Understanding the Efficacy of U-Net & Vision Transformer for Groundwater Numerical Modelling

Maria Luisa Taccari [1] Oded Ovadia [2] He Wang [3] Adar Kahana [4] Xiaohui Chen [1] Peter K. Jimack [3]

Abstract

This paper presents a comprehensive comparison of various machine learning models, namely U-Net (Ronneberger et al., 2015), U-Net integrated with Vision Transformers (ViT) (Dosovitskiy et al., 2021), and Fourier Neural Operator (FNO) (Li et al., 2020), for time-dependent forward modelling in groundwater systems. Through testing on synthetic datasets, it is demonstrated that U-Net and U-Net + ViT models outperform FNO in accuracy and efficiency, especially in sparse data scenarios. These findings underscore the potential of U-Net-based models for groundwater modelling in real-world applications where data scarcity is prevalent.

1. Introduction

Groundwater numerical models, such as MODFLOW (Hughes et al., 2022), are crucial for water resource management, although they are computationally demanding. To alleviate this, surrogate modelling through data-driven methods offers efficient approximations of these complex numerical techniques.

Neural Operators (Lu et al., 2021; Seidman et al., 2022), particularly the Fourier Neural Operator (FNO) (Li et al., 2020), have been at the forefront of recent advances, having shown potential to approximate arbitrary continuous functions. However, the computational demand of FNO is particularly high during training phase while these neural operators require architectural enhancements to deliver promising results in subsurface problems (Wen et al., 2022; Jiang et al., 2023). This is evident in the work of Wen et al. (Wen et al., 2022), where the integration of FNO with U-Net architecture showed improved accuracy, speed, and data efficiency in multiphase flow problems. However, Gupta and Brandstetter's work (Gupta & Brandstetter, 2022), showing that U-Net outperforms FNOs across various fluid mechanics problems, raises a question about the necessity of neural operators when the vanilla U-Net architecture already exhibits remarkable performance.

Recently, transformers (Vaswani et al., 2017) have seen considerable success in various fields, including physical systems (Cao, 2021; Li et al., 2023), for which the datasets are typically smaller compared to other domains. Only one study explores the use of transformers in groundwater modeling (Mellouli et al., 2022), demonstrating that the models were outperformed by both GRU and LSTM models to predict groundwater levels across various stations in France with meteorological and hydrological data.

Finally, the integration of U-Net with Transformers, as exemplified in studies like TransUNet (Chen et al., 2021) and ViTO (Ovadia et al., 2023), has demonstrated their utility across a broad range of applications, particularly in the field of medical image segmentation and operator learning for inverse PDE problems. Yet, the applicability of these combinations in addressing time-dependent forward problems, real-world data scenarios, and in situations with sparse data, remain areas yet to be fully explored.

Several studies, such as the one by Brakenhoff et al. (Brakenhoff et al., 2022; Mellouli et al., 2022), primarily focus on individual time series when analysing the impact of various hydrological stressors, including pumping rates, precipitation excess, and river stage variations, on groundwater levels of individual monitoring wells. While this approach provides valuable insights, it does not account for spatial correlations, thereby limiting its use to existing time series or monitoring wells. Similarly, previous comparisons have been predominantly limited to specific models like LSTM, CNNs and NARX in the context of groundwater level forecasting (Wunsch et al., 2021), leaving room for broader explorations.

In this paper, we present a comprehensive comparison

[*]Equal contribution [1]School of Civil Engineering, University of Leeds, Leeds, UK [2]Department of Applied Mathematics, Tel-Aviv University, Tel-Aviv, Israel [3]School of Computing, University of Leeds, Leeds, UK [4]Division of Applied Mathematics, Brown University, Providence, U.S.A. Correspondence to: Maria Luisa Taccari <cnmlt@leeds.ac.uk>.

Accepted after peer-review at the 1st workshop on Synergy of Scientific and Machine Learning Modeling, SynS & ML ICML, Honolulu, Hawaii, USA. July, 2023. Copyright 2023 by the author(s).

among models—specifically U-Net, U-Net integrated with Vision Transformers (U-Net+ViT), and Fourier Neural Operator (FNO)—for their efficacy in modeling time-dependent forward and inverse problems in groundwater systems. We test our model extensively on synthetic datasets, simulating conditions from the Overbetuwe region in the Netherlands, including sparse data scenarios. We show that both U-Net and U-Net+ViT are particularly well-suited to these important sparse data scenarios, with the addition of the Transformer providing enhanced predictive capability in many cases.

2. Methodology

2.1. Example of study site and data

This subsection provides context and rationale for our study via an example case study based upon the polder region of Overbetuwe in the Netherlands (Figure 1). This region showcases the characteristic Dutch system of water management where the area is divided into several polders in a mix of agriculture, nature, and urban environments. Alongside its sparse data and heterogeneous soil, these unique characteristics underscore the inherent complexities of water management in similar settings, making this dataset a suitable choice for our research. The subsoil is primarily composed of clay and sandy clay, with soil properties being determined via borehole and cone penetration tests. The study area features numerous observation wells for monitoring groundwater heads while well fields (indicated as groundwater usage facilities in the figure) are utilized for the extraction of drinking water. The work of Brakenhoff et al. (Brakenhoff et al., 2022) considers a dataset consisting of 250 head time series, with daily recordings starting from the year 1990 and drawdown attributed to the extraction from up to four well fields.

For the purposes of this study, we employ synthetic data to validate the proposed methodology, with the intention to subsequently apply the validated method to the real-world data of the Overbetuwe region. Figure 2 represents a sample of the high-fidelity labeled dataset, which is constructed using the U.S. Geological Survey (USGS) finite-difference flow model, MODFLOW. The model is composed of a single-layer representation of a confined aquifer with a 128×128 grid. The aquifer's heterogeneity is reflected through varying horizontal hydraulic conductivity within the bounds $k \in [0.1, 0.5]$ m/d. The hydraulic conductivity fields in our study are created using random fields which are then thresholded to delineate different classes. A maximum of ten pumping wells are extracting water with variable rates in the range $Q \in [0, 30]$ m^3/d over a simulation period of $T = 10$ days. The pumping wells are located in random locations which vary for each sample. The boundary conditions are delineated as Dirichlet, with the head equal to zero, mimicking a polder encircled by ditches where a stable water level is maintained through a comprehensive network of pumping stations.

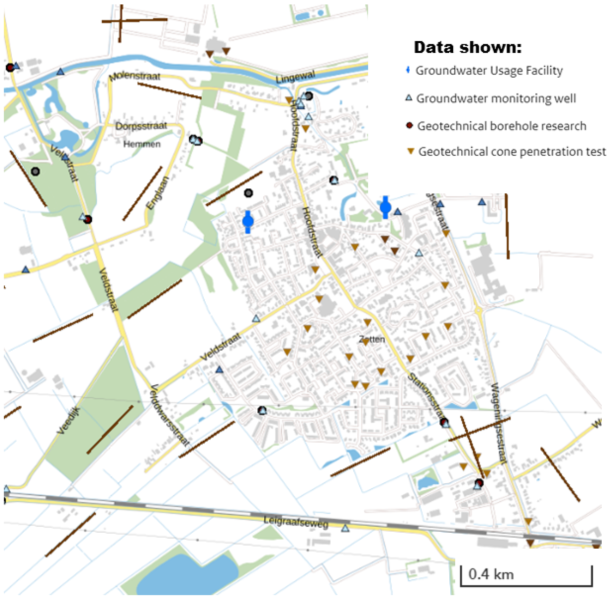

Figure 1. A representation of the target study area located in the polder region of Overbetuwe, Netherlands. (DIN).

The datasets consist of $N_{train} = 5000$ training instances and $N_{test} = 1000$ testing instances. To mirror the inherent sparsity of real-world data, a data selection strategy is adopted for the test dataset. The locations of the boreholes for estimating the hydraulic conductivity are chosen following a radial distribution pattern, and a helical pattern is used for the wells monitoring hydraulic head (Figure 2).

2.2. Architectures

The architectures of the three models under comparison in this study encompass the U-Net structure, a U-Net with attention mechanism in the bottleneck, and the Fourier neural operator (FNO).

The U-Net architecture is designed with an encoder-decoder structure where the decoder receives the upsampled feature map, which is then concatenated with the corresponding feature map from the encoder through a skip connection. Detailed diagrams of the the U-Net encoder and decoder can be found in Figures 6 and 7 in Appendix A. The encoder consists of three bottleneck blocks, where each block utilizes three layers of Conv2d, Instance Normalization, and GELU activation to extract spatial features. These blocks increase the number of channels by a factor of 2 and perform downsampling with a stride of 2. The decoder is composed

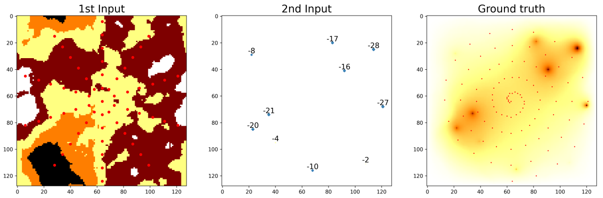

Figure 2. An test sample of the high-fidelity dataset constructed using MODFLOW. The first column showcases the heterogeneous hydraulic conductivity, the second column presents the randomly positioned pumping wells, and the third column depicts the resulting hydraulic head. The red dots (column 1 and 3) represent a selection of data points—following radial and helical patterns, respectively—emulating the sparse observations in real-world scenarios.

of a series of upsampling blocks, where each block consists of a bilinear upsampling operation (Upsample), followed by a double convolution operation. Each convolution within the decoderis followed by Instance Normalization and GELU activation function. The bottleneck consists on a single convolutional layer. In the time-dependent scenario, the time series data of the historical pumping rates is processed through two layers of feed-forward neural network (FNN) prior to being concatenated to the input for the latent space representation (Figure 6).

The second model, here called UNet+ViT, employs the Vision Transformer (ViT) (Dosovitskiy et al., 2021), in the latent space representation of the U-Net, as per implementation of TransUNet (Chen et al., 2021) and ViTO (Ovadia et al., 2023). The input is tokenized into a sequence of flattened 2D patches, each of size 1×1. Positional information is retained by employing trainable convolutional projection to learn and add specific position embeddings to the patch embeddings. The structure of the Transformer includes L blocks, with each block comprising Multi-Head Attention (MSA) and FNN. This configuration involves the use of 2 blocks, each with 2 Multihead Self-Attentions, and a FNN composed of 128 neurons. For a more detailed visualization of the Vision Transformer, attention block, and multihead attention, please refer to Appendix A, Figure 8.

The Fourier neural operator (FNO) (Li et al., 2020) model leverages the fast Fourier Transform to parameterize the integral kernel directly in the Fourier space. The implementation of FNO for the $2D$ Darcy Flow problem as presented in (Li et al., 2020) is followed in this study. The total amount of parameters of FNO corresponds to 2.38 million, that is 15 times more than UNet+ViT (151k) and 17 times more than UNet (137k).

3. Results

3.1. Forward problem with sparse observations

This section presents the prediction of the hydraulic head at sparse monitoring wells after a constant 10-day pumping period under two different training conditions. We employ distinct sampling strategies for both input and output data in our methodology. Our training data is sampled from a regular quadratic grid, while for testing we have explored other arrangements, such as radial and helical, to understand their potential impact on the prediction performance. In the first scenario, training is conducted using sparse data, with a spacing of 20 grid points for the input hydraulic conductivity field and a spacing of 8 for the output hydraulic head. Testing is then carried out on sparse data points, following the radial and helical patterns delineated in subsection 2.1. The resulting root mean square error (RMSE) is found to be 5.2×10^{-2}, 3.5×10^{-2} and 8.1×10^{-2} for the vanilla U-Net, the UNet+ViT models and FNO respectively. These results underline the superior performance of the UNet+ViT model in handling sparse data, exhibiting a lower RMSE compared to both the vanilla U-Net and the FNO models.

In contrast, when training is performed using the entire field and testing on the same sparse dataset, the error marginally escalates to 3.9×10^{-1} for FNO, 3.8×10^{-1} for UNet and 3.6×10^{-1} for UNet+ViT model. This outcome is anticipated considering the training set exhibits sparsity in the first scenario, but not in the latter. Additionally, Figure 3 displays the prediction over the entire domain, resulting in a lower RMSE of 1.0×10^{-2} for FNO, 1.7×10^{-2} and 1.9×10^{-2} for the vanilla U-Net and UNet+ViT models, respectively. The FNO model, while superior when dealing with full data, exhibits the highest predictive error under sparse data observations. These results highlight the practical advantages of the U-Net and especially UNet+ViT model in real-world scenarios for which data sparsity is common.

It should be noted that traditional simpler neural networks and other machine learning techniques may not provide adequate solutions for this specific problem. This assertion is backed by a comparison of the results from a fully connected neural network, a linear regression model and a random forest, detailed in Appendix B. Despite the substantial number of trainable parameters, reaching 51.17 million, inherent to the fully connected neural network and the application of linear regression and random forest, these methods significantly underperform compared to the U-Net, the UNet+ViT models, and FNO.

3.2. Identification of pumping wells

In this section, we focus on an inverse problem: specifically the identification of pumping wells. This task requires determining the locations and rates of pumping wells based

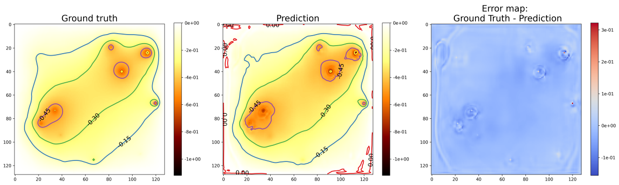

Figure 3. Predictions over the entire domain are shown, with the UNet+ViT model exhibiting a RMSE of 1.9×10^{-2}. The figure demonstrates the model's ability to accurately capture the spatial distribution of the hydraulic head across the entire domain.

on the observed hydraulic heads. Throughout these experiments, we employ a single hydraulic conductivity field, which, while spatially varying, remains identical across all samples within the dataset.

In evaluating the performance of our models, we use both RMSE and accuracy. The RMSE calculates the average difference between the true and the predicted value for each pump location in the test dataset, giving a quantitative measure of the prediction error. Complementing this, the accuracy was determined by counting the proportion of correct pump predictions, where a prediction is considered correct if the predicted and actual pump locations align. This gives a sense of how often the model correctly identifies the location of pumps.

The U-Net model performs optimally, achieving an RMSE of 5.6×10^{-2}. Interestingly, the integration of the Vision Transformer with the U-Net model does not confer any additional precision in this scenario, yielding a near RMSE of 6.1×10^{-2}. The FNO model exhibits a higher RMSE of 1.1×10^{-1}, indicating a somewhat lower accuracy in identifying the pumping well locations.

To visually illustrate these results, Figure 4 presents a test sample using the U-Net + ViT model. It demonstrates an accuracy of 93% in locating the pumps, calculated across the entire test dataset. The figure visualizes the model's ability to accurately identify the positions and the pumping rate of the wells. In comparison, the FNO model achieved a notably lower detection accuracy of 79% in the same task.

3.3. Example results for time series data

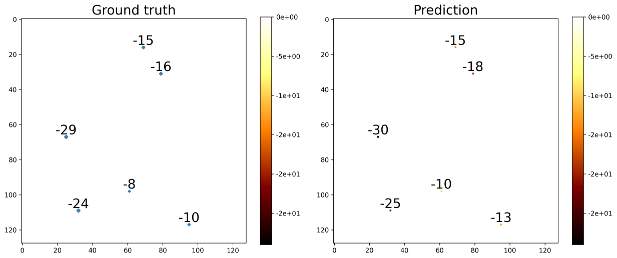

Figure 4. Test sample results for the identification of pumping wells using the U-Net + ViT model. The model accurately identifies both the positions of the pumping wells with an accuracy of 93 % across the entire test dataset.

This section unveils the results achieved from the analysis of time series data, starting with a simplified scenario, for which the inputs are the varying hydraulic conductivity field and the pumping rate of a single pump which varies over a 10-day simulation period. Results are evaluated in terms of root mean square error (RMSE) with a focus on the comparison of different configurations of the U-Net architecture with transformers. Figure 5 presents a comparison of results over 5 time frames for the U-Net with the Vision Transformer under autoregressive testing conditions.

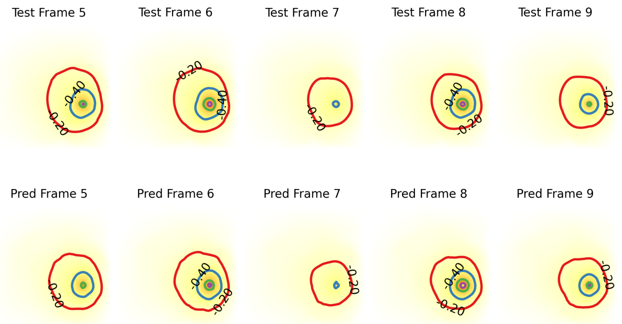

Figure 5. Depiction of autoregressive testing results from the UNet+ViT model over the final five frames of a ten-frame test sequence. This sequence simulates a single variable-rate pump operating over a ten-day period within a region characterized by a diverse hydraulic conductivity field. The top row presents the ground truth, while the lower row displays the predicted outcomes. Contour lines in the images represent groundwater levels, which are color-coded for enhanced visual clarity. Each successive frame is employed as input to the U-Net-Vision Transformer system, aiding in the prediction of the subsequent frame.

The RMSE for each method was calculated to quantify the models' performance. The U-Net architecture alone yielded an RMSE of 1.79×10^{-2}. When supplemented with a Vision Transformer, consisting of 2 attention blocks and 2 heads, the performance improves, registering an RMSE of 1.67×10^{-2}. However, increasing the complexity of the Vision Transformer to 8 blocks and 8 heads did not further improve the performance, instead, it led to a slight degradation in the RMSE (1.77×10^{-2}). Adding an Axial Transformer (Ho et al., 2019) to the U-Net architecture also did not enhance the performance, yielding an RMSE of 1.83×10^{-2}. These results suggest that while adding a Vision Transformer to the U-Net architecture leads to performance improvement, increasing the complexity of the latent space does not necessarily do so.

4. Conclusion

This paper explores and evaluates the capabilities of different machine learning models, with a particular focus on U-Net, U-Net integrated with Vision Transformers (ViT), and Fourier Neural Operator (FNO), in the context of predicting hydraulic head in groundwater studies.

Our analysis and testing, conducted on synthetic datasets designed to simulate the conditions from the Overbetuwe region in the Netherlands and including scenarios with sparse data, firmly establish that both U-Net and U-Net + ViT models are particularly adept at dealing with such tasks. Importantly, these models are also preferred due to their fewer requisite parameters.

Specifically, in the case of sparse observation scenarios, the vanilla U-Net and the U-Net + ViT models outperformed the FNO model. In particular the performance of the UNet+ViT model was superior when handling sparse data, highlighting the potential of the model in real-world applications, where data scarcity is a common issue. The U-Net model demonstrated optimal performance in identifying pumping wells. Interestingly, the integration of the Vision Transformer with the U-Net model did not confer any additional accuracy in this scenario. As for the analysis of time series data, supplementing the U-Net architecture with a Vision Transformer improved the model performance, recording an RMSE of 1.67×10^{-2} compare to 1.79×10^{-2} of the vanilla U-Net. However, increasing the complexity of the Vision Transformer did not further enhance the model performance, indicating that a more complex architecture does not necessarily yield better results.

Future research will involve applying this validated methodology to real-world data, beginning with the Overbetuwe region in the Netherlands. This will offer an opportunity to further validate and refine the model, accounting for the sparsity and uncertainties inherent in real-world data.

Broader impact

The implications of this research span a wide range of potential societal impacts, with a primary focus on improving the efficiency and reliability of groundwater level forecasting. Given that groundwater is a crucial resource for approximately 2.5 billion people worldwide, fulfilling their daily water needs, and a significant source of global irrigation water, the importance of reliable forecasts cannot be overstated. Our work, through enhancing the performance of groundwater numerical models, offers an opportunity to revolutionize the management and distribution of this vital resource. By providing more accurate and data-efficient predictions, we can aid in the formulation of informed and sustainable water management strategies. This is particularly crucial considering the pressing challenges of population growth and climate change.

Acknowledgements

This work was carried out with support of the Leeds-York-Hull Natural Environment Research Council (NERC) Doctoral Training Partnership (DTP) Panorama under grant NE/S007458/1. Our sincere appreciation is extended to Professor Karniadakis of Brown University. The financial assistance provided by the Leeds Institute of Fluid Dynamics and Deltares, which made possible the research visit to Brown University, is also gratefully acknowledged. Lastly, we would like to express our gratitude to the reviewers. Their critiques and suggestions have greatly enhanced the overall clarity of our work.

References

DINO loket. URL https://www.dinoloket.nl/en/subsurface-data.

Brakenhoff, D. A., Vonk, M. A., Collenteur, R. A., Van Baar, M., and Bakker, M. Application of time series analysis to estimate drawdown from multiple well fields. *Frontiers in Earth Science*, 10, 2022. ISSN 2296-6463. doi: 10.3389/feart.2022.907609. URL https://www.frontiersin.org/articles/10.3389/feart.2022.907609.

Cao, S. Choose a transformer: Fourier or galerkin, 2021.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Gupta, J. K. and Brandstetter, J. Towards multi-spatiotemporal-scale generalized pde modeling, 2022.

Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. Axial attention in multidimensional transformers, 2019.

Hughes, J. D., Russcher, M. J., Langevin, C. D., Morway, E. D., and McDonald, R. R. The modflow application programming interface for simulation control and software interoperability. *Environmental Modelling & Software*, 148:105257, 2022.

Jiang, Z., Zhu, M., Li, D., Li, Q., Yuan, Y. O., and Lu, L. Fourier-mionet: Fourier-enhanced multiple-input neural operators for multiphase modeling of geological carbon sequestration, 2023.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Li, Z., Meidani, K., and Farimani, A. B. Transformer for partial differential equations' operator learning, 2023.

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, mar 2021. doi: 10.1038/s42256-021-00302-5. URL https://doi.org/10.1038%2Fs42256-021-00302-5.

Mellouli, N., Rabah, M. L., and Farah, I. R. Transformers-based time series forecasting for piezometric level prediction. In *2022 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pp. 1–6, 2022. doi: 10.1109/EAIS51927.2022.9787530.

Ovadia, O., Kahana, A., Stinis, P., Turkel, E., and Karniadakis, G. E. Vito: Vision transformer-operator, 2023.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015.

Seidman, J. H., Kissas, G., Perdikaris, P., and Pappas, G. J. Nomad: Nonlinear manifold decoders for operator learning, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.

Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A., and Benson, S. M. U-fno – an enhanced fourier neural operator-based deep-learning model for multiphase flow, 2022.

Wunsch, A., Liesch, T., and Broda, S. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (lstm), convolutional neural networks (cnns), and non-linear autoregressive networks with exogenous input (narx). *Hydrology and Earth System Sciences*, 25(3):1671–1687, 2021. doi: 10.5194/hess-25-1671-2021. URL https://hess.copernicus.org/articles/25/1671/2021/.

A. Appendix A

This appendix provides detailed diagrams of the model structures.

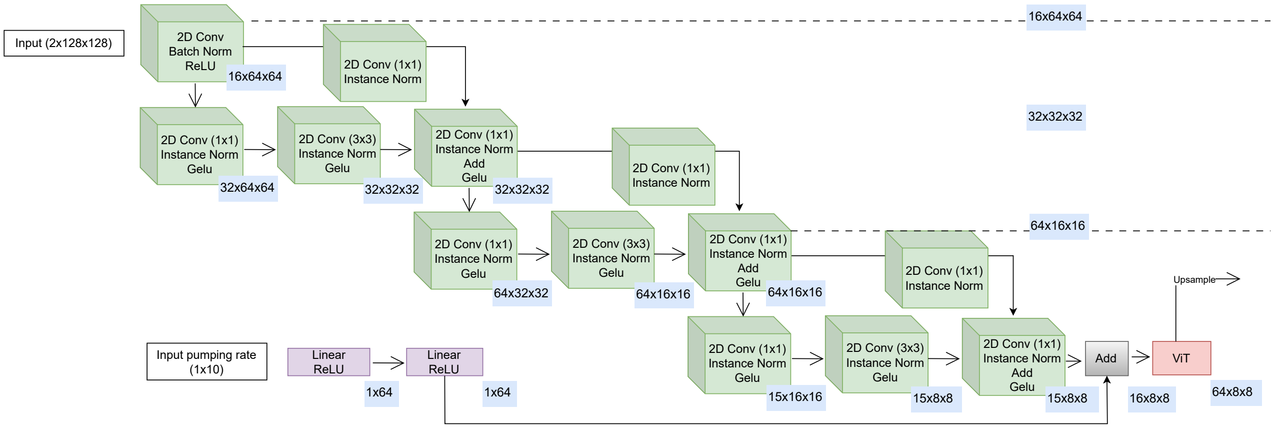

Figure 6. U-Net Encoder: The encoder consists of 3 bottleneck blocks, each comprising 3 convolutional blocks (Conv2d, Instance Normalization, and GELU). These blocks increase the channel dimension by a factor of 2 and perform downsampling with a stride of 2, extracting spatial features. In time-dependent scenarios, the pumping rate time series undergoes processing through a two-layer feed-forward neural network. The resulting processed data is then concatenated with the input, creating the latent space representation.

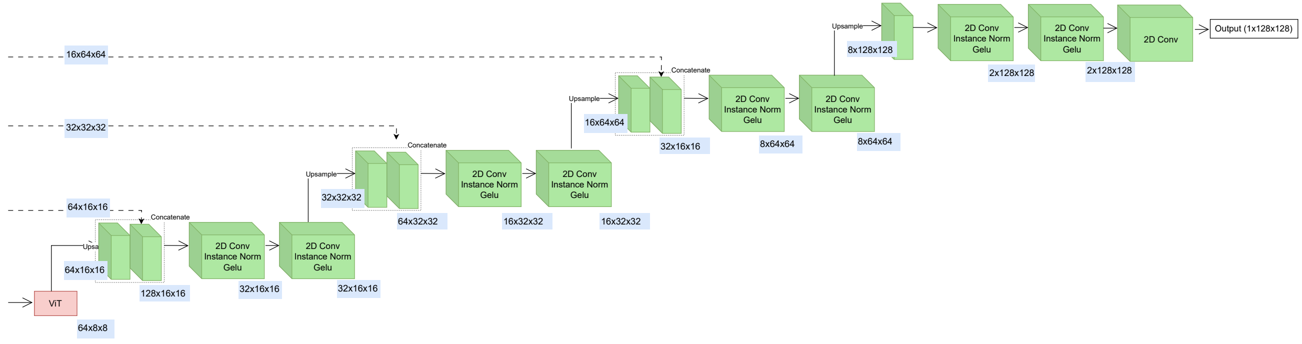

Figure 7. U-Net Decoder: The decoder is composed of a series of upsampling blocks, where each block consists of a bilinear upsampling operation followed by a double convolution operation.

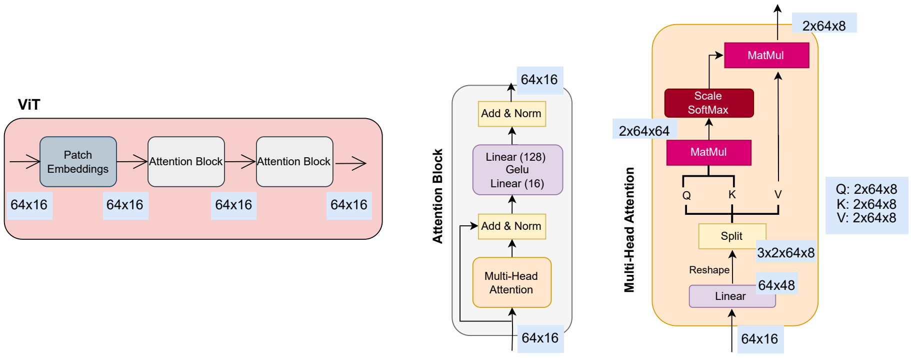

Figure 8. The Vision Transformer (ViT) and its components: The input is tokenized into a sequence of flattened 2D patches, each of size 1×1. The structure of the Transformer encoder includes 2 attention blocks, each containing Multi-Head Attention block and fully connected layers.

B. Appendix B

This appendix sets out to examine whether simpler machine learning models, specifically a fully connected neural network, a linear regression model, and a Random Forest model, can achieve the same level of accuracy as more advanced models like the U-Net, the UNet+ViT models, and FNO in predicting groundwater levels.

The particular Random Forest model tested here used 30 estimators. The fully connected neural network, employed for this comparison, comprises three hidden layers, each containing 1000 nodes and using ReLU activation functions. The model holds an impressive count of 51.17 million trainable parameters.

Unfortunately, none of the models was able to predict accurately the groundwater levels neither capturing the location of the wells. Specifically, the fully connected neural network and the linear regression model yielded high RMSEs of 1.17×10^{-1} and 1.24×10^{-1}, respectively. The Random Forest model fared slightly better, achieving a lower RMSE of 1.02×10^{-1}, but it still fell short of the U-Net, the UNet+ViT models, and FNO.

Figure 9 visually contrasts the predictions of these simpler models gainst the ground truth. Their significant underperformance becomes evident when compared to more sophisticated models. For a comparison of these results with accurate outcomes produced by the UNet+ViT model, the reader is directed to Figure 3.

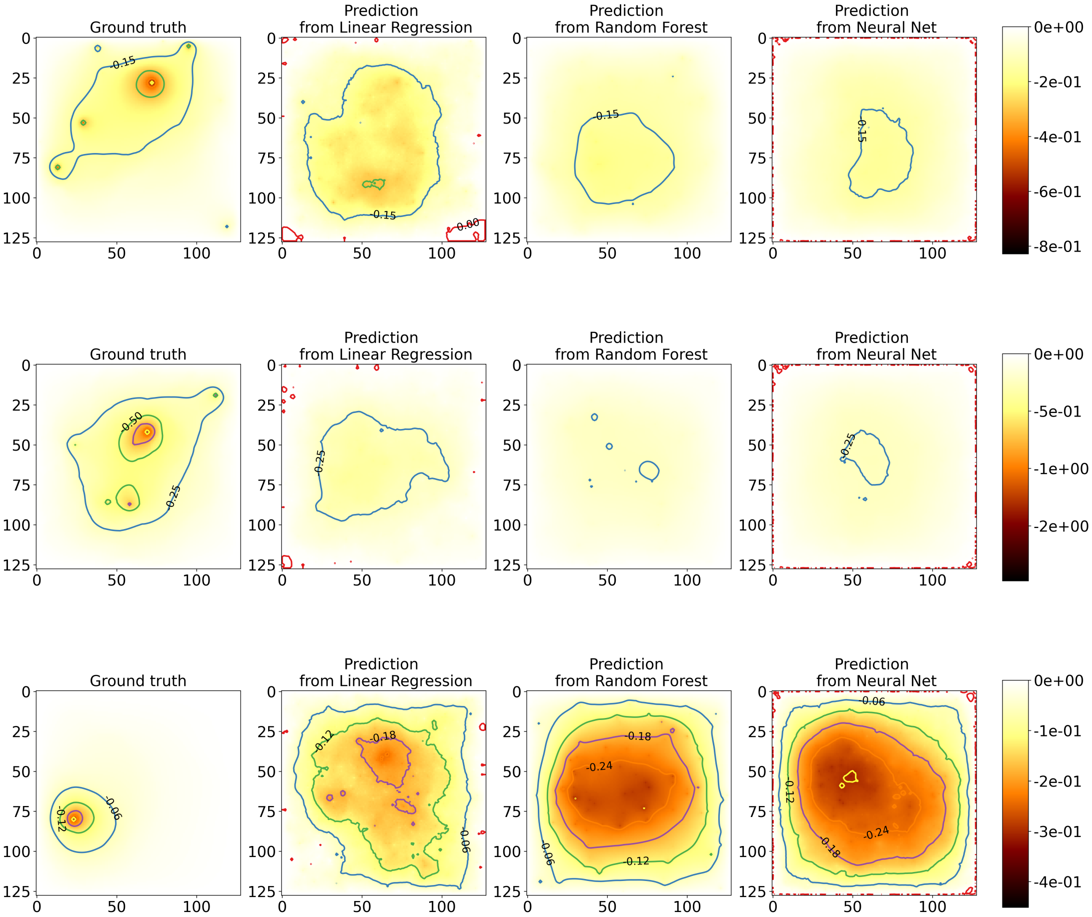

Figure 9. These figures illustrate the testing results using a fully connected neural network, a linear regression model, and a Random Forest model. The top row of each figure presents the ground truth, while the remaining rows display the predicted outcomes generated by each respective model. Contour lines represent groundwater levels, color-coded for clarity.