Imputation of Partially Observed Water Quality Data Using Self-Attention LSTM

Onatkut Dagtekin
Department of Computer Science and Technology
University of Hull
Hull, UK
O.Dagtekin-2019@hull.ac.uk

Nina Dethlefs
Department of Computer Science and Technology
University of Hull
Hull, UK
N.Dethlefs@hull.ac.uk

***Abstract*—Possible sensory failures on monitoring systems result in partially filled data which may lead to erroneous statistical conclusions which may affect critical systems such as pollutant detectors and anomaly activity detectors. Therefore imputation becomes necessary to decrease error. This work addresses the missing data problem by experimenting with various methods in the context of a water quality dataset with high miss rates. Compared models chosen make different assumptions about the data which are Generative Adversarial Networks, Multiple Imputation by Chained Equations, Variational Auto-Encoders, and Recurrent Neural Networks. A novel recurrent neural network architecture with self-attention is proposed in which imputation is done in a single pass. The proposed model performs with a lower root mean square error, ranging between 0.012-0.28, in three of the four locations. The self-attention components increase the interpretability of the imputation process at each stage of the network, providing information to domain experts.**

***Index Terms*—self-attention, imputation, recurrent neural network, water quality, missing data**

I. Introduction

Water has a significant role in environment and public health. Therefore continuous monitoring of water quality is crucial to detect pollution, to ensure that various natural cycles are not disrupted by anthropogenic activities and to assess the effectiveness of beneficial management measures taken under defined protocols such as the EU Water Framework Directive (WFD) and The Marine Strategy Framework Directive (MSFD). With increasing capability and low cost of sensors, constant monitoring has become widespread within research programmes providing high quality and in situ data. Partial or incomplete data may be returned from in situ monitoring networks due to external factors such as biofouling, electrical/mechanical failures or refinement of data due to quality assurance procedures, leading to misconstrued statistical analysis of the gathered data.

There are three different types of missing data; Missing completely at random (MCAR), Missing at random (MAR) and Missing not at random (MNAR). MCAR occurs when missingness does not depend on any variables. MAR occurs when missingness depends on the observed variables. MNAR occurs when missingness depends on both observed and unobserved variables. The occurrence of MAR indicates that the missing variables in a dataset can be derived from known ones by modelling the relationship between the missing and present variables. The challenge of imputation is the identification of the missingness mechanism and the possibility of multiple existence of missing mechanisms occurring together. In the context of water quality, the statistical methods show that the conditions of MAR are satisfied [9]. The model parameters for imputation can be learned by randomly omitting parts of complete samples and modelling the relationship between missing and known variables. After the imputation process, predictions about variables can be done for tasks such as pollutant detection in drinking water, early detection of sea snot or algal blooms in water bodies and water consumption monitoring.

There are different methods for addressing the problem of missing data imputation. The simplest solution would be to remove the rows of missing data. However, such a solution might affect the quality of the remaining data depending on the miss percentage and the temporal modelling of variables. There are simple methods such as mean/median imputation or constant/zero imputation. Multivariate solutions include Multiple Imputation by Chained Equations (MICE) and regression [14], [28]. With the increasing popularity and availability of deep learning and machine learning models, models such as random forests (RFs) and neural networks (NNs) are used for imputation as well [15], [31].

The following work proposes a novel architecture that uses a self-attention component in combination with Long Short-Term Memory (LSTM) to improve the effectiveness and interpretability of data imputation in the context of water quality. The proposed model is compared to different imputation methods; mean imputation, MICE with Bayesian ridge regressor [29] and k-Nearest Neighbour (kNN) [11], Generative Adversarial Networks (GAN) [38], Variational Autoencoders (VAE) [5], Recurrent Neural Networks (RNN) [40]. These models were chosen to observe the performance for the task of imputation under different assumptions of data distribution and data modelling. VAE and Bayesian Ridge assume that the data is normally distributed and model the data with Bayesian probability. GANs aim to learn the latent distribution of data using Nash Equilibrium. k-Nearest Neighbors (kNN) uses distance as a similarity metric for imputation. Recurrent models expose the temporal properties of the data. The proposed model outperforms the baselines in three of the four sites.

Neural network models are black-box processes by default where there is no information given about the prediction process due to the sheer number of calculations. This results in the reduction of interpretability of the process by non-experts. The self-attention component gives insight into how samples interact with each other at different stages of the network, increasing interpretability, as opposed to other neural network models and guides the model to increase its performance. The model is also tested on three other locations with different properties to test the generalisability of the model. The model performs imputation with a single pass imputing multiple variables.

II. Related Work

Deep learning and machine learning methods have been used extensively for imputation. Zhang et al. [39] use kNN and linear regression for imputation. Folguera et al. [8] use Self-Organizing Maps (SOMs) for imputing missing variables. Mulia et al. [24] use artificial neural networks (ANN) in combination with a Genetic algorithm for imputation. Autoencoders have been extensively applied to the task of data imputation in several domains [2], [5], [34]. The transformer architecture has been used for imputation as well [32]. Bansal et al. [1] uses a combination of kernel regression, convolutions, and multi-head attention for data imputation. GAN architectures have been used for data imputation [19], [21], [38]. Cao et al. [7] use recurrent components for imputation and assumes the missing values belong to the RNN graph. Variations of SVD have been used for imputation [6], [22], [35]. Spatio-temporal approaches have been used as well [20], [37]. Shu et al. and Papadimitriou et al. [27], [30] use PCA based approaches for data imputation. The aforementioned methods with the exception of the transformer architecture do not give insight into how the imputation process is done.

Imputation of missing data for the subject of water quality has been done in several approaches.Zhang et al. [40] use an encoder-decoder LSTM model with attention and sliding window approach for imputation. The model is further modified by using the differnet context vectors for different directions for the time series [41]. Kim et al. [16] compare ANN, SOM and a Soil and Water Assessment Tool to data from Taehwa River, South Korea. Rodriguez et al. [29] compare inverse distance weighting, RF regressor, Ridge regression, Bayesian Ridge (BR) regression, AdaBoost, Huber regressor, SVR and kNN regressor for data imputation for Santa Lucía Chico River, Uruguay. Tabari et al. [33] compare Multilayer Perceptron and Radial Basis Function networks in the context of water quality data imputation. Random forest and SVM have been applied to the task of imputation of water quality data [17]. Ratolojahanary et al. [28] compares RFs, Boosted Regression Trees (BRT), kNN and Support Vector Regression (SVR) using water quality data from Oursbelille, France. Nieh et al. [25] compare mean, median and multiple imputation in the context of microbial water quality data. Osman et al. [26] compare Gaussian Process Regression, Principal Component Analysis, Decision Trees, ANNs, Multiple Imputation and EM models.

The majority of the mentioned methods for water quality data imputation focus on improving the performance of the model for a single water body. The model we propose achieves better performance on different monitoring locations with different properties. The attention component used also provides information between the elements of input of the model from start to finish providing a different explanation than majority of the approaches. The testing of the model is done with eleven different values within the range of $[5\%, 95\%]$ miss rates. In previous work, the majority of the models are tested within the range of $[10\%, 80\%]$ miss rates or discrete values such as $20\%, 50\%, 80\%$ miss rates. Our experimentation setting reflects the real world phenomenon where datasets might have high miss rates and data become unusable.

III. Methods

A. Dataset

The data was collected by ESM2 and ESMx data loggers at four different moorings depicted in Figure 1. The data was collected as a part of ”The National Marine Monitoring Programme” (NMMP) to monitor eutrophication regarding ”The Convention for the Protection of the Marine Environment of the North-East Atlantic” (OSPAR) and ”Marine Strategy Framework Directive” (MSFD) assessments. The dataset was partitioned into four fractions based on location. Each of the locations is expected to have different characteristics due to their locations such that the Liverpool buoy is near a maritime route, WestGab is near wind farms, TH1 is near the mouth/delta of the Thames and Dowsing is in the open sea. Models may be able to expose these spatial differences between the buoys and the temporal properties in general. The periodicity and the relationship between the variables were analysed by [3], [4], [10] with varying date ranges and locations by performing wavelet analysis. The periodicities of variables depend on the season and range between 6 hours to 24 hours.

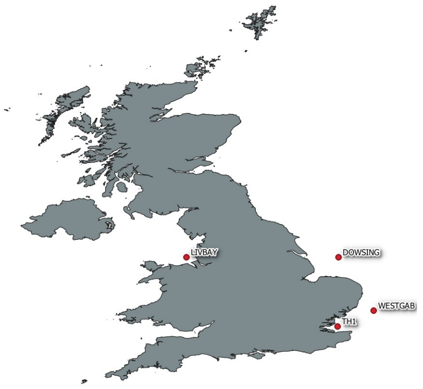

Fig. 1. Locations of moorings

Table I contains a summary of the whole dataset. Chlorophyll fluorescence is caused by algal activity through photosynthesis. Turbidity is the cleanliness of the water. Dissolved oxygen increases with photosynthetic activity and is used for respiration and decomposers. Salinity measures the concentration of salt in water. photosynthetically active radiation (PAR) is the light received by algae that can be used for photosynthetic activity. The data was collected at 30-minute intervals at each station between the dates 01/01/2009 and 04/08/2019. Base refers to the missingness in datasets between the observed dates. Before normalisation, PAR columns of the data was imputed with zero imputation with regard to the sunset and sunrise time according to the observation date.Due to sensor availability at certain times, chunks from the start and the end of some monitoring locations was removed. The operations result in 54.05% miss rate for TH1, 65.99% miss rate for LIVBAY, 56.39% miss rate for DOWSING, 56.59% miss rate for WESTGAB. Miss percentages reported relates to rows with at least one value missing.

B. Proposed Architecture

1) Imputation Model: The proposed model, named self-attention imputer (SAI), in Figure 2 uses the attention model introduced by [36] with the addition of LSTMs for temporal analysis and a linear layer. Similar to [7] and [41], a backward pass through the data is done but this is executed at the same pass using only the input batch. Instead of using a single self-attention component for a biLSTM layer exposing the periodical information known previously, using separate self-attention components enables the model to give different weights in the backward and forward direction. The self-attention component increases the interpretability of the neural network by assigning weights between samples given as input. Moreover, this entails the relationship between samples might not be linear depending on the missingness of the variables.

The model was based on the evidence that water quality data had MAR properties and the statistical analysis of periodicity [3], [4], [10] which justifies the use of LSTMs for this task. The data is normalized with min-max normalization and initially imputed with -1s before it is fed into the model. The models were tested with varying missing rates ranging from 5% to 95%. Compared baselines with their parameters (Earlystopping with a patience of 20 epochs with 10^{-5} tolerance was used during training for deep learning models):

- mean imputation
- MICE with kNN - k=25
- MICE with Bayesian Ridge regressor - # of iterations = 100, tolerance = 10^{-5}
- VAE - batch size = 32, Adam used as optimizer with learning rate = 10^{-4}, two linear layers with ReLU activation for encoder and decoder. Hidden size of four for μ and σ.
- GAIN - batch size = 32, Adam used as optimizer, discriminator learning rate = 10^{-4}, generator learning rate = 10^{-5}, discriminator trained every 5 epochs, discriminator with three linear layers, two with ReLU activations and one with sigmoid, generator with three linear layers, all with ReLU activation.
- GAIN-LSTM - same hyperparameters as GAIN except discriminator with an LSTM and a linear layer with sigmoid activation, generator with four LSTMs and a linear layer with ReLU activation.
- Luong attention model - batch size = 32, Adam used as optimizer with learning rate = 10^{-4}, encoder hidden size = 16, # of encoder/decoder layers = 1, attention type used = general

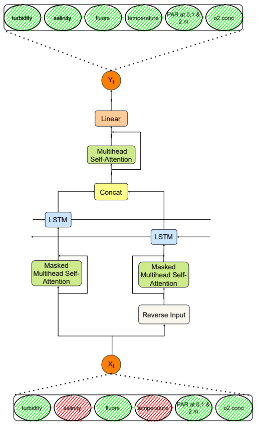

Fig. 2. Proposed architecture for imputation. The input passed through masked multi-head attention layers in forward and backward directions resulting in different attention weights for each bi-LSTM layer direction. The resulting tensors of bi-LSTM layer are concatenated and fed into a multi-head self-attention and a linear layer respectively. The output is the imputed vector.

2) Prediction Model: A model that predicts oxygen values given the current observations is also created to further support the quality of the imputation done by SAI. The prediction model consists of a 1-D convolution layer, a bidirectional LSTM layer, and a linear layer similar to [12] as depicted in Figure 3. We use the SAI model that was trained for 60% miss rate to impute the data using the WestGab data for this prediction task. The WestGab data was chosen due to the high percentage of non-imputed dissolved oxygen variable. The kernel used for the convolution layer is $2x2$ with a stride of 1. By predicting the dissolved oxygen we may be able to detect phytoplankton bloom patterns.

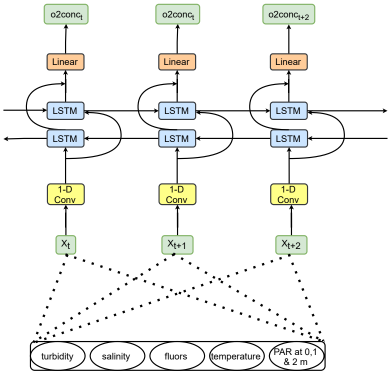

Fig. 3. Proposed architecture for prediction. The input is passed through a 1-D Convolutional layer, a bi-LSTM layer and a linear layer. The output is a single float variable.

TABLE I
SUMMARY OF THE DATASET VARIABLES

	mean	std	min	max	Description	Unit
fluors	1.16	1.76	0.01	42.320	Chlorophyll Fluorescence	arb. unit
ftu	8.52	11.54	0.01	221.22	Turbidity	Formazin Turbidity Unit (FTU)
o2conc	9.19	1.00	5.40	16.04	Dissolved oxygen concentration	mg/l
sal	33.92	1.13	25.76	35.459	Salinity	PSS78 (Practical Salinity Scale)
temp	11.56	4.33	1.74	21.330	Temperature	°C
depth_0	225.70	384.91	0.00	2566.80	PAR at 0 meter	$\mu \mathrm{E} m^{-2} s^{-1}$
depth_1	69.15	171.46	0.00	1622.70	PAR at 1 meter	$\mu \mathrm{E} m^{-2} s^{-1}$
depth_2	44.47	116.16	0.00	1617.50	PAR at 2 meters	$\mu \mathrm{E} m^{-2} s^{-1}$

IV. Results

The complete data points were randomly set to missing according to a certain percentage by masking. The data was normalized using min-max normalisation using all the available locations. All the models were trained with 70% of the WestGab data to observe the imputation performance of the model across datasets with different time ranges and spatial properties while yielding information only from a single dataset. Mean Square Error (MSE) was used as the loss function for training the models. For the kNN regressor, the neighbour count was set to 25 with uniform weights and the iteration count was set to 100 for the Bayesian Ridge Regressor. Sequence length for LSTM based models was chosen to be 32. This meant the model would be able to learn the true distribution of the target variables instead of the imputations. An Adam optimiser was used for all of the deep learning models [18]. Min-max normalisation was used to force the data to be defined in $[0, 1]$. The missing values were initially imputed with -1 as a placeholder. Using this type of normalisation enables models to use the softplus activation function at the end of linear layers, denoted by Equation 1 where $beta$ is a hyperparameter. Use of a ReLU was avoided due to PAR features having a substantial difference between min and max values and softplus function provides a low rate of reduction for lower values.

$$\text{Softplus}(x) = \frac{1}{\beta} * \log(1 + \exp(\beta * x)) \qquad (1)$$

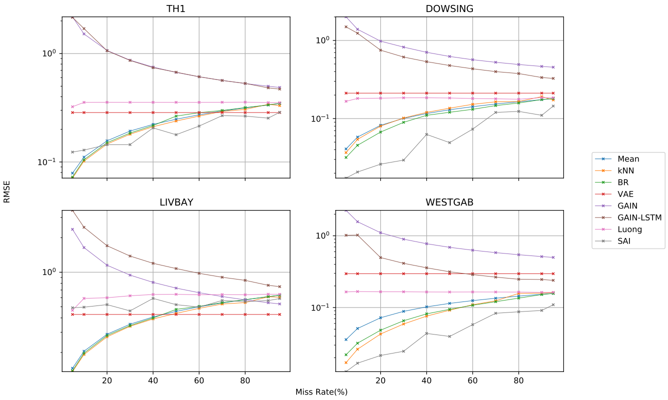

Fig. 4. Comparison of imputation performance of models for four datasets at different missing percentages. SAI outperforms other models in all miss rates for WestGab and Dowsing and the majority of miss rates for TH1.

The models' results were compared by using root mean square error (RMSE). All neural network models were trained with an early stopping criteria of patience 20 and delta of 10^{-5}. If early stopping was not applied after 300 epochs, training was terminated. During training, all 8 features were imputed.

The GAIN model was tested in two different settings one with linear layers and another one with LSTM layers which included a linear layer at the end, named GAIN and GAIN-LSTM respectively. VAE was trained with imputation and reconstruction error without using a missingness matrix simultaneously contrary to [13], [23]. It should be noted that the neural network models do reconstruction and imputation whereas MICE and mean only perform imputation. Data with MAR properties assumes that the missing values can be imputed with the observed variables so the reconstruction loss of the overall network has to be taken into account for deep learning models whereas for MICE and mean imputation no such assumption is necessary since they do not modify observed variables.

Figure 4 visualizes the results of experimentation where the proposed model outperforms the other models for all miss rates in Dowsing and WestGab and for majority of the miss rates in TH1. The exact values of RMSE can be found in Tables III to X. Table II refers to the prediction task of dissolved oxygen in four datasets after the missing data was imputed. The reconstructed values by SAI was replaced with original values before training for the prediction task.

TABLE II
RMSE OF PREDICTION FOR ALL DATASETS

	Error(RMSE)			
	TH1	DOWSING	LIVBAY	WESTGAB
Conv-LSTM	0.0840	0.0806	0.1289	0.0740

V. Discussion

The GAIN algorithm is used for imputing MCAR data [38]. RMSE of GAIN and GAIN-LSTM show that water quality data is not MCAR due to the model's performance on the supplied locations. Exposing the temporal properties

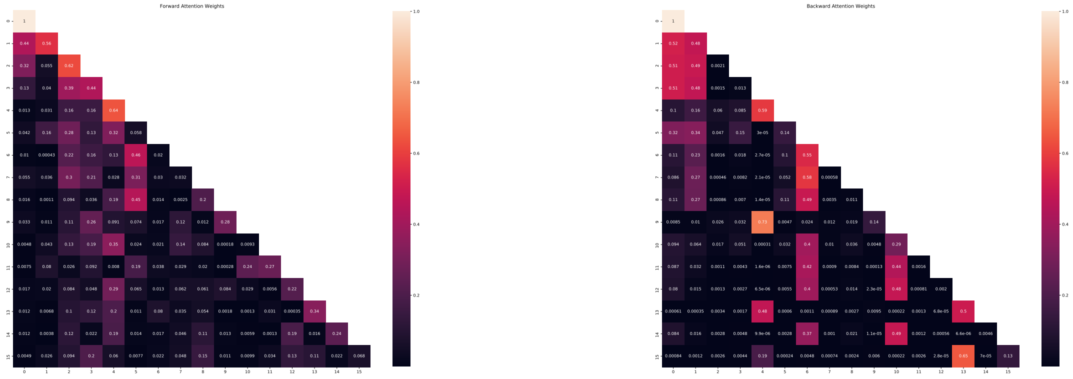

Fig. 5. Sample heatmap of self-attention scores for a sequence length of 16. The higher scores indicate more attention to that part of the input. The component produces different scores in backward and forward directions.

of the data by using GANs under the assumption of MCAR mechanism does not aid the imputation performance except WestGab and Dowsing. The data was assumed to be MAR, as seen from the other models, given more evidence of the data i.e., lower miss rates, RMSE always decreases. The poor performance of the GAIN imputer at low rates of missing values shows that the model is not fit for reconstruction purposes. The differences between the models come from the limits to understand the data with the lowest amount of evidence. At lower miss rates the models apart from VAE and GAIN perform better since non-imputed data is abundant and the model is able to model the missingness.

Different Bayesian approaches were applied with VAE and BR. Both of the models map the distributions of data to a Gaussian distribution, however VAE maps it to a lower distribution by doing encoding to a lower dimension, sampling from this distribution and decoding to the original distribution. For this task, VAE maps the distribution of the data together with the missingness whereas BR assumes that the data and its parameters are normally distributed in its original space and does no reconstruction. It should be noted that the VAE model shows signs of underfitting as the training is terminated after 19 epochs for all rates of missingness and did not improve under early stopping limits. Therefore, VAE was not considered a suitable model for imputation as it is obvious that it does not learn from the data. SAI focuses on the important sections of the input instead of modelling the latent distribution of the samples as a whole, therefore it is less prone to underfitting and does not encode the data to latent dimensions.

The scope of the dataset for experimentation has high percentages (>50%) of missing data in all of the datasets, even after data treatment. The proposed model is aimed to focus on a higher percentage of missing data. Previous work [3], [4] has shown that there are semi-daily and daily cycles, in spite of skips in the training data, the proposed SAI is able to impute the data effectively regardless of miss rates in majority of the locations.

Using a different attention mechanism benefits the performance of the model. Luong attention focuses on the relationship between input and output whereas the proposed model uses a mechanism of Vaswani et al. [36] which shifts the focus solely to the input of the component. Figure 5 visualizes the attention mechanism used before the two LSTMs. The multi-head attention component uses ReLU as an activation function which results in weights with ≥ 0 where no attention is paid to components with 0 weights. This also shows that the bidirectionality of the model helps it focus on different aspects of the data in different directions and forces the focus on key components of the data. The attention mechanism used by Luong focuses on all of the encoder hidden states and current decoder hidden states. The self-attention component focuses only on the input whereas Luong attention focuses on the relationship between the input and the output. Application of different neural networks architectures results in different RMSE values such that Luong's RMSE has less deviation depending on the dataset.

The kNN model shows that the data points show similar properties at low missing percentages as seen from Figure 4 since the model uses nearest neighbours where feature X is not missing. As the complete data points are decreasing the performance drops drastically to 0.12-0.16 between 70-95% miss rates for WestGab and to 0.52-0.61 for LivBay between the same miss rates for the kNN model. The high missing % of the problem makes the kNN model unsuitable for this task compared to SAI. For lower missing percentages (<%40), the neural network models have to shift the focus on reconstruction rather than imputation, still the model is able to do both tasks in majority of the cases presented. Since MICE and mean models do not need to do reconstruction, as information is removed from the data, RMSE increases.

The overall performance of SAI gives insights about the missingness properties of these locations. The missingness mechanism of Dowsing and WestGab are similar as the RMSE values of SAI, kNN, and BR show the same pattern. The missingness pattern of LivBay differs from the other three sites since each tested model had higher RMSE rates for that specific location.

The prediction model was trained and tested on both imputed and non-imputed data. From the results in Figure II, it can be deduced that the imputation model is able to generalise the different distributions to an extent as the highest RMSE was attained by LivBay data with an RMSE of 0.1289. It shows that the locations have different distributions relating to the dissolved oxygen concentration.

VI. Conclusion

This paper introduced a novel architecture for the task of data imputation in the context of water quality and compared various machine learning and deep learning methods. By introducing a different architecture and attention mechanism, the performance of imputation is improved where data is missing above 50%. The attention mechanism increases the interpretability of the model at different stages, aiding data understanding.

Future directions of research include the usage of different loss functions to reduce the effect of reconstruction loss on the model and broader experimentation on well-known datasets to test the generalisability of the architecture. Ensembles of neural network architectures could be applied together to minimize the effect of reconstruction loss. Transfer learning techniques could be applied to improve the prediction of dissolved oxygen and imputation. A limitation of our architecture is volatility of the model due to the initial imputation value. Experimentation with different initial imputation values is required to test the generalisability of the architecture.

The data used for this work was obtained through in situ measurements which is highly frequent. Other forms of data such as ship-based data obtain measurements less frequently, the proposed model could be tested on such data in the future.

References

[1] P. Bansal, P. Deshpande, and S. Sarawagi, "Missing value imputation on multidimensional time series," *arXiv preprint arXiv:2103.01600*, 2021.

[2] B. K. Beaulieu-Jones, J. H. Moore, and P. R. O.-A. A. C. T. CONSORTIUM, "Missing data imputation in the electronic health record using deeply learned autoencoders," in *Pacific symposium on biocomputing 2017*. World Scientific, 2017, pp. 207–218.

[3] A. N. Blauw, E. Beninca, R. W. Laane, N. Greenwood, and J. Huisman, "Dancing with the tides: fluctuations of coastal phytoplankton orchestrated by different oscillatory modes of the tidal cycle," *PLoS One*, vol. 7, no. 11, 2012.

[4] A. N. Blauw, E. Benincà, R. W. Laane, N. Greenwood, and J. Huisman, "Predictability and environmental drivers of chlorophyll fluctuations vary across different time scales and regions of the north sea," *Progress in Oceanography*, vol. 161, pp. 1–18, 2018.

[5] G. Boquet, J. L. Vicario, A. Morell, and J. Serrano, "Missing data in traffic estimation: A variational autoencoder imputation method," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2882–2886.

[6] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[7] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "Brits: bidirectional recurrent imputation for time series," in *Advances in Neural Information Processing Systems*, 2018, pp. 6775–6785.

[8] L. Folguera, J. Zupan, D. Cicerone, and J. F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices," *Chemometrics and Intelligent Laboratory Systems*, vol. 143, pp. 146–151, 2015.

[9] C. Güler, G. D. Thyne, J. E. McCray, and K. A. Turner, "Evaluation of graphical and multivariate statistical methods for classification of water chemistry data," *Hydrogeology journal*, vol. 10, no. 4, pp. 455–474, 2002.

[10] J. Heffernan, J. Barry, M. Devlin, and R. Fryer, "A simulation tool for designing nutrient monitoring programmes for eutrophication assessments," *Environmetrics: The official journal of the International Environmetrics Society*, vol. 21, no. 1, pp. 3–20, 2010.

[11] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019.

[12] X. Jin, X. Yu, X. Wang, Y. Bai, T. Su, and J. Kong, "Prediction for time series with cnn and lstm," in *Proceedings of the 11th International Conference on Modelling, Identification and Control (ICMIC2019)*. Springer, 2020, pp. 631–641.

[13] E. Jun, A. W. Mulyadi, and H.-I. Suk, "Stochastic imputation and uncertainty-aware attention to ehr for mortality prediction," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–7.

[14] M. H. Khalifeloo, M. Mohammad, and M. Heydari, "Multiple imputation for hydrological missing data by using a regression method (klang river basin)," *International Journal of Researchin Engineering and Technology*, vol. 4, no. 06, 2015.

[15] H.-G. Kim, G.-J. Jang, H.-J. Choi, M. Kim, Y.-W. Kim, and J. Choi, "Recurrent neural networks with missing information imputation for medical examination data prediction," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2017, pp. 317–323.

[16] M. Kim, S. Baek, M. Ligaray, J. Pyo, M. Park, and K. H. Cho, "Comparative studies of different imputation methods for recovering streamflow observation," *Water*, vol. 7, no. 12, pp. 6847–6860, 2015.

[17] W. Kim, W. Cho, J. Choi, J. Kim, C. Park, and J. Choo, "A comparison of the effects of data imputation methods on model performance," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2019, pp. 592–599.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "Collagan: Collaborative gan for missing image data imputation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2487–2496.

[20] Y. Liu, R. Yu, S. Zheng, E. Zhan, and Y. Yue, "Naomi: Non-autoregressive multiresolution sequence imputation," *arXiv preprint arXiv:1901.10946*, 2019.

[21] Y. Luo, X. Cai, Y. Zhang, J. Xu *et al.*, "Multivariate time series imputation with generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[22] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[23] J. T. McCoy, S. Kroon, and L. Auret, "Variational autoencoders for missing data imputation with application to a simulated milling circuit," *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 141–146, 2018.

[24] I. E. Mulia, T. Asano, and P. Tkalich, "Retrieval of missing values in water temperature series using a data-driven model," *Earth Science Informatics*, vol. 8, no. 4, pp. 787–798, 2015.

[25] C. Nieh, S. Dorevitch, L. C. Liu, and R. M. Jones, "Evaluation of imputation methods for microbial surface water quality studies," *Environmental Science: Processes & Impacts*, vol. 16, no. 5, pp. 1145–1153, 2014.

[26] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, "A survey on data imputation techniques: Water distribution system as a use case," *IEEE Access*, vol. 6, pp. 63 279–63 291, 2018.

[27] S. Papadimitriou, J. Sun, C. Faloutos, and S. Y. Philip, "Dimensionality reduction and filtering on time series sensor streams," *Managing and Mining Sensor Data*, pp. 103–141, 2013.

[28] R. Ratolojanahary, R. H. Ngouna, K. Medjaher, J. Junca-Bourié, F. Dauriac, and M. Sebilo, "Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset," *Expert Systems with Applications*, vol. 131, pp. 299–307, 2019.

[29] R. Rodríguez, M. Pastorini, L. Etcheverry, C. Chreties, M. Fossati, A. Castro, and A. Gorgoglione, "Water-quality data imputation with a high percentage of missing values: A machine learning approach," *Sustainability*, vol. 13, no. 11, p. 6318, 2021.

[30] X. Shu, F. Porikli, and N. Ahuja, "Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3874–3881.

[31] D. J. Stekhoven, "missforest: Nonparametric missing value imputation using random forest," *Astrophysics Source Code Library*, 2015.

[32] I. Sucholutsky, A. Narayan, M. Schonlau, and S. Fischmeister, "Pay attention and you won't lose it: a deep learning approach to sequence imputation," *PeerJ Computer Science*, vol. 5, p. e210, 2019.

[33] H. Tabari and P. Hosseinzadeh Talaee, "Reconstruction of river water quality missing data using artificial neural networks," *Water Quality Research Journal of Canada*, vol. 50, no. 4, pp. 326–335, 2015.

[34] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1405–1414.

[35] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[37] X. Yi, Y. Zheng, J. Zhang, and T. Li, "St-mvl: filling missing values in geo-sensory time series data," 2016.

[38] J. Yoon, J. Jordon, and M. Van Der Schaar, "Gain: Missing data imputation using generative adversarial nets," *arXiv preprint arXiv:1806.02920*, 2018.

[39] A. Zhang, S. Song, Y. Sun, and J. Wang, "Learning individual models for imputation," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 160–171.

[40] Y.-F. Zhang, P. Thorburn, W. Xiang, and P. Fitch, "Ssim-a deep learning approach for recovering missing time series sensor data," *IEEE Internet of Things Journal*, 2019.

[41] Y. Zhang and P. J. Thorburn, "A dual-head attention model for time series data imputation," *Computers and Electronics in Agriculture*, vol. 189, p. 106377, 2021.

Appendix

TABLE III
RMSE for Mean Imputation

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.3456	0.1803	0.6298	0.1575
90	0.3355	0.1747	0.6119	0.1533
80	0.3152	0.1643	0.5760	0.1444
70	0.2952	0.1536	0.5402	0.1349
60	0.2725	0.1420	0.4998	0.1250
50	0.2478	0.1297	0.4559	0.1143
40	0.2223	0.1160	0.4076	0.1023
30	0.1929	0.1006	0.3526	0.0885
20	0.1567	0.0820	0.2881	0.0721
10	0.1109	0.0580	0.2051	0.0511
5	0.0789	0.0409	0.1457	0.0358

TABLE IV
RMSE for KNN Imputation

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.3310	0.1729	0.6061	0.1619
90	0.3402	0.1912	0.6107	0.1589
80	0.3055	0.1653	0.5439	0.1575
70	0.2926	0.1639	0.5284	0.1241
60	0.2651	0.1523	0.4846	0.1095
50	0.2386	0.1359	0.4378	0.0919
40	0.2114	0.1200	0.3918	0.0760
30	0.1804	0.1017	0.3380	0.0590
20	0.1463	0.0801	0.2735	0.0428
10	0.1019	0.0542	0.1919	0.0263
5	0.0710	0.0367	0.1361	0.0171

TABLE V
RMSE for BR Imputation

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.3475	0.1808	0.6306	0.1573
90	0.3360	0.1751	0.6133	0.1518
80	0.3174	0.1583	0.5768	0.1354
70	0.2988	0.1466	0.5439	0.1207
60	0.2849	0.1314	0.5044	0.1077
50	0.2649	0.1204	0.4746	0.0945
40	0.2169	0.1101	0.3997	0.0817
30	0.1849	0.0895	0.3429	0.0655
20	0.1505	0.0672	0.2808	0.0484
10	0.1048	0.0455	0.1972	0.0319
5	0.0726	0.0318	0.1386	0.0220

TABLE VI
RMSE for VAE Imputation

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.2857	0.2117	0.4287	0.2964
90	0.2857	0.2117	0.4287	0.2964
80	0.2857	0.2117	0.4287	0.2964
70	0.2857	0.2117	0.4287	0.2964
60	0.2857	0.2117	0.4287	0.2964
50	0.2857	0.2117	0.4287	0.2964
40	0.2857	0.2117	0.4287	0.2964
30	0.2857	0.2117	0.4287	0.2964
20	0.2857	0.2117	0.4287	0.2964
10	0.2857	0.2117	0.4287	0.2964
5	0.2857	0.2117	0.4287	0.2964

TABLE VII
RMSE for GAIN Imputation

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.6441	0.7275	0.6483	0.6680
90	0.5464	0.5417	0.5746	0.5561
80	0.5462	0.5226	0.5865	0.5602
70	0.5684	0.5297	0.6153	0.5845
60	0.6106	0.5662	0.6630	0.6296
50	0.6695	0.6220	0.7276	0.6901
40	0.7488	0.6972	0.8157	0.7712
30	0.8705	0.8088	0.9434	0.8917
20	1.0735	1.0073	1.1655	1.0961
10	1.5355	1.4342	1.6666	1.5594
5	2.1704	1.9781	2.3393	2.2077

TABLE VIII
RMSE for GAIN-LSTM Imputation

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.5029	0.3409	0.7694	0.2506
90	0.5167	0.3504	0.7921	0.2538
80	0.5950	0.4335	0.8630	0.3582
70	0.5625	0.3965	0.9035	0.2645
60	0.6108	0.4257	0.9753	0.2871
50	0.6648	0.4852	1.0746	0.3171
40	0.7476	0.5306	1.1984	0.3536
30	0.8637	0.6076	1.3842	0.4061
20	0.8147	0.4764	1.4490	0.2793
10	1.1593	0.6569	2.0263	0.3798
5	1.6811	0.9848	2.8946	0.5307

TABLE IX
RMSE FOR LUONG IMPUTATION

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.3524	0.1847	0.6422	0.1628
90	0.3530	0.1855	0.6433	0.1628
80	0.3564	0.1770	0.6361	0.1644
70	0.3553	0.1791	0.6393	0.1647
60	0.3548	0.1807	0.6363	0.1649
50	0.3548	0.1834	0.6419	0.1646
40	0.3549	0.1846	0.6400	0.1649
30	0.3549	0.1844	0.6235	0.1659
20	0.3553	0.1825	0.5994	0.1659
10	0.3546	0.1812	0.5915	0.1672
5	0.3233	0.1662	0.4659	0.1652

TABLE X
RMSE FOR SAI IMPUTATION

Miss Rate(%)	*TH1*	*Dowsing*	*Liverpool*	*WestGab*
95	0.2873	0.1447	0.5898	0.1095
90	0.2538	0.1101	0.5664	0.0911
80	0.2650	0.1236	0.5550	0.0873
70	0.2676	0.1199	0.5675	0.0833
60	0.2144	0.0733	0.4971	0.0578
50	0.1784	0.0492	0.5217	0.0396
40	0.2062	0.0627	0.5917	0.0436
30	0.1446	0.0294	0.4605	0.0246
20	0.1444	0.0261	0.5221	0.0215
10	0.1287	0.0207	0.4960	0.0168
5	0.1234	0.0173	0.4903	0.0128