Response to referee comments

Reviewer 2

General Remarks: The paper presents an interesting use of deep learning with LSTM Networks for infilling groundwater data. The article is timely and tries to make a comprehensive description and explanation of how the Deep learning technique is implemented using statistical and machine learning techniques. The paper is a welcome contribution to the field of groundwater and hydrological earth sciences. However, I cannot recommend publication in the present form due to the comments and questions raised. The paper needs major revision.

Response: Thanks for the reviewer's careful consideration of our manuscript and positive assessment. We address individual comment below to improve our manuscript for possible publication.

Reviewer Comment 2.1 — The paper states that long-term spatiotemporal changes in subsurface hydrological flow is usually quantified using a network of wells. However this paper does not deal with the long-term trend or analysis. Hourly data is hardly interpreted or used for the long term. Hourly information for sure contains noise that would be advisable to remove for the long term analysis.

Response: There seems to be some confusion between long-term changes and long-term trends. We will rephrase to clarify that monitoring networks capture the dynamic system behaviors over long time windows, which allows the discovery of signals over a spectrum of time scales as revealed by the spectral analysis we presented in the manuscript. While long-term trends, e.g., low-frequency variations, are usually smoother and could be captured by existing time series analysis method like ARIMA (see our comparison analyses between LSTM- and ARIMA-based gap filling results), our focus is on capturing high-frequency dynamics that are important signatures for understanding managed systems. We will clarify those points in our objectives and reiterate in conclusions to avoid confusion.

Reviewer Comment 2.2 — Observations are mentioned to be spatially sparse, and temporal gaps exist. Many papers have solved the same type of problem, without using the term spatiotemporal. Almost every course in hydrology deals in one chapter with the issue of using spatial correlation and temporal correlation to fill in data. So in this respect, the authors are invited to clearly indicate what innovation is brought by this work to spatiotemporal analysis.

Response: We agree with the reviewer that a lot has been done in hydrology for spatial and temporal analyses. However, there have been very few studies that address spatial and temporal correlations simultaneously due to the difficulty in parameterizing the spatial and temporal correlations all together. Deep neural networks provide an alternative way to represent such correlations without assuming the explicit form of correlations a priori, which is the innovation our work originally aimed to bring and demonstrate. However, we found there are multiple steps towards accomplishing that goal as it involves merging two types of deep neural networks to represent both the spatial and temporal correlations and evaluate various configurations thoroughly. Related to the comments Reviewer #1 provided, we will be focusing on the temporal component of the data and use LSTM for capturing multi-scale signatures, which we feel is appropriate scope for a technical note. One unique advantage of using LSTM to represent temporal correlations is that we do not pre-assume a correlation form. We will make this clear

in the discussion. We will also add spatio-temporal analyses as our next step to achieve the ultimate goal.

Reviewer Comment 2.3 — Following point two, it is known that in most of the cases, aquifers with little or no human intervention have low variability. Conventional guidelines and measures in hydrogeological science are typically based on monthly data.

Response: It is true that some aquifers with no or little human intervention have low variability, for which monthly data could be sufficient to understand the system behavior. However, anthropogenic activities, in particular, dam operations, have increasingly impacted the river and aquifer systems by altering the exchange patterns between river water and groundwater, and the associated thermal and biogeochemical processes [Song et al., 2018, Shuai et al., 2019, Zachara et al., 2020]. Due to significant, high-frequency (hourly) stage variations caused by dam regulations to meet power generation needs, it is insufficient to use monthly data to understand such systems as have been demonstrated in numerous studies performed at our study site. Our study site is representative of many dam-regulated gravel-bed rivers across the world. Therefore, our study will have broader impacts to many other systems. We will make this point clear in our introduction and conclusion sections.

Reviewer Comment 2.4 — In the present paper the idea of nonlinear dynamics is mentioned almost everywhere in the introduction and justification of the work. This is somewhat surprising and needs better justification, since groundwater dynamics, in many cases, can be represented with linear models. As it is concluded in this paper results, ARIMA can approximate the system quite well.

Response: This comment is related to the earlier comment 2.3 as high-frequency dynamics lead to higher level of nonlinearity in system responses, especially in the specific conductance that is a result of mixing of water from various sources. We have shown that a linear model like ARIMA was not able to capture such nonlinearity, while LSTM could. We will explain this point in the revised manuscript. Please also refer to our response to comment 2.3 for the importance of capturing high-frequency dynamics for many dam-regulated systems, which will also be better articulated in the revised manuscript.

Reviewer Comment 2.5 — The particular case study presented here shows a relative complex dynamic nature indeed, but it seems it is due to human intervention (however I could be wrong). Can you comment on this and the uncertainties associated?

Response: The reviewer is partially right that human intervention contributes to the complexity of system behavior by creating high-frequency flow dynamics. However, the full complexity is a result of interactions between such human-induced variations and the natural heterogeneity of aquifer physical properties [Zachara et al., 2020]. There is significant uncertainty associated with aquifer physical heterogeneity at our study site as revealed by previous studies. We will add these additional discussion about the system complexity in the revised manuscript.

a The human intervention might affect your calculation and therefore, extractions might not be following a random but more human induced behaviour. So data understanding or replicability used in one year might not be the same in another. It would be advisable first to check how much and when extraction took place. Is this data filled in for a long term analysis, or short-term? This question arises since the hourly step is used.

Response: Please refer to our response to Comment 2.1 for explaining our use of long-term data versus long-term trend analyses. The reviewer is right that high-frequency flow variations are mainly caused by the dam operations while the seasonal and interannual variabilities are controlled by climatic forcing like precipitation and snow pack in the headwater systems. We will clarify the drivers of the high-frequency and low-frequency variations in the revision. We used multiple years of training data from dry, normal and wet hydrologic years to capture potential operational patterns associated with various conditions. LSTM units include a 'memory cell' that can maintain information in memory for long periods of time. We will look into the memory cells for more explanations.

b If indeed human intervention influence the dynamics of the groundwater system, the logical approach would be to find a variable to represent direct or indirect measurement of extractions.

Response: Thanks for the suggestion. Please refer to our response to your earlier comment. We will look into both memory cells and state cells to illustrate how and where the extraction occurs. We will expand our analyses accordingly.

c It is suggested to read the paper by Amaranto et al. (2018) "Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland". J Hydroinformatics, 20 (6): 1227–1246. DOI: `https://doi.org/10.2166/hydro.2018.002` and - Amaranto et al. (2019). A spatially enhanced data driven multimodel to improve semiseasonal groundwater forecasts in the High Plains aquifer, USA. Water Resources Research, 55, 5941– 5961. `https://doi.org/10.1029/2018WR024301`

Response: Thank you for the paper suggestions. Both of the papers listed above use data-driven approaches to improve groundwater forecasts. The MuMoC framework select neighboring wells to assist groundwater predictions is of our interest. Although the authors used data with coarser temporal resolution (daily or monthly) to make monthly predictions, which is different from our purpose of filling short gaps (up to 3 days) for capturing high-frequency dynamics, the idea of using information from neighboring wells applies to our case. We will include an additional architecture to enable multi-well setup. We will review and discuss these two papers in our revision.

Reviewer Comment 2.6 — The regional aquifer and geology might play a more significant role in the study, since not only the river but the size and other interventions and hydrometeorological recharges might be correlated.

Response: We agree that the regional aquifer and geology play an important role as shown in previous studies performed by our colleagues [Chen et al., 2012, 2013, Zachara et al., 2020]. The aquifer 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 and the heterogeneity within the Hanford formation. The understanding we developed from these earlier studies is that the physical heterogeneity contributes to the different response behaviours at different locations while the river stage

dynamics lead to multi-frequency dynamics in those responses. We will add more information to our system description to help readers understand.

Reviewer Comment 2.7 — The stations are so close, and the hourly variation appears to be periodic with an amplitude of 4 or 5cm, according to Figure 1 (and on other graphs). It is intriguing, the question I would have is what happens every hour? and if this hourly variation is noise on the measurement device or data? What is the precision of the measurement device? What is the volume of water extracted to reach the variation of 1 cm? Where the recharge water comes from(has this been studied in the past)? Is this 5 cm recharge volume feasible in one hour? Could be the water from the river affecting your measurements (interflow)? It is advisable to present the time series of the river flow. It would be also useful to have a few hydrological balances (note that this is a hydrological journal). The problematic still can be questioned due to its apparent complex dynamics with the river and human intervention (not a typical, natural aquifer).

Response: The reviewer is right that the water table elevation difference is small due to the close distance between wells and the highly permeable aquifer material (hydraulic conductivity in the range of 4000-7000 m/d). The rapid change in water table is at first due to pressure wave propagation from the river stage variation, and the recharge water comes from the river or displacement of groundwater from other parts of the aquifer depending on flow directions and locations of interest. Our pressure transducers (Stainless-Steel CS451 by the Campbell Scientific) is accurate enough to capture cm-scale variations. With a full scale (FS) of 102m,and resolution of 0.0035% FS which means with a full scale of 102m, the sensor will be able to detect a 0.375cm change in pressure. The measurement range of the pressure transducer set up in our study site is 0 to 10.2m, the standard accuracy is $\pm 0.1\%$ [Scientific] which leads the accuracy to about 1cm. In this case, the pressure changes with an amplitude of 4 5cm are the actual measurements and our consistent 10-year data has proven this point. In our revision, we will refer to numerous hydrologic modeling studies performed at the site to help readers understand the flow conditions and where the water comes from.

Reviewer Comment 2.8 — On the model setup, Please explain why you use Mx128.

Response: We use 128 (i.e. 128 units for each LSTM layer) because this number of units showed better performance after we experimented with different model architectures with different number of units. We will add this rationale to the manuscript.

Reviewer Comment 2.9 —

Page 7, line 10, mentions the supplemental material, but I cannot find it in the paper.

Response: The supplemental materials can be found using this link
https://www.hydrol-earth-syst-sci-discuss.net/hess-2019-196/hess-2019-196-supplement.pdf

Reviewer Comment 2.10 — Important: choice of (a very complex model) LSTM has to be justified, since it seems AR-type models is enough. Frankly, I don't see the need for complex models like LSTM, but if you have arguments to defend your position, please present them to convince the readers.

Response: We are interested in using an LSTM for this problem because LSTMs have had success in predicting values in time-series data without assuming explicit temporal dependence forms . We cite several examples of this on P3 L12. We wanted to explore whether an LSTM would provide improved performance over traditional methods (i.e. ARIMA). As far as we are aware of, there have not been other applications of LSTMs to groundwater well data. Related to our response to other relevant comments, we will make sure we clearly state the advantages of using LSTM in our revision.

Furthermore, there have been several recent applications of LSTMs to hydrology and earth sciences. Kratzert et al. [2018] used LSTMs to predict rainfall-runoff from meteorological observations. Zhang et al. [2018] used LSTMs for predicting and monitoring sewer overflow. Additionally, Fang et al. [2017] used LSTMs to predict soil moisture with high fidelity. We will update our paper (and references) to further explain our interest in applying LSTMs to this hydrological domain.

Old section (P3 L13): This makes LSTMs well suited for the problem at hand, particularly for data where multiple timescales of variability can affect responses [Liu et al., 2016, Song et al., 2017].

New section (P3 L13): This makes LSTMs well suited for the problem at hand, particularly for data where multiple timescales of variability can affect responses [Liu et al., 2016, Song et al., 2017]. Furthermore, LSTMs have been applied to earth science and hydrology domains. Kratzert et al. [2018] used LSTMs to predict rainfall-runoff from meteorological observations. Zhang et al. [2018] used LSTMs for predicting and monitoring sewer overflow. Additionally, Fang et al. [2017] used LSTMs to predict soil moisture with high fidelity.

Reviewer Comment 2.11 — On page 14, it states that other configurations of LSTM can be further explored; however, it is not clear why this was not done before. Not sure why the selected configuration was just tried to see if it works or not, without any analysis what is the best structure. This relates to comment 8 and 9.

Response: We acknowledge that we did not explain this point as well as we could have. We performed hyperparameter searches on: Number of LSTM layers, number of units per LSTM layer, number (and size of) dense layers, activation functions. This was performed for data on one well (399-1-1) with a smaller subset of input and output prediction windows, experimenting with different architecture configurations. However, this was not an exhaustive search of all possible configurations. We will add some additional description to this effect to the manuscript.

Reviewer Comment 2.12 — I am a bit in confusion how to interpret the statements made in conclusion. The ARIMA is not suited or less suited for filling high frequency (hourly, or short gaps) and more suitable for a long term period (24, 48 and 74 hours). It is suggested we need deep learning for filling high-frequency gaps (of one hour)?. Maybe is good to elaborate on the simplicity of what this translates to, I am not sure if the meaning is right.

Response: We acknowledge a potential source of confusion in terms of high-frequency fluctuations in temporal scale and the dominant frequency in the wavelet transform. In our study, we find that the ARIMA method would work well if the dominant frequencies in the wavelet transform are from seasonal cycles, whereas an LSTM could work better if the dynamics are dominated by daily and subdaily (high-frequency) fluctuations. We will better state this in the paper to avoid further confusion.

Reviewer Comment 2.13 — Not sure if there is an idea of how high is the overall error; in the figure 8, with well 1-15 it seems almost perfect representation (zero error in the validation data for many points). Also in the same well, it appears like high negative correlation up to 128 hours

Response: Well 1-15 has a different pattern than the rest of wells in that most of its higher intensities fall under lower frequency and its SpC correlates well with river stage towards longer periods and less persistent in time. Both ARIMA and LSTM have highly accurate predictions for well 1-15 (i.e., relative errors are about 1%), which is consistent with our conclusions that smooth changes in the observations are easier to be captured. In the wavelet spectral plot in column 3 of figure 8, where log10 of Wavelet Power Spectrum (WPS) is shown, the blue color under 128 hours still represent weak positive values.

References

Xingyuan Chen, Haruko Murakami, Melanie S. Hahn, Glenn E. Hammond, Mark L. Rockhold, John M. Zachara, and Yoram Rubin. Three-dimensional bayesian geostatistical aquifer characterization at the hanford 300 area using tracer test data. *Water Resources Research*, 48(6), 2012. doi: 10.1029/2011WR010675. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011WR010675`.

Xingyuan Chen, Glenn E. Hammond, Chris J. Murray, Mark L. Rockhold, Vince R. Vermeul, and John M. Zachara. Application of ensemble-based data assimilation techniques for aquifer characterization using tracer data at hanford 300 area. *Water Resources Research*, 49(10):7064–7076, 2013. doi: 10.1002/2012WR013285. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2012WR013285`.

Kuai Fang, Chaopeng Shen, Daniel Kifer, and Xiao Yang. Prolongation of smap to spatiotemporally seamless coverage of continental u.s. using a deep learning neural network. *Geophysical Research Letters*, 44(21):11,030–11,039, 2017. doi: 10.1002/2017GL075619. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL075619`.

F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018. doi: 10.5194/hess-22-6005-2018. URL `https://www.hydrol-earth-syst-sci.net/22/6005/2018/`.

Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

Campbell Scientific. Cs451:stainless-steel pressure transducer. URL `https://www.campbellsci.com/cs451`.

Pin Shuai, Xingyuan Chen, Xuehang Song, Glenn E. Hammond, John Zachara, Patrick Royer, Huiying Ren, William A. Perkins, Marshall C. Richmond, and Maoyi Huang. Dam operations and subsurface hydrogeology control dynamics of hydrologic exchange flows in a regulated river reach. *Water Resources Research*, 55(4):2593–2612, 2019. doi: 10.1029/2018WR024193. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024193`.

Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, volume 1, pages 4263–4270, 2017.

Xuehang Song, Xingyuan Chen, James Stegen, Glenn Hammond, Hyun-Seob Song, Heng Dai, Emily Graham, and John M. Zachara. Drought conditions maximize the impact of high-frequency flow variations on thermal regimes and biogeochemical function in the hyporheic zone. *Water Resources Research*, 54(10):7361–7382, 2018. doi: 10.1029/2018WR022586. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022586`.

John M. Zachara, Xingyuan Chen, Xuehang Song, Pin Shuai, Chris Murray, and C. Tom Resch. Kilometer-scale hydrologic exchange flows in a gravel-bed river corridor and their implications to solute migration. *Water Resources Research*, n/a(n/a):e2019WR025258, 2020. doi: 10.1029/2019WR025258. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR025258`. e2019WR025258 2019WR025258.

Duo Zhang, Geir Lindholm, and Harsha Ratnaweera. Use long short-term memory to enhance internet of things for combined sewer overflow monitoring. *Journal of Hydrology*, 556:409 – 418, 2018. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2017.11.018. URL `http://www.sciencedirect.com/science/article/pii/S0022169417307722`.