Predicting and Visualizing Tidal Trends in Vietnam using an Automated Geographical Data Pipeline

Ankit Gupta
University of Virginia
Charlottesville, Virginia, USA
ag8mp@virginia.edu

N. Rich Nguyen
University of Virginia
Charlottesville, Virginia, USA
nn4pj@virginia.edu

ABSTRACT

Vietnam is one of the most natural disaster-prone countries in the world, with the country being exposed to floods, cyclones, and severe storms frequently. When developing a smart city in Vietnam, it's key to be able to take advantage of this historical data to model and forecast future events. One of the key indicators for many natural disasters in Vietnam is the change in high tides over seasons, but many current models simply analyze current data rather than making intelligent predictions for future data. In addition, the algorithms that do exist are often difficult to visualize, and stay local to one's desktop, thus making them harder to understand. In this paper, we introduce an LSTM-based model to predict Tidal severity in 14-day intervals for the 20 largest cities in Vietnam. In addition, we automatically retrieve tidal information across the past 5 years for each of these cities and automatically encode the tidal predictions for 5 14-day intervals from 04/01 to 05/28 in 5 separate visualizable GeoJSON files. Each of these files can then be uploaded and visualized as styles using modern visualization tools like MapBox. By doing so, we are not only able to forecast critical tidal information for Vietnam, but we are able to create an automated pipeline consisting of three scripts for collecting, predicting, and visualizing data. This pipeline allows users to define a set of cities and their coordinates and fully automate the process of finding relevant data, making predictions on that data, and displaying those intelligent predictions through geographic visualizations.

CCS CONCEPTS

• **Tide Prediction**; • **Machine Learning** → *Neural Networks*; • **LSTM Neural Networks**; • **Smart Cities** → Natural Disaster Prediction;

KEYWORDS

machine learning, smart city, tide prediction

ACM Reference Format:
Ankit Gupta and N. Rich Nguyen. 2022. Predicting and Visualizing Tidal Trends in Vietnam using an Automated Geographical Data Pipeline. In *Proceedings of (KDD Undergraduate Consortium '22).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

KDD Undergraduate Consortium '22, August 14–18, 2022, Washington, DC

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Vietnam is one of the five most prone countries in the world to climate change [1]. A large part of this is the fact that the country has an extensive coastline at 3,260 kilometers in length, along with the fact that Southeast Asia has a rising sea level at four times faster than the global average. Vietnam's geography is also one of the most vulnerable, since it has a low-lying coastline, which means that changes in temperature patterns could result in the tide rising and floods happening.

Changes in high tides have been shown to be directly correlated to the natural disasters, such as coastal flooding [2]. When the high tide becomes much higher, there is a trend that floods occur more, so better understanding these trends can directly help cities and researchers become better at preparing for these natural disasters. Other factors that also impact flooding include precipitation, temperature, humidity, elevation, and more, but for this study, we focus mainly on tidal information.

The main goals for this project were to develop a model to predict tidal changes in Vietnam based on historical tidal information, and to be able to visualize that knowledge through an automated pipeline. While others have worked on analyzing tidal changes before, by focusing on Vietnam and passing this data into a custom-built pipeline for visualizing geographical data, we see this work as leading the way for any other important geographical data that could help a great number of people by being visualized.

One key aspect of this project is that it is mainly focused on the major cities in Vietnam. To identify these cities, we took the 20 most populated cities in Vietnam. We picked 20, because we needed to train 20 individual LSTM models, and we wanted a sustainable number that would also be able to be visualized. By picking the most populated cities, we are able to show a solid proof of concept for the model that works for a large percentage of the population.

The total sum of the populations for the 20 major cities in Vietnam, ranked in terms of the population, is 24.571 million. There are a total of around 98 million people in Vietnam, which means that the top 20 major cities covers exactly 25 percent of Vietnam. Thus, this would be a useful proof of concept for tidal prediction.

The contribution of this paper is that it presents a scalable, end-to-end system for going from cities to predictions, to visualizations. This automated pipeline can be applied to any geographical prediction system, meaning that it can be extremely helpful for smart city planning and development. In this case, the application was to 20 major cities in Vietnam, but this end-to-end pipeline could be expanded to any country in the world with any type of prediction algorithm.

2 RELATED WORK

There have been several other papers that have worked on tidal prediction in the past. There are two main areas of research related to tidal prediction, but none of them have an end-to-end system that automatically pulls data, trains a model, and visualizes the model.

The first area of related work is tidal and sea level prediction. For example, in one study, Imani et al. were able to use an extreme learning machine and relevance vector machine to predict sea level along the Chiayi coast in Taiwan, and achieved an R-squared coefficient of 0.93 [3]. In another study, French et al. were able to look specifically at ports and used an artificial neural network to create short-term forecasts for extreme water levels at ports [4]. In a study by Granata et al., researchers were able to create several machine learning models for tide level prediction in Venice, specifically a M5P regression tree, Random Forest, and Multilayer Perceptron, and they achieved a coefficient of determination between 0.924 and 0.996 using the M5P regression tree method [5]. Clearly, there has been successful work done in the past with tide prediction.

The second area of related work is natural disaster prediction in Vietnam. As opposed to the first area, this area of related work is much more closely related to Vietnam and to making important predictions for natural disasters such as flooding and typhoons. In one study, Luu et al, used several different machine learning techniques to create a flooding susceptibility assessment map, and found that there was an area that spans 829 square kilometers in Vietnam that has a very high risk for flooding [6]. In another study, Mai et al. created a hydraulic model for flood flows using geospatial analysis tools, but they did not create any machine learning-based model to make future predictions [7]. In a third study, Loi et al. worked on real-time flood forecasting for the Vu Gia-Thu Bon river basin using several geospatial analysis tools and were able to predict stream flows in the basin with a R-squared score of 0.95 [8]. There have also been several other studies that have looked into natural disaster monitoring in different parts of Southeast Asia [9-11]. Clearly, there have been several successful attempts at flood monitoring specific to Vietnam.

For the first area of related work, there are three key gaps in the related work for tide and sea level prediction. The first is the fact that none of these cities were done for Southeast Asia, which is easily one of the most natural disaster prone parts of the world. In this paper, we focus on Vietnam, which is much more strongly affected by tidal events due to its low elevation, which means that this study would be more valuable to those in similar geographic regions. The second gap that this paper addresses is the fact that there is no scalable or automated way of retrieving data. In each of these related work sections, the authors found data by manually getting it, but they did not have an automated approach to retrieving data. If the Imani et al. team, for example, wanted to be able to use the same algorithm on a different country, they would need to manually find the data and insert it into their model, meaning that the model wouldn't be as easily scalable. The third key gap is that no work has been done on visualization. Each of these pieces of related work makes a prediction, but they do not discuss how those predictions can be visualized in an integrated way that is explainable, which is another area that this paper addresses.

For the second area of related work, there are two key gaps that have not yet been explored. To begin with, many of the research papers for flood prediction in Vietnam just focus on one river, or one specific region in Vietnam. This results in the predictions not being scalable to other parts of Vietnam, which means that they would not be able to benefit all major cities across Vietnam. The second area that this related work still leaves unaddressed is the concept of visualizing these results. While these papers have used geospatial mapping tools like ARC-GIS, these tools do not scale for larger applications, or when it comes to integrating visualizations in other websites.

In this paper, we address all five of these missing gaps covered across both of the papers. Our pipeline makes accurate predictions using an LSTM model, is easily scalable across any number of major cities in Vietnam, includes a GeoJSON conversion that, when combined with MapBox, results in flexible visualizations stored in the cloud, and is specific to Vietnam's tidal prediction. In addition, since tidal prediction is a key indicator for several different natural disasters, this work presents a step forward in an indicator that can be applied to several different types of natural disaster prediction and can be scaled across the world.

3 METHODS

3.1 Methods Overview

The diagram in Figure 1 shows the entire data pipeline, from collecting the data, to processing it, to analyzing it, to visualizing it. Each of the parts of the pipeline is linked to its corresponding section in the paper as well.

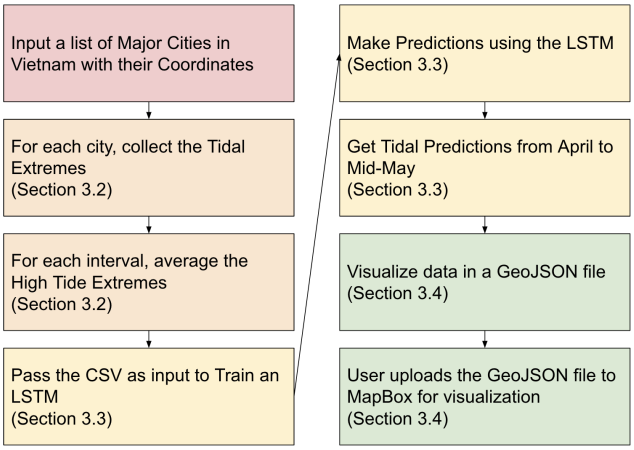

Figure 1: Methods Overview Diagram

3.2 Retrieving Tidal Data

We gathered the tidal information data from the StormGlass API, specifically from the Tidal Extremes API. The input to the API is the latitude and longitude of a given point, and the output is the set of tidal extremes, including lows and highs, from that given day at that location. The data is sourced from a variety of water stations from throughout the world [12].

To determine which 20 major cities we wanted to use, we found the population for each of the major cities in Vietnam, ranked. We found that the three most populated cities in Vietnam are Ho Chi Minh City, Hanoi, and Da Nang. Within these three cities, Ho Chi Minh City has 8.993 million people, Hanoi has 8.053 million people, and Da Nang has 0.989 million people.

Once we had the list of cities, we also supplied the latitude and longitude for each of those cities, sourced from online. Next, we got the average of the tidal highs and got that information stored as a CSV file.

For the dates, for each city, we got the tide information for the days at 01, 14, and 28, from each of the months 01-12 from each the years 2017, 2018, 2019, 2020, and 2021. This resulted in a total of 3*12*5 = 180 data points for each of the 20 major cities.

To justify the choice for 180 data points, 5 years of data is the enough data to be able to generalize patterns, while making sure that the data is relatively recent.

In addition, tidal information was retrieved for days 1, 14, and 28, for a few key reasons. The first reason is that daily data is extremely variable. The weather on one given day is not something that can be generalized across years, but weather every few weeks is more indicative of seasons, and thus is something that can be generalized. In addition, retrieving data for every day could result in the model overfitting based on daily patterns from one year, when those same patterns would not be able to be applied to another year.

```
"Day","Tide"
2017-01-01,1.0385110023476865
2017-01-14,1.2303949778619292
2017-01-28,1.101360295180071
2017-02-01,0.982299962714531
2017-02-14,0.9598551801841988
2017-02-28,1.1746122881875463
2017-03-01,1.1504960699490994
2017-03-14,1.046490142938796
2017-03-28,1.1987541445108714
2017-04-01,1.0723655739653606
2017-04-14,0.9399715223843609
2017-04-28,1.2999455039935452
2017-05-01,1.0062152179074655
2017-05-14,0.9550373044605857
```

Figure 2: Tide Data CSV file for Ho Chi Minh City.

3.3 Predicting Tidal Data

In order to predict tidal data, there are several different types of methods that could be used. One of the most popular methods for time-series based prediction is known as a LSTM, or Long Short Term Memory Neural Network. The LSTM is extremely effective at time series based prediction, since the structure of the neural network supports sending and receiving signals over time. Given that this project was focused on making predictions every 2 weeks for 5 years, the LSTM was one of the best choices possible to come up with a simple yet effective model to predict time-series based tidal data.

For each city, we created and trained an LSTM on the information for the city. The LSTM was created as shown in Figure 3.

In terms of parameters, the LSTM was trained using a look back window of 10 time intervals, the LSTM was created with an input 5 units, and the LSTM has a dense layer with an input of 1 unit. The model was compiled with mean squared error with the Adam Optimizer. The choice for the LSTM structure was mainly based on

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 5)                 320

 dense (Dense)               (None, 1)                 6

=================================================================
Total params: 326
Trainable params: 326
Non-trainable params: 0
```

Figure 3: LSTM Structure.

the fact that the data is linear and only has the one tide variable, which means that simpler models perform much better, which resulted in the LSTM structure shown in Figure 2.

Prior to training the model, the data was split into 70 percent for training, and 30 percent for validation. The model was trained for 25 epochs, since further increases in the number of epochs did not significantly increasing the accuracy of the model. The model's predictions were directly compared to the ground truth values, and the RMSE for training and testing were both calculated for each of the 20 major cities. These results are described in more detail in the results section. We retrieved the predictions for 04/01, 04/14, 04/28, 05/01, and 05/14, since those are 5 of the key dates that have the highest likelihood of natural disasters.

3.4 Visualizing Tidal Data

Once we had the predictions, we stored them all in 5 separate files. The visualization script takes each of the files and converts them to GeoJSON formatted files with points, that are encoded based on the predicted tide. These encodings assign a color to different ranges of tides, where tide values less than 0.2 are assigned to be "White", tide values between 0.2 and 0.4 are assigned to be "Yellow", and so on for the colors "Light Orange", "Orange", "Light Red", and "Red". These severity categories are then included in the GeoJSON format along with the latitude and longitude for the actual geographical data points.

4 RESULTS

4.1 Model Results

After training and validating the model, one key method for visualizing results is to display the actual and predicted data to see if there are any common trends. For these graphs, we had three different lines, each with a different color, as shown in Figure 4. For each graph, we plotted how the tide level changes over time.

The Predicted vs Actual graphs are shown for Ho Chi Minh City, Hanoi, and Da Nang in Figure 4. One interesting trend is that Hanoi's predictions are much more spiked with a higher frequency than the others, which indicates that tides are much more volatile in that region. Another interesting trend was from Nha Trang, which has one of the most irregular patterns, and is shown in Figure 4. By contrast to Nha Trang, Ho Chi Minh City and Da Nang have very stable predictions that seem to follow similar patterns for increases and decreases over time.

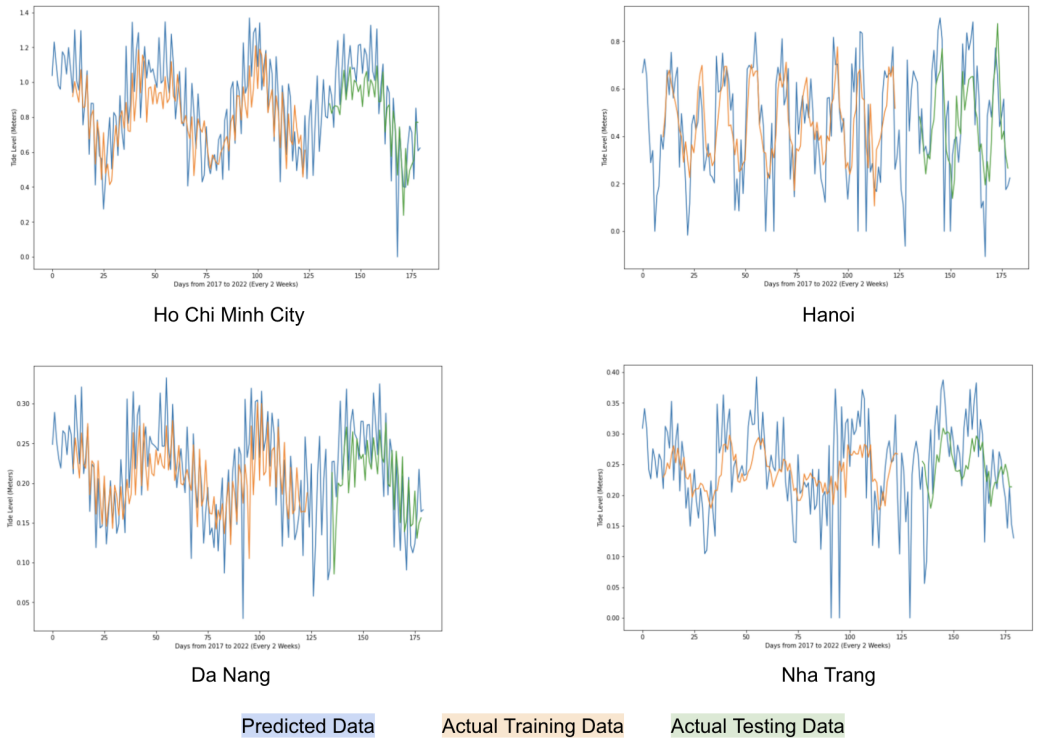

Figure 4: Predicted vs Actual Tide for different cities in Vietnam

4.2 RMSE Results

As an evaluation metric, we used RMSE, or Root Mean Square Error, to measure the deviation between the actual and predicted values for each of the data points. The RMSE values for the most populated cities in Vietnam are shown in Figure 5.

City Name	Train RMSE	Test RMSE
Ho Chi Minh City	0.1654	0.1687
Hanoi	0.1646	0.1934
Da Nang	0.0415	0.0450
Haiphong	0.1683	0.1961
Bien Hoa	0.1548	0.1554
Can Tho	0.1754	0.1848

Figure 5: RMSE results for the largest cities

4.3 Visualization Results

There were 5 different graphs that were created, for each of the 14-day intervals during the peak flooding season in Vietnam. Each of these graphs was created using MapBox Studio after uploading the generated GeoJSON file from the last step. The visualizations for the graphs are shown in Figure 6.

4.4 Comparison of Results

As shown in Figure 5, the lowest RMSE for testing was 0.0450. The great majority of the results from Figure 5 were between an RMSE of 0.15 and 0.2, which shows that the model is able to make relatively accurate predictions overall, even though there could be a great deal of variability for certain cities. In addition, the fact that the RMSE was under 0.05 for DaNang indicates that the prediction algorithm is extremely effective for cities with reliable weather patterns, which means that the model also has the potential to make accurate tidal predictions for smaller cities in Vietnam that have stable weather patterns.

Comparing these results to those of the related work, one key distinction is that the related work used different algorithms, like relevance vector machines, random forest, multilayer perceptrons, and more. In several cases, the related work used R-squared scores, and they were able to achieve scores above 0.9, which demonstrates that their models were very accurate. Since this research study is focused on time-series based predictions rather than real-time predictions or hydraulic models, one important contribution that this paper brings is that it shows the results of using an LSTM Neural Network, as opposed to other algorithms that are more focused on predicting data at one given point in time.

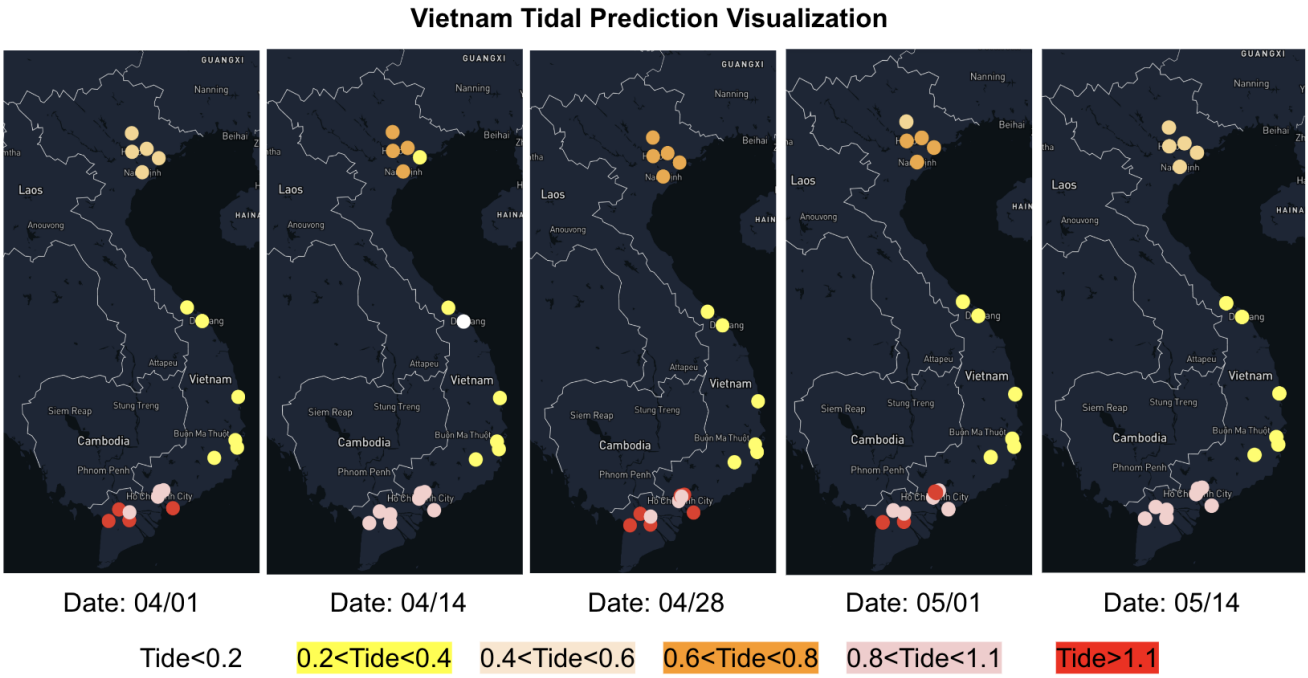

Figure 6: Vietnam Major Cities Tidal Severity Visualization

Another key difference in the results between this study and others is in the ability to visualize predictions. While other studies focused on being able to make predictions, none of them focused on finding a scalable, end-to-end method for going from input data to visualizations of those predictions. As shown in Figure 6, the final result from this research was not just an accurate prediction, but also a visualization on a map that can be created for multiple different time periods. The visualization results from this study are different than those from the results of other studies, because the visualization results from this study cover multiple different points in time.

5 DISCUSSION

5.1 Comparison to Related Work

In comparison to the related work, this paper addresses all five major gaps from the two areas of related work.

For the first area of related work, tidal prediction, this paper presents an approach that can be done for cities in Southeast Asia, specifically Vietnam. As opposed to other studies that were mainly focused on areas that are not as prone to flooding, this paper presents a system that has the potential to be much more impactful, since it can be applied to areas that would be more prone to flooding, and thus would benefit more from information about tidal predictions. Another important contribution this paper presents is the ability to automatically retrieve data, and thus be easily scalable by allowing users to enter more input cities. The studies from the related work were mainly focused on being able to make predictions for one area, but this paper presents an approach that can be applied to any area, which is an important contribution. Lastly, this paper also includes visualizations of the final predictions over multiple time series points, as shown in Figure 6. As opposed to the related work, this study presents a method for visualizing time-series based data, which can be extremely helpful for researchers that are trying to find tidal patterns over time.

For the second area of related work, natural disaster prediction, this research is easily scalable beyond just one river or city in Vietnam, and can be used in any part of the world. One key aspect of the automated data pipeline is the fact that the system can collect tidal extremes for any city in the world. This means that the input cities provided to the model could easily be changed, which would result in the tidal data being retrieved for those cities, and the model training and testing on the data for that city. As opposed to many of the related work that used a fixed region or country, this is an important contribution, since it allows the system to be easily applied to areas that needs visualizations of tidal predictions. Secondly, this study also allows results to be visualized, rather than just predicting natural disasters. Although some of the related work did have visualizations, many of these visualizations were done using geospatial analysis tools. In this paper, the visualizations are stored in a GeoJSON format, and then can be uploaded to visualize, which makes it much easier to get information from different data points and visualize them across many different types of visualization platforms.

5.2 Impact

The main benefit this research presents is the unlimited scalability of such an end-to-end pipeline. Although this project focused mainly

on tidal predictions for 20 of Vietnam's major cities, the cities could easily be expanded to hundreds, and the locations could be placed throughout the world. This level of scalability allows others to not just find data, but to make intelligent, real-time predictions for that data, and visualize those predictions.

In terms of tidal information, this project also involved develop an LSTM model that accurately predicts tides. As displayed in Figure 5, the test RMSE for Da Nang was 0.0450, which means that the LSTM model has the potential to be extremely accurate for certain large cities. Even for cities with larger RMSE values, like Ho Chi Minh City, the RMSE values are still relatively low, at 0.1654, given that most of the tide numbers are close to 1 for their value.

The main impact of this research is that it directly helps city planners who want to be able to see which areas are more prone to flooding for faster response. By creating visualizations of tidal severity in different areas across Vietnam and seeing how those trends change every few weeks, it can be much easier to see which areas would need more resources for faster response. Thus, the main benefit of this end-to-end platform is in being able to create an intelligent visualization that can be helpful for people across cities to keep track of tidal patterns through visualizations.

5.3 Limitations

One key limitation is that this paper focused on 20 major cities, which means that the results are not fully representative of smaller cities in Vietnam. In addition, the program retrieves information every 14 days, which means that it might not be as good at predicting more granular trends that would happen daily. Another limitation is that the RMSE is not as low as possible for more variable cities, like Hanoi, which means that the model would need to be further improved to become more accurate for different cities. In order to address these limitations, more data can be collected across more cities, covering more days, and multiple different types of models can be tested in the future to increase the accuracy of the final machine learning model.

5.4 Future Work

In the future, there is a great deal of work that can continue to be done. To address the limitation on the number of cities, more cities in Vietnam, such as minor cities, can be included in the research. To address the limitation of retrieving information every 14 days, the data collection stage can iterate through every day in the year, which would help develop more granular data trends. This could further be improved by making predictions hourly based on tidal trends.

Another interesting area for future work would be combining tidal predictions with other indicators. This paper mainly focused on predicting tides, but other indicators could be used to detect natural disasters, such as sudden changes in temperature, humidity, or precipitation. In addition, more work can be done on detecting natural disasters not just in Vietnam, but to other countries across the world, due to this pipeline's scalability.

6 CONCLUSION

Overall, the only inputs to this data pipeline were the list of major cities and the locations of those cities. The tidal information from the past 5 years was retrieved automatically, 20 different LSTM models were trained automatically, and the predictions were converted to visualizable GeoJSON formatted files for each date, automatically. As a result of this automated pipeline, all stages of the machine learning process, from finding data, to passing data into a model, to training the model, validating the model, and converting it to a visualizable format, have all been automated. This presents a step forward for end-to-end based machine learning prediction tasks, and the application of this created pipeline to tidal prediction shows the scalable impact this research can have on the world.

7 ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Division Of Computer and Network Systems under Grant Award #2026050.

REFERENCES

[1] Disaster management reference handbook - Vietnam (December 2021). ReliefWeb. (2021, December 10). Retrieved June 15, 2022, from https://reliefweb.int/report/viet-nam/disaster-management-reference-handbook-vietnam-december-2021

[2] Taherkhani, M., Vitousek, S., Barnard, P. L., Frazer, N., Anderson, T. R., & Fletcher, C. H. (2020). Sea-level rise exponentially increases coastal flood frequency. Scientific reports, 10(1), 1-17.

[3] Imani, M., Kao, H. C., Lan, W. H., & Kuo, C. Y. (2018). Daily sea level prediction at Chiayi coast, Taiwan using extreme learning machine and relevance vector machine. Global and planetary change, 161, 211-221.

[4] French, J., Mawdsley, R., Fujiyama, T., & Achuthan, K. (2017). Combining machine learning with computational hydrodynamics for prediction of tidal surge inundation at estuarine ports. Procedia IUTAM, 25, 28-35.

[5] Granata, F., & Di Nunno, F. (2021). Artificial Intelligence models for prediction of the tide level in Venice. Stochastic Environmental Research and Risk Assessment, 35(12), 2537-2548.

[6] Luu, C., Bui, Q. D., Costache, R., Nguyen, L. T., Nguyen, T. T., Van Phong, T., ... & Pham, B. T. (2021). Flood-prone area mapping using machine learning techniques: A case study of Quang Binh province, Vietnam. Natural Hazards, 108(3), 3229-3251.

[7] Mai, D. T., & De Smedt, F. (2017). A combined hydrological and hydraulic model for flood prediction in Vietnam applied to the Huong river basin as a test case study. Water, 9(11), 879.

[8] Loi, N. K., Liem, N. D., Tu, L. H., Hong, N. T., Truong, C. D., Tram, V. N. Q., ... & Jeong, J. (2019). Automated procedure of real-time flood forecasting in Vu Gia–Thu Bon river basin, Vietnam by integrating SWAT and HEC-RAS models. Journal of Water and Climate Change, 10(3), 535-545.

[9] Chen, C., Jiang, J., Liao, Z., Zhou, Y., Wang, H., & Pei, Q. (2022). A short-term flood prediction based on spatial deep learning network: A case study for Xi County, China. Journal of Hydrology, 607, 127535.

[10] Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. Water, 10(11), 1536.

[11] Li, X., Yan, D., Wang, K., Weng, B., Qin, T., & Liu, S. (2019). Flood risk assessment of global watersheds based on multiple machine learning models. Water, 11(8), 1654.

[12] Storm Glass: API documentation. Storm Glass | API Documentation. (n.d.). Retrieved June 15, 2022, from https://docs.stormglass.io/