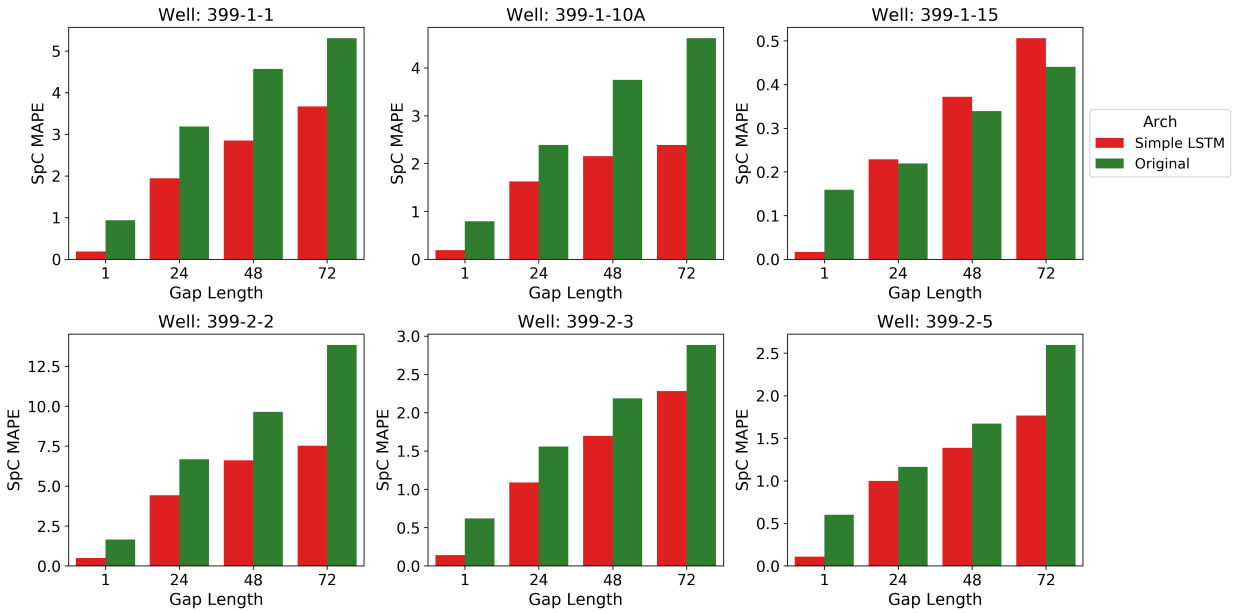


Figure 2.1: Comparison of the 2011 SpC MAPE scores of the best single LSTM model per gap-length of each well against the best original model per gap-length. The results for the new single LSTM are shown in red, and the results for the original architecture are shown in green.

Additionally, we trained multi-well models with the simplified architecture using the same procedure described in section 4.3 and compared them to the single-well modified architectures. Similar to the single-well models, the new multi-well models perform better than the original architecture across all gap-lengths by MAPE (Figure 2.3) and relative error distribution (Figure 2.4).

We modified the description on hyperparameter search accordingly in the manuscript under section 3.1.1.

Reviewer Comment 2.2 — This is very related to the comment above: The authors argued in their answer to Reviewer Comment 3.1.d that the (one) reason for the convolutional layer is to map from a sequence with M time steps to a sequence of N time steps. I don't know how this slipped my eye in my previous review, but an important question is "Why do you even map to N?". On Page 9 L1ff. you say you actually only map M to 1 and, then move M by one time step (integrating the last prediction into the shifted input sequence M) to predict the next time step and so one. So why is the LSTM-based model not trained to do exactly this? This setting is the most common LSTM setting (called sequence-to-one), and you would simply use the LSTM output at the last time step, to predict the next time step. During inference (= gap filling) you would do exactly what you do now: passing one sequence of M time steps through the model, get the prediction for the M+1 time

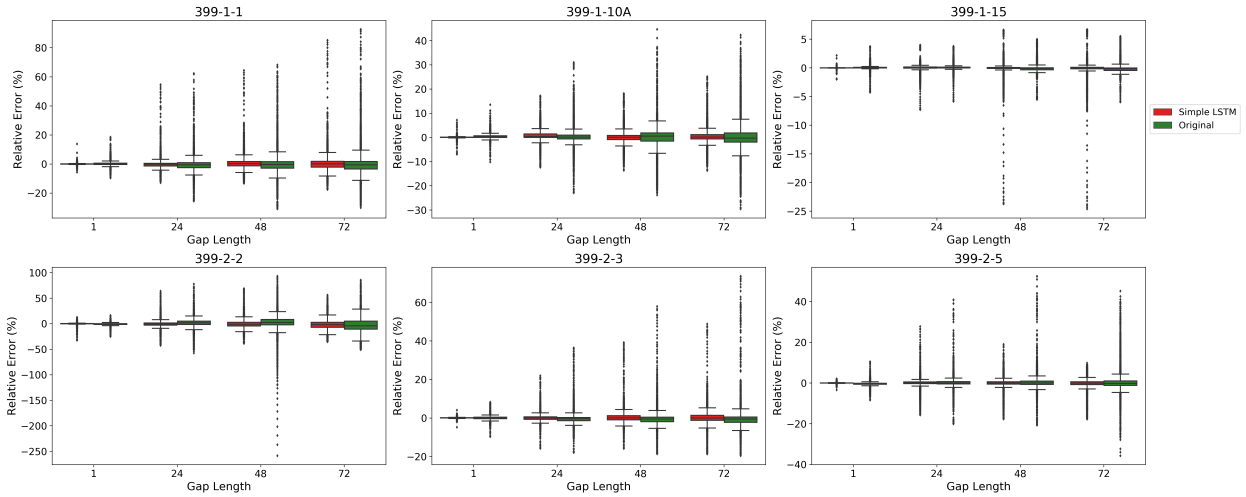


Figure 2.2: Comparison of the distribution of relative errors for 2011 of the best single-layer LSTM model per gap-length of each well against the best original model per gap-length. The results for the new single LSTM are shown in red, and the results for the original architecture are shown in green.

step, shift M by one time step and include the previous prediction, pass the new sequence again to get the prediction for the M+2 time step, and so on. The convolutional filter is also not, what makes you model account for spatial correlations (related to the answer of the authors to reviewer comment 1.4), since the LSTM can already account for those correlations. So the framing of the manuscript can remain unchanged.

Response:

Along with the changes we made to the LSTM architecture, we have followed the reviewer's suggestion to only map from M to 1 for gap filling. We also evaluated the performance of gap filling using M to N mapping (sequence-to-sequence) using the new architecture, which didn't perform as well as the M to 1 mapping. The new methodology section reflects both changes in architecture and in gap filling strategy.

Reviewer Comment 2.3 — I can not follow the conclusion in L7 P 15ff, especially that " ARIMA cap capture [...] but not changes that occur over a short time window (i.e., at higher frequencies) ". As the authors note themselves, ARIMA is better in every error statistic. It is argued that the LSTM does better at higher frequencies and it is pointed to Figure 9, the first two columns. The figures are small so it might be hard to see, but from what I can see, I don't see the LSTM being better in any well at any point during the entire period. The blue line, which shows the relative error, seems to be always worse for the LSTM, also during periods with higher variance. At this point, I can't see any evidence that backs the statement of the authors and I think, additionally to these plots, some quantification (using some metrics) are needed to support the statement that the LSTM has some advantage over the ARIMA model.

Response: We improved Figure 9 (Now Figure 8 in the revise manuscript) by combining the predicted time series by both the LSTM and ARIMA approaches in one plot so that the differences between them

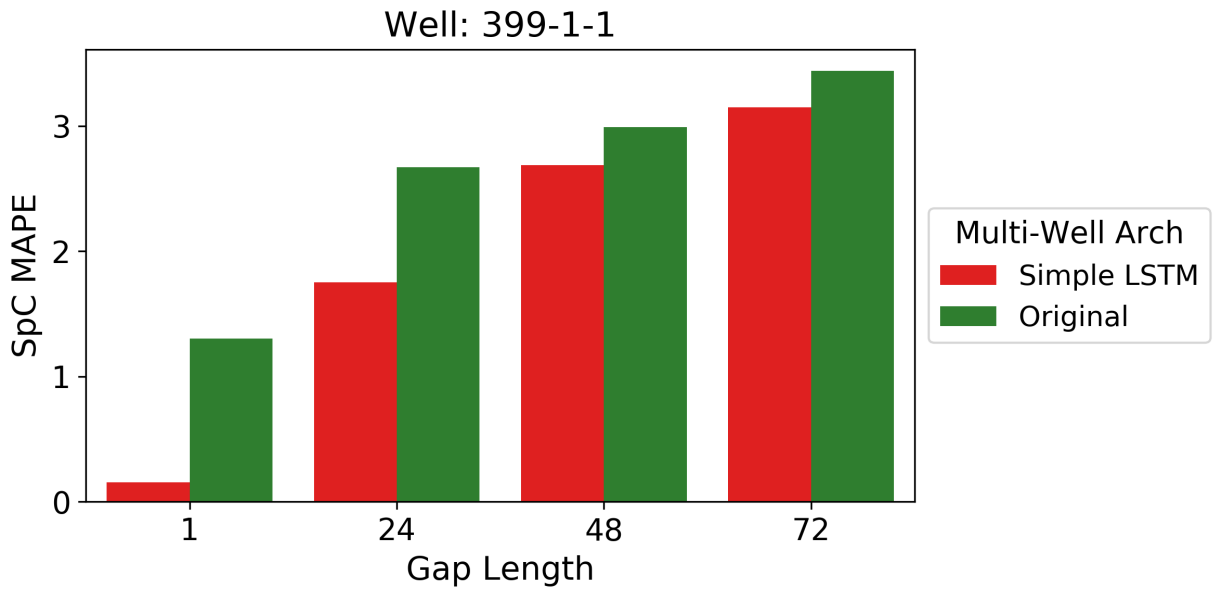


Figure 2.3: Comparison of the 2011 SpC MAPE scores of the best multi-well single-layer LSTM model per gap-length of each well against the best original multi-well model per gap-length. The results for the new multi-well single-layer LSTM are shown in red, and the results for the original multi-well models are shown in green.

can be better visualized. We moved the relative error plots to the supplemental material as they still provide a different perspective in comparing these two methods. By doing so, we were also able to significantly enlarge the time series plot to further help visualize the differences. The new Figure 9 is shown in Figure . It is now clear that there are unrealistic spikes in ARIMA predictions over multiple times in all the wells.

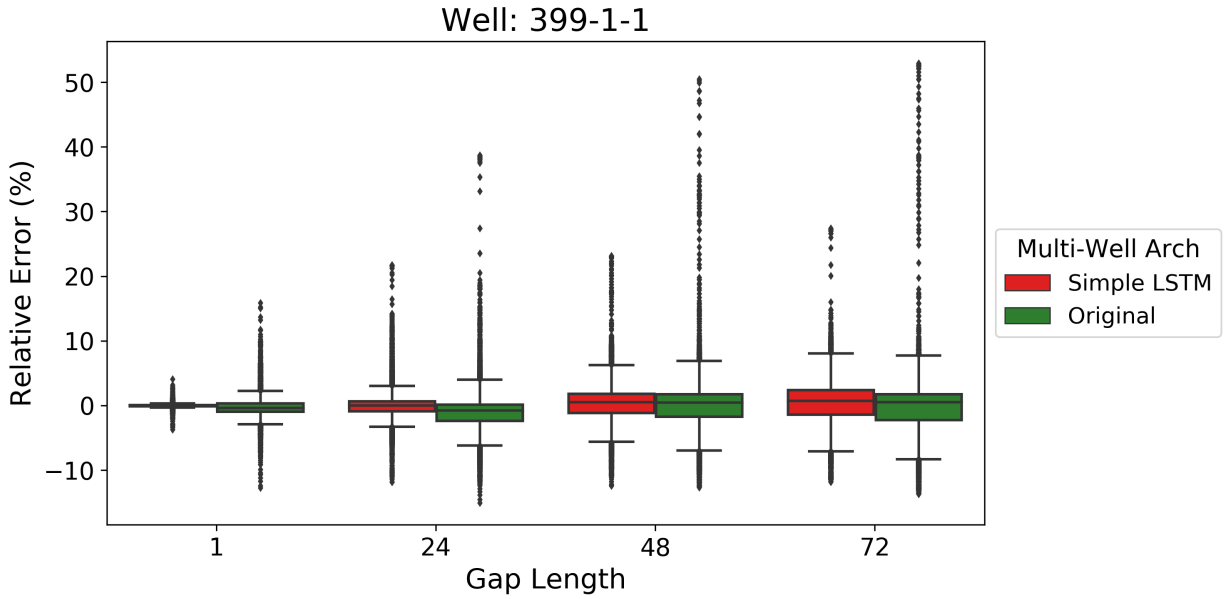
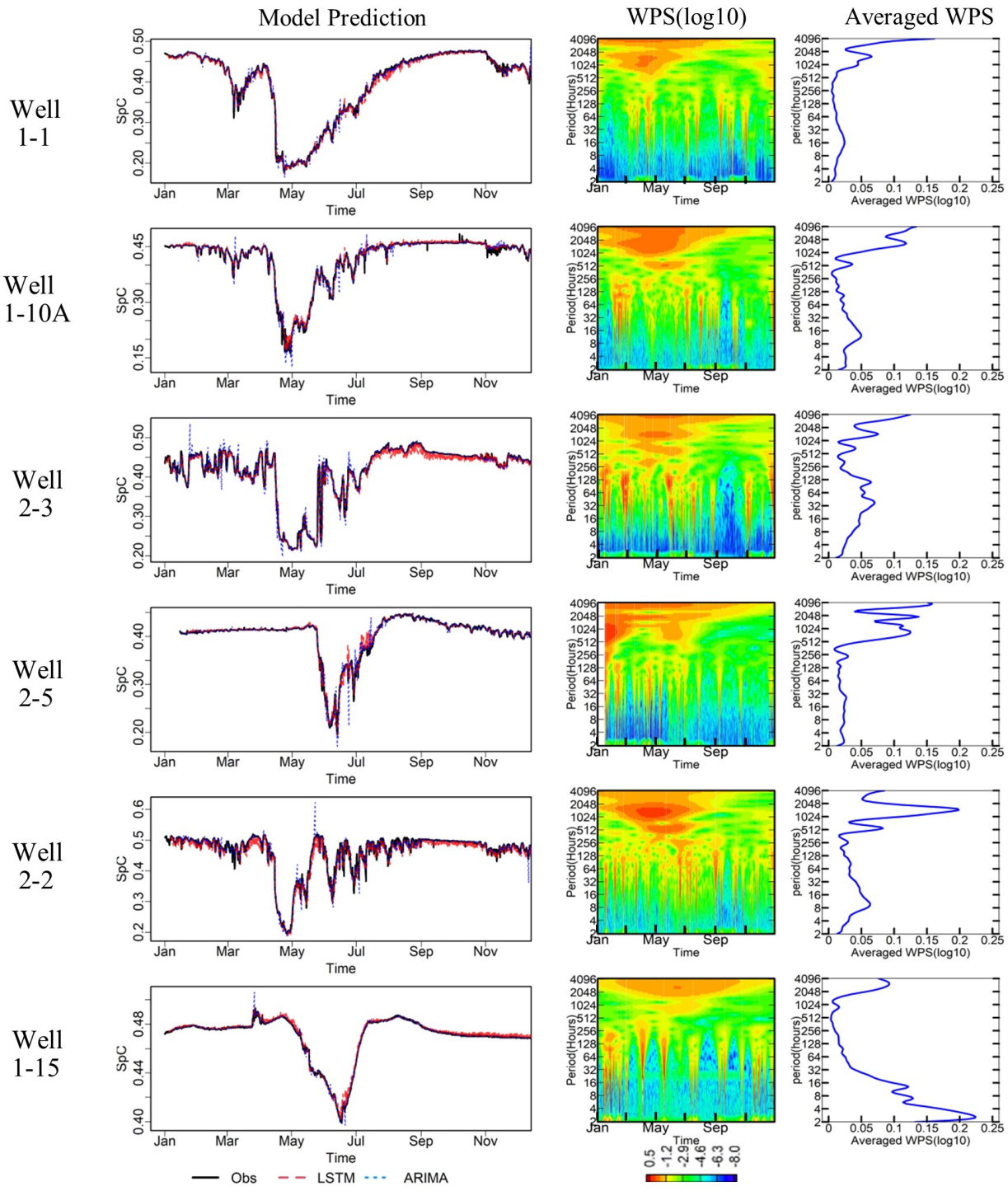


Figure 2.4: Comparison of the distribution of relative errors for 2011 of the best multi-well single-layer LSTM model per gap-length against the best original multi-well model per gap-length. The results for the new multi-well single-layer LSTM are shown in red, and the results for the original multi-well models are shown in green.


Model Prediction
WPS(log10)
Averaged WPS
Well 1-1
Well 1-10A
Well 2-3
Well 2-5
Well 2-2
Well 1-15
SpC
Time
Jan
Mar
May
Jul
Sep
Nov
Period(Hours)
period(hours)
Averaged WPS(log10)
Obs
LSTM
ARIMA

figureRevised Figure 9 (Now Figure 8 in the revise manuscript): Columns 1 shows time series of model predictions from LSTM (in red) and ARIMA (in blue) methods, respectively, assuming 24-hour synthetic gap in the SpC data, compared with the observations (in black). The best model configurations were used for all models. The testing data come from year 2016 for wells 1-1, 1-10A, 2-2, and 2-3, from year 2017 for well 1-15 and from 2008 for well 2-5. Column 2 is the spectrogram of each well and column 3 is the WPS averaged over for the corresponding year.

Reviewer Comment 2.4 — P18 L11: "significantly" I agree that the improvement seems obvious, however, the use of significant should always be supported by the result of a significance test. Otherwise, maybe rephrase this sentence.

Response: We have rephrased the sentences by removing "significantly" to avoid any confusion.

Reviewer Comment 2.5 — Isn't it possible to train a multi-well ARIMA(X) model as well? This would be an interesting benchmark for the experiment in Section 4.3, since in the single site the ARIMA model showed superior performance. If the LSTM would be better in the multi-well setting, this would certainly be an interesting result.

Response: Following the suggestion, we have built multi-well ARIMA models by following the similar setup with multi-well LSTM models, e.g., adding the information from well 1-10A and 1-16A in ARIMA model as extra regressors. The boxplots of relative errors under different gap lengths for well 1-1 has been listed in Figure 2.5 (same as Figure 9 in the revised manuscript). This comparison did reveal interesting results on how ARIMA can only benefit from additional spatial information for small gaps, while the LSTM method can achieve considerable performance improvement by including neighboring wells for filling larger gaps. We revised the discussion in Section 4.3 accordingly.

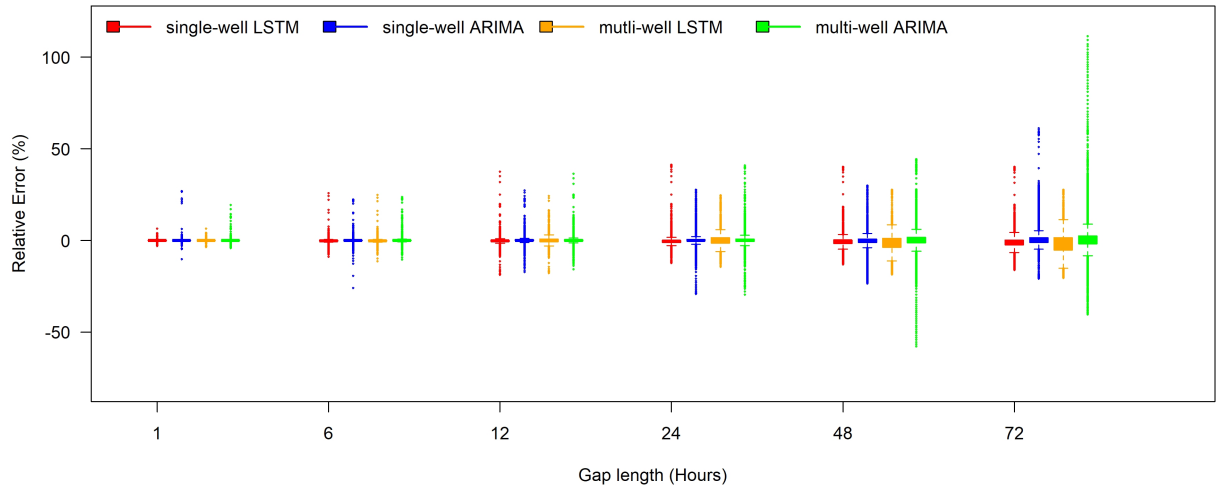


Figure 2.5: The boxplot of relative errors for filling gaps lengths of 1, 6, 12, 24, 48, and 72 hours for single-well LSTM model, single-well ARIMA model, multi-well LSTM model, and multi-well ARIMA on well 1-1.

Reviewer Comment 2.6 — The two sentences in P19 L3-4 seem to contradict each other. " LSTM excels in dealing with high-frequency dynamics (daily and subdaily) or nonlinearities, although they require more training data and computational resources. The LSTM approach also appeared to overestimate the high-frequency (daily and subdaily) fluctuations in some wells near the river (i.e., wells 1-1, 1-10A, and 2-2), which was likely caused by the variability in dynamics signatures among the training, validation and test periods. ". They " excel " but " also appear[ed] to overestimate ".

Response: We agree that we didn't elaborate well on this seemingly contradictory statement. We have rephrased the sentence to "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 lacks such dynamics."