**Predictive modelling of sea debris around Maltese coastal waters**

**The Data**

**u:** The east-west component of the current velocity, representing movement in the eastward (positive u) or westward (negative u) direction. It's a measure of how fast the water is moving across the Earth's surface in the horizontal plane from west to east or vice versa.

**v:** The north-south component of the current velocity, representing movement in the northward (positive v) or southward (negative v) direction. Like ‘u’, it measures the speed of water moving in the vertical plane from north to south or the opposite.

Together, u and v components allow for the representation of the direction and speed of sea surface currents at a given point. By combining these two components, one can calculate the total velocity and direction of the current at that point.

The magnitude (or speed) of the current can be calculated from the u and v components using the Pythagorean theorem. If you consider u and v as the legs of a right triangle, the magnitude of the current (hypotenuse of the triangle) is the square root of the sum of the squares of u and v.

The direction of the current, measured as an angle from the north (often in degrees or radians), can be calculated using the arctangent of v over u, which gives the angle θ.  
  
**Dimensions:**

**time:** Hourly Intervals (amount depends on period)

**lat:** 52 latitude points, indicating the geographic scope in the north-south direction.

**lon:** 43 longitude points, indicating the geographic scope in the east-west direction.

**Original Boundaries: (***DOUBLE CHECK)*  
Latitude Range in the Dataset: 35.74470138549805 to 36.88019943237305

Longitude Range in the Dataset: 13.676799774169922 to 15.380399703979492

**Updated for Lagrangian:**lon\_min, lat\_min, lon\_max, lat\_max = 14.15, 35.79, 14.81, 36.3

**For simulation to prevent stuck boundaries:**

extent = [14.15, 14.71, 35.79, 36.2]

**Final polygon for predictions:**

polygon\_coordinates = [

(14.6, 35.87),

(14.35, 36.01),

(14.35, 36.09),

(14.6, 36.09),

(14.6, 35.87)

]Add some information about the sea surface currents preprocessing file! (Remember that the nan values are not removed in this) this notebook is only to merge the data we have into a single file!

Add graphs for how the data is and explain why it was taken from 2020 till august 2023. (since there is a lot of missing data at the end of 2023). Actual Data starts at February 25th 2020 since that is the first date of the data that was given.

* Would be a good idea to do something similar to:  
  A graph of a number of data

  Description automatically generated

**Lagrangian Model Simulation**

* Mention pre-processing of the data, how the current data was merged and the length (no. of days), as of writing is 6 months of 2023.
* Simulation can be for same length used to train AI model or for brevity just simulate for only 7 days.
* Originally the idea was to use and combine wind and current data, but we decided to only use the sea surface currents data. This decision was based on two factors. Firstly, sea surface debris is proven to be more affected by sea surface currents instead of wind as seen in *[https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0111913]… In this study, researchers analysed data collected during five expeditions across all major ocean basins between 2014 and 2015 using a variety of methods including visual surveys, net sampling, and acoustic imagery. They found that "the majority of microplastics were concentrated in subtropical gyres" and that "wind-driven processes played only a minor role in determining the spatial distribution of microplastics." This finding suggests that while wind does contribute to some extent, it has less significance compared to ocean currents when discussing the distribution of sea surface debris.* Although it must be said that the wind data would make this more accurate in a real-life scenario. Secondly, we decided to do this to lessen the amount of work that had to be done.
* Original simulation was for 3 years but after some consideration, this proved to be too long and pointless. This decision was originally taken since initially we wanted to make predictions for a month but after some consideration this was not possible to implement and therefore we decided to simulate the Lagrangian model for a shorter period and make a prediction for a total of 1 day instead.
* First step was to create a land-sea mask. This was done by using a shapefile, This was then rasterised, and land is represented by 1 and sea by 0. Then it was converted to NETCDF format and then added within the grid boundaries to match with the boundaries of the data set. Show outputted image here. Can also mention how shapefile was used as map of Malta.
* Important to mention that due to the radars only north part of the island there is data available.
* Created the Lagrangian pipeline, and custom kernels (explain more in detail how they are deleted at boundaries and how the land sea mask reflects or deletes.. find reference for probability of being deleted or reflected).
* From simulations we noticed how particles were getting stuck at boundaries. This was found out to be since we were using all the dataset there was no data at the borders, so the particles were not affected. This was solved by using a bit smaller boundary (by like 00.1).
* Originally I was going to have multiple particles random but then it was more realistic to have around 50 (1 particle represents a group of debris) starting from very close proximity to simulate clusters of debris how they move.
* Simulation run for period of dataset, dt (elaborate) saved every hour? Elaborate as well.
* Decided to pre-process the input data further to remove any nan (not a number) values from the ‘u’ and ‘v’ components. This was done since nans can lead to numerical integration issues, interpolation problems, propagation error and also physical interpolation issues.
* When it comes to the missing data in the Lagrangian model, (find reference) the model has built in way of interpolating the data based on values around it so no preprocessing is necessary when it comes to this.

**Lagrangian Model Evaluation**

* Use trajectories for drifters, start particle (drifter) with the same time, date, place and visualise.
* Compare trajectories with a particle/s.

**AI Model**

* Original was to train on 3 years single model and then predict 1 months. This was unfeasible. Then the next test was to train model on 7-day dataset and predict 1 day but making one model was hard due to the dimensionality of the data (4-d) and the predicted results were not accurate at all. Therefore, we decide to move to a smaller portion of the data (smaller boundaries) but longer period to predict the ‘u’ and ‘v’ components individually (for every pair of lon and lat) and then combine the predictions into one file to make the simulation from it.
* For this we decided to try and use 6 months of data and then also we decided we could test using 6 months but skipping a month (jan,mar,may…) because of the seasonal changes in the data.
* The prediction is for 24 hours. The prediction will be converted into NetCDF format so that we can simulate it using the Lagrangian model.
* First we built a normal LSTM to make a single target prediction ‘u’. Then we tried the same LSTM for two target predictions ‘u’ and ‘v’. Then we tried to find the best model (LSTM,GRU…) and the bets hyperparameters. Finally, we run a model for every pair and then merge into single output.

ACTUAL IMPLEMENTATION:

* Data extracted from 2023-01-01 till 2023-06-30 (6 months)
* Pre-processed data by dropping useless columns and removing rows where u and v are nan. First experiment we also dropped v so that data will be time and u to predict u.
* Removing nan values resulted in some points having way less data this could be seen on the visualisation.
* Number of pairs was reduced from 41 to 37…
* Because dropping nans resulted in a lot of missing data, we decided to use 1 year (22-01-01 till 22-12-31) worth of data to ensure there was enough data for every pair (since different pairs had different amount of data)
* This resulted in a fair number of rows per pair of lat and long.
* Due to several nans the model input ended up only having on average around 4k inputs per lat and lon pair, this led to a very small val and training loss value in the first epoch. To counter this we decided to use 2 years’ worth of data (2021 and 2022).
* The model for a single target prediction gave very good results. When the model was updated to have two targets (u & v) to predict the results were not as good. This gave thought to have separate models for both variables… to ensure that the results are accurate as possible.
* After multiple tests, it was discovered that a sequence length (window\_size) of 1 gives the best results.
* We also tried to use ‘v’ and ‘u’ (both as features) and from this we got slightly better results, so we concluded that they have a better effect and thus we decided to shift focus on using 2 feature to predict a single feature.
* From multiple tests it was concluded that a learning rate of 0.001 performs the best.
* Run same tests but this time with GRU and it was concluded that GRU has overall better performance.
* Regularization techniques (e.g., dropout, L1/L2 regularization) was added help mitigate overfitting since small window sizes might lead to overfitting.
* From further testing, we can conclude that having two features to predict a single target will yield very similar results as using a single feature to predict a single target. Although accuracy was always slightly better when using two features for a single target.
* Due to the Nans in the dataset we also tried to use interpolation to fill these values (spline) but it was concluded that the results were not good and worse than when we drop the nans. This is because… (give reason).
* Important to mention that all these tests where run on the same pair of lon and lat:

target\_lat = 36.03409957885742

target\_lon = 14.528599739074707

* Also tried LSTM with diff dropout and regularisation and got very similar results so we decided to go with the results that yielded the best overall results.
* Ideal batch size was 10 after some testing.
* It was decided that the prediction pipeline should change. Instead of creating a model to predict the next hour based on the previous hour, the new model takes the last 72 hours and make 1 prediction for 24 hours… *for now I decided to fill in nans with mean value, but this might change after some testing. I am also thinking of adding rolling forecasting… basically completely changing the sequence and feature inputs!*
* After some consideration and testing, a LSTM was created that was able to make predictions based on the last 24 hours and then use rolling forecast to predict the next 24 hours. The model had reasonably good results (add results) and a discovery was made, since we are working with sea surface currents it will be very hard to predict the next day’s weather (even with rolling predictions) just because of the fact of how unpredictable it can be. For example, one day can, be calm and suddenly, not visible over couple hours the values are way higher. This could be studied further where other weather conditions like weather could be used with the data and the training so that it is more evident if there is going to be a switch in the weather (add more useful input features).
* Changed to a longer window size of 72 days instead of 24 and got better results. Added interpolation to the input file since we need 72 and will not work if we drop nans.
* The new dates are:
* 2020/01/01 till 2023/08/01 🡪 The training data
* Input (72 hours from last sequence in test data) 🡪 2023/08/01 till 2023/08/03
* Actual (To compare) 🡪 2023/08/04.
* Explain exactly why I took (started from) the last sequence of the test set… although it can be any random day and it will work… we did this for brevity.
* Now experimenting with using GRU Layers instead of LSTM since they can be better for short sequence forecasts and take less time to train which will be useful when we make the pipeline for the 37 models.
* After some tests it was concluded that, although GRU’s take less time to train, the results were not as good as the tests that were done on the LSTM. GRUs simplify calculations compared to LSTMs, potentially speeding up training. However, this simplicity might limit their ability to model complex sequences with long-term dependencies as effectively as LSTMs, resulting in reduced performance on tasks where such intricacies are crucial.
* Since we are going to have to run an estimate of around 37 separate models to get a combined file that can be used for the Lagrangian simulation, the time that the models took to train was taken into consideration. Therefor a time of 20 mins was chosen to be the ideal time so that in total it will not take more than 12 hours to train the final multiple models. Early stopping was used to ensure that every model will have the best possible training and to prevent over fitting if the data for the specific point is less.
* It is for this reason that a learning rate of 0.001 was chosen instead of smaller rates like 0.0001 or 0.0005. This balance aims to reduce convergence time and improve results, as overly small learning rates may lead to slower progress and potentially getting stuck in less optimal minima.
* Created a loop pretty much does everything the same like the lstm we worked with before… the only difference is that it does this for every csv file corresponding to every pair of the coordinates.
* Even though there is a single model, it is re defined every loop. This ensures each dataset trains on a fresh model without previous weights. Declaring the model inside the loop ensures that a new instance of the model is created for each dataset. This approach is crucial when dealing with multiple datasets to avoid any data leakage or influence from previously trained models (clean slate training). It helps maintain the integrity of the learning process for each distinct dataset.
* Pretty much have a loop to extract the 3 days for the input per pair of coordinates and then the same thing but for the output 1 day for comparison.
* Important to mention that these have nans, so I had to interpolate since we need 72 time periods for the input to make a prediction and 24 to make a comparison… but it must be said that this means that the values are not that accurate.
* Mention why 70-15-15 split. Get reference.
* Since the training is in a loop and takes long, it will be done in batches of 10-10-10-7. The code was updated to use a more computationally expensive model architecture to allow for better predicted results. The output, metrics on test set and the graph for the test set is saved so it is not an issue to go back to them to use for evaluation. **Ignore what I said before about computational time!**

**AI Model Evaluation**

* Compare predictions (24 hours) to actual data of the same date of 24 hours.
* Can compare to trajectories as well.
* Create two Lagrangian simulation for the actual 24 our and for the predicted 24 hour and then have 2 visualisations on a single plot.
* To evaluate the accuracy of our models' predictions, we encountered a significant challenge. Our initial approach involved using data from January 1st to January 7th, 2023, as a basis for predicting values for January 8th. However, the presence of Nans in the dataset presented complications, as these gaps in the data made it difficult to generate hourly predictions for the subsequent day. To address this issue, we adopted a strategy that involved preprocessing the data, specifically by interpolating missing values to ensure a complete 24-hour dataset for the day preceding the prediction. This approach allowed us to proceed with making predictions, albeit with the caveat that the interpolated data may not perfectly reflect reality, potentially affecting the accuracy of our predictions. Consequently, this introduces a degree of uncertainty regarding the reliability of these predictions, suggesting that our evaluations might not be very accurate.
* Mention that the 3 days that the prediction was used is the last day of the test set and the next two dates. This was done so that the model will be more accurate since it uses close dates from the training even though it is unseen will give better predictions… Find a reference about this.
* The best thing to do is take an average of the error metrics for all the models. I can compare the whole combined predictions to the whole combined actual values, but this can be problematic since there is so much missing data.
* Evaluation can be done visually by using the Lagrangian Evaluation notebook to see a video simulation of the actual values and the predicted values overlapped.

**Future Work**

* Incorporate wind into the simulation and the prediction to make the model more accurate in real life scenarios.
* Jellyfish/movement and search and rescue.
* Can move onto ensemble learning (elaborate).
* Individual models with shared learning, train separate models for individual (lat & lon), then leverage insights from clustering for better models. Can also experiment with clustering common data together and making predictions per cluster. (Elaborate).
* Can make area bigger and have more than the 37 total models. Also, can have a more complex model like a transformer.
* Get more weather data that will influence surface debris movement and create a method with more features for a more realistic and accurate prediction! These can also be predicted as new targets and the Lagrangian to be updated to use this new data.
* Model that predicts the missing values in the datasets.

**Miscellaneous**

* Put Adam in Acknowledgments.
* REDP Application ID: ICT-2023-00203
* Mention Rubber Duck Incident.
* Tidal height was ignored (Reference – Godin 1983)
* Nan values where ignored (elaborate)
* Mention that there exists ARIMA for example for same situation, this will emphasize how different and why we took a different approach.

**Results**

LSTM Single Target U (2 years):

|  |  |
| --- | --- |
| Test MSE | 0.0024190667630645227 |
| Test R-squared | 0.8156669780585895 |
| Test Explained Variance | 0.816099923155643 |
| Test MAE | 0.021644903820475958 |

LSTM Single Target V (2 years):

|  |  |
| --- | --- |
| Test MSE | 0.0026758533046433593 |
| Test MAE | 0.032873101353487616 |
| Test R-squared | 0.796099829302783 |
| Test Explained Variance | 0.8053555550911925 |

LSTM Single Target U Simple (3 years):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0018392425899589357 |
| Test MAE for 'u' component | 0.028609757062382415 |
| Test R-squared for 'u' component | 0.6981919036542692 |
| Test Explained Variance for 'u' component | 0.6981958637429437 |

LSTM Single Target V (3 years):

|  |  |
| --- | --- |
| Test MSE for 'v' component | 0.0034144461552393186 |
| Test MAE for 'v' component | 0.0397711281401556 |
| Test R-squared for 'v' component | 0.7681236805037378 |
| Test Explained Variance for 'v' component | 0.7688760337222249 |

LSTM Single Target U (24 Sequence):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0036544685574883087 |
| Test MAE for 'u' component | 0.04077346492489306 |
| Test R-squared for 'u' component | 0.7214374535512165 |
| Test Explained Variance for 'u' component | 0.7528849249465257 |

LSTM Single Target U (12 Sequence):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0029143808739715525 |
| Test MAE for 'u' component | 0.034527283585515066 |
| Test R-squared for 'u' component | 0.7777920622305681 |
| Test Explained Variance for 'u' component | 0.7784308998641449 |

LSTM Single Target U (1 Sequence):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0026107904748576068 |
| Test MAE for 'u' component | 0.03225322956613726 |
| Test R-squared for 'u' component | 0.8011538230713655 |
| Test Explained Variance for 'u' component | 0.8101298180496459 |

LSTM Single Target U (0.001 LR):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.002342229529907343 |
| Test MAE for 'u' component | 0.029819278266926652 |
| Test R-squared for 'u' component | 0.8216082860740367 |
| Test Explained Variance for 'u' component | 0.8216656680894614 |

LSTM Single Target U (with ‘v’ as feature):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.002516725745286363 |
| Test MAE for 'u' component | 0.03264639806964096 |
| Test R-squared for 'u' component | 0.808318094597248 |
| Test Explained Variance for 'u' component | 0.82340737786933 |

**From here on always 2 features and a single target**

LSTM Single Target U (with adaptive LR starting at 0.001):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0023592781916186665 |
| Test MAE for 'u' component | 0.03052777880046811 |
| Test R-squared for 'u' component | 0.8203098053128676 |
| Test Explained Variance for 'u' component | 0.8203161559186333 |

LSTM Single Target U complex (3 years):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.001898580755587542 |
| Test MAE for 'u' component | 0.029220607353491038 |
| Test R-squared for 'u' component | 0.6886662720026921 |
| Test Explained Variance for 'u' component | 0.6945530437051486 |

GRU Single Target U (3 years, LR=0.0001):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0017845362995307333 |
| Test MAE for 'u' component | 0.02757052929739459 |
| Test R-squared for 'u' component | 0.7073675495528291 |
| Test Explained Variance for 'u' component | 0.7073919490494525 |

GRU Single Target U (2 years, LR=0.0001):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0024379899291823042 |
| Test MAE for 'u' component | 0.03051282151001504 |
| Test R-squared for 'u' component | 0.8143148669044942 |
| Test Explained Variance for 'u' component | 0.8168623228711309 |

GRU Single Target U Less complex(2 years, LR=0.001):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0022538887932932 |
| Test MAE for 'u' component | 0.02898659598288193 |
| Test R-squared for 'u' component | 0.8283365999360444 |
| Test Explained Variance for 'u' component | 0.8283392773560619 |

LSTM Single Target U Less complex (Interpolation, single feature):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.007416813195800896 |
| Test MAE for 'u' component | 0.04529537497181993 |
| Test R-squared for 'u' component | 0.519584878029119 |
| Test Explained Variance for 'u' component | 0.5203720938677607 |

LSTM Single Target U Higher dropout 0.3 (Adding onto best model):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0024214458955741304 |
| Test MAE for 'u' component | 0.03134766232455302 |
| Test R-squared for 'u' component | 0.8155749135706839 |
| Test Explained Variance for 'u' component | 0.8209414886247651 |

Batch size = 8

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.002179373371083336 |
| Test MAE for 'u' component | 0.02803178147733633 |
| Test R-squared for 'u' component | 0.8340119335069861 |
| Test Explained Variance for 'u' component | 0.834027284219115 |

Batch size = 16

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0023517817452135712 |
| Test MAE for 'u' component | 0.03077673077260262 |
| Test R-squared for 'u' component | 0.820880758716658 |
| Test Explained Variance for 'u' component | 0.8208942235949539 |

Batch size = 32

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.002271727808319735 |
| Test MAE for 'u' component | 0.028695319344842898 |
| Test R-squared for 'u' component | 0.8269779233312538 |
| Test Explained Variance for 'u' component | 0.8291700086264638 |

Batch size = 64

Very similar to normal

Batch size = 128

Very similar to normal

Batch size = 256

Very similar to normal

LSTM Double Target U,V (2 years):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.004009208428052819 |
| Test MSE for 'v' component | 0.008702654008220605 |
| Test MAE for 'u' component | 0.04196935631070439 |
| Test MAE for 'v' component | 0.0579934010020833 |
| Test R-squared for 'u' component | 0.6944980947116437 |
| Test R-squared for 'v' component | 0.40923099614634095 |
| Test Explained Variance for 'u' component | 0.722836121476568 |
| Test Explained Variance for 'v' component | 0.43345408386505246 |

LSTM Double Target U,V (3 years):

|  |  |
| --- | --- |
| Test MSE for 'u' component | 0.0022882551865674927 |
| Test MSE for 'v' component | 0.005073492373509212 |
| Test MAE for 'u' component | 0.03262867246795131 |
| Test MAE for 'v' component | 0.05051134610980628 |
| Test R-squared for 'u' component | 0.6245117715403712 |
| Test R-squared for 'v' component | 0.655457229349919 |
| Test Explained Variance for 'u' component | 0.6271222424285632 |
| Test Explained Variance for 'v' component | 0.6628467139706005 |