Temporal EMM for Time Series Prediction

Prathamesh Samal
Eindhoven University of Technology
Eindhoven, the Netherlands

Arthur Sliwinski
Eindhoven University of Technology
Eindhoven, the Netherlands

Divyansh Purohit
Eindhoven University of Technology
Eindhoven, the Netherlands

Myrna van 't Hof
Eindhoven University of Technology
Eindhoven, the Netherlands

Marceli Morawski
Eindhoven University of Technology
Eindhoven, the Netherlands

Abstract

Predictive analysis of time series data knows many application domains ranging from finance to healthcare. This field previously dominated by statisticians and econometricians, has experienced an influx of machine and deep learning models. Next to the fact that forecasting tasks are highly complex for time series data due to temporal complexities, these new Deep Learning methods like LSTMs are considered "black-box" models, which are notoriously hard to interpret due do to a lack of transparency and explainability. Local Pattern mining,or more specifically Exceptional Model Mining, subfields of data mining, might help tackle some of these problems by providing key insights into the subgroups of data for which a predictive model tends to perform exceptionally well or poorly.

This study broadens the SCaPE model class for Exceptional Model Mining (EMM) to temporal regression contexts to find subgroups where the performance of regression models differs from their overall behavior. This is done by the implementation of a LSTM model and the review and redefinition of the SCaPE inspired quality measure, which makes the method more robust for temporal data. Finally the adapted SCaPE, Temporal Q-Based Subgroup Discovery, is evaluated on real buoy data of the *National Oceanic and Atmospheric Administration*, which yields a variety of interesting subgroups.

Keywords

EMM, SCaPE, time series prediction, Black-Box interpretability, LSTM, LPM

ACM Reference Format:
Prathamesh Samal, Arthur Sliwinski, Divyansh Purohit, Myrna van 't Hof, and Marceli Morawski. 2025. Temporal EMM for Time Series Prediction. In *Proceedings of Conference on Knowledge Discovery and Data Mining (KDD 2026) (ACM SIGKDD '32).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

ACM SIGKDD '32, Jeju, Korea

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

A significant portion of global trade and industrial activity relies on maritime operations, including shipping, oil extraction, and naval logistics. Accurate wave height prediction therefore remains critical for ensuring the safety, efficiency, and scheduling of offshore activities [Cerqueira 2023]. Although, there has been advances in data-driven forecasting, the prediction of ocean dynamics remains challenging task due to its non-stationary and highly context-dependent nature [Hewamalage et al. 2022].

Deep learning models, particularly recurrent and sequence-based architectures, have shown strong predictive power for such tasks [Lindemann et al. 2021], but under changing environmental conditions and seasonal variations their performance often fluctuates. This poses a fundamental question: under what kind of conditions does a forecasting model perform particularly well or poorly? Answering this requires a deeper understanding of model behavior across different temporal and environmental contexts, besides high predictive accuracy.

Exceptional Model Mining (EMM) provides a framework for exactly this kind of model introspection that identifies subgroups in which a predictive model behaves differently from its overall trend. However, existing approaches such as SCaPE (Subgroup Characterization and Performance Explanation) [Duivesteijn and Thaele 2014] are limited to static datasets and cannot capture how subgroup behavior evolves over time. Temporal dependencies that are crucial in time series forecasting, remain unaddressed in this framework.

To bridge this gap, we extend the SCaPE framework into the temporal domain. In our proposed framework, we evaluate model performance across time windows, and enable the discovery of subgroups that exhibit distinct and consistent temporal behaviors. This adaptation allows EMM to work with sequential data, and provide insights into when and why model errors change over time.

2 Related Work

2.1 Time series prediction

The forecasting of time series knows applications in a variety of domains. Use cases vary and include finance [Dakhore et al. 2024], climatology [Sha and Guha 2023], [Cerqueira 2025] and healthcare [van der Schaar and Maxfield 2023]. A large number of varying techniques exist for performing predictive tasks on time series data including statistical, time domain, machine and deep learning methods. However, the violation of assumptions, non-stationarities and noise, do pose significant hordes for time series estimation

techniques [Hewamalage et al. 2022]. Furthermore models can be difficult to interpret. This is a major problem because important decisions, which require high levels of trust and transparency, are made based on these analysis and forecasts. Statistical models can already be difficult to interpret due to mathematical complexity, but in the case of so called black-box models, understanding model behavior becomes even more difficult. So although modern machine and deep learning techniques show great predictive performance, black-box models require careful consideration given the importance of interpretability, especially in domains like healthcare. [Sendak et al. 2020]

2.2 Local Pattern Mining (LPM)

Local Pattern Mining is a data mining field which aims to discover interpretable subsets of the data which behave significantly different from the dataset as a whole. These so called subgroups are defined by some predicate on a variable or a set thereof. A specific framework called Subgroup Discovery (SD) considers those subgroups who's predicate contains one variable, whereas Exceptional Model Mining (EMM) is able to find exceptional subgroups considering any number of explanatory variables. For example, the EMM method can be used to find those groups of patients which are at higher risk to develop certain complications after surgery [van den Biggelaar et al. 2025]. But it can also be used to evaluate model performance, the paper by [Du et al. 2025] proposes a framework named *Conformalized Exceptional Model Mining* which utilises EMM to discover those subsets within the data for which a model encounters difficulties or perhaps performs exceptionally well. It does this by presenting interpretable subgroups according to a quality measure, which in this case is designed to capture exceptional model performance, either good or bad.

2.3 EMM on Time-series

Through their explainability and interpretability the aforementioned methods might help overcome some of the problems in time series forecasting mentioned earlier. A well performing EMM framework could highlight data where a time series forecasting model might encounter predictive problems in a way that is interpretable to the person developing the model within a specific domain. This does not directly influence performance of the model itself, but it can provide the modeler with very useful insights into the model that otherwise would not be available.

A paper closely related to this use case of EMM, proposes *SCaPE* [Duivesteijn and Thaele 2014]. This is an EMM framework which performs a similar task to the previously mentioned *Conformalized Exceptional Model Mining*, but it is tailored to classifiers for which the ground truth is known. Capturing how a soft classifier model performs on specific subsets of data, and how this performance within subsets compares to performance on entire dataset. This more comprehensive analysis of data and model performance which is able to take into account multiple explanatory variables in describing the exceptional subgroups, makes SCaPE a very useful framework for modeling. The aim of our work is to expand the utility of this framework (SCaPE) allowing it to perform the same task on predictive time series models, and help alleviate some of the problems encountered in this type of modeling.

3 Methodology

In this section, we integrate an LSTM model with the SCaPE framework to analyze when and where the regression model shows unreliable performance. The pipeline mainly involves two consecutive phases

- A time series predictor based on LSTM for forecasting ocean wave heights
- A TQ-SD (Temporal Q-based Subgroup Discovery) system to identify notable temporal and environmental subgroups where the model's performance diverges from its overall behavior.

3.1 SCaPe

The SCaPE (Soft Classifier Performance Evaluation) [Duivesteijn and Thaele 2014] framework identifies areas within a feature space, where a classifier performs or does not perform well. Every subgroup in SCaPE is described as a combination of attribute-value conditions and is expressed as:

$$S_{SCaPE} = \{t \in D \mid A_1 \in I_1 \wedge A_2 \in I_2 \wedge \dots\}.$$

Assuming a dataset D of size n, where each record is of the form $x = \{a_1 \dots . a_k, r, b\}$. Here, $\{a_1 \dots . a_k\}$ represent the "*descriptive*" variables. On the other hand b and r represent the "*targets*". With r representing the soft classifier output, and b the ground truth.

In order to quantify model performance and define subgroup significance, the authors develop a few key and interesting metrics. First the *ARL*, *Average Ranking Loss* is defined. This metric quantifies how well r captures b. The *ASL* represents the *ARL* of a specific subgroup within the dataset N. A search algorithm then explores a variety of subgroups which are then ranked based on their quality measure, indicating exceptional performance, either good or bad. However, this framework assumes the following:

- Soft classifier with probabilistic outputs
- Static and i.i.d. dataset

Due to these presumptions, it is not suitable for time series prediction and regression, where the target variables are temporally associated and continuous. In order to overcome these constraints, we develop a temporal regression-based extension of SCaPE called Temporal Quality-based Subgroup Discovery (TQ–SD). TQ–SD captures temporal consistency and variability in model performance across time, in contrast to SCaPE, which finds static exceptional regions in a classification task.

3.2 Long Short Term Memory (LSTM) Model Class

A Long Short-Term Memory, or LSTM, network [Hochreiter and Schmidhuber 1997] is a type of neural network that aims to solve long-term gradient problems of the recurrent neural network. This recurrent neural network is an artificial neural network that leverages closed loop connections to feed previous outputs back into the network, allowing it to take previous values into account when predicting the next. This type is network is thus well equipped for predicting temporal data.

LSTM networks employ a special cell architecture to control the flow of information from long-term and short-term memories. One

such cell consists of three consecutive gates: the forget gate, the input gate, and the output gate. This helps regulate the flow of information and overcomes the issues with standard RNNs.

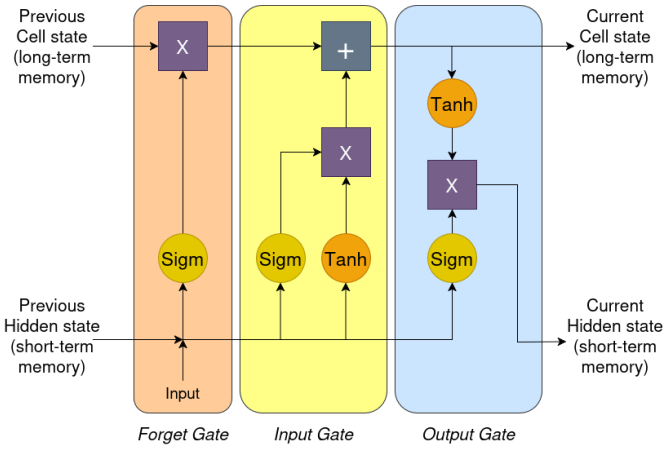

Figure 1: LSTM architecture

The ocean wave dataset can be represented as a time series

$$\{D = (\mathbf{x}_t, y_t)\}_{t=1}^{T},$$

where $\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(p)}]^\top \in \mathbb{R}^p$, p represents the number of features such as wind speed (WSPD), air temperature (ATMP), water temperature (WTMP) and $y_t \in \mathbb{R}$ represents the target variable, i.e the corresponding wave height (WVHT) at time t.

The LSTM model uses a sliding window of the previous L time steps:

$$\mathbf{X}_t = [\mathbf{x}_{t-L}, \mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_{t-1}],$$

and predicts the next wave height as

$$\hat{y}_t = f_\theta(\mathbf{X}_t),$$

where f_θ represents the trained LSTM model with parameters θ. The predictions $\hat{y}_t$ are then obtained for the test dataset. The model parameters is given in Table 1. The dataset is split into 80% and 20% training and testing parts respectively. It is optimzed using the Adam optimizer with mean squared error as the loss function. After training the model, residuals $(y_t - \hat{y}_t)$ are used for subgroup discovery.

3.3 Subgroup Construction and Quality Evaluation

The subgroups in this study are defined as below

$$S = \{t \in D \mid x_t^{(j_1)} \in I_1, x_t^{(j_2)} \in I_2, \ldots, x_t^{(j_k)} \in I_k\},$$

where I_j represents an interval that describes the range of the corresponding feature $x_t^{(j)}$. Continuous attributes are divided into quartile-based bins, where each bin represents an environmental characteristic. Each subgroup here represents a set of features from the dataset, using which the LSTM model performance can be assessed. An example of a subgroup where Wind Temperature is between 10°C to 20°C and Wind Speed is between 0 to 5 m/s is represented as follows

$$S = \{t \mid WTMP \in [10, 20], WSPD \in [0, 5]\},$$

The model performance of a subgroup is assessed using Root Mean Squared Error (RMSE) between the actual and predicted wave heights.

$$\text{RMSE}_S = \sqrt{\frac{1}{|S|}\sum_{t \in S}(y_t - \hat{y}_t)^2}$$

The ranking-loss measure used in SCaPE is replaced by this regression-based metric, which helps the framework to work with continuous variables.

Table 1: LSTM Parameter Configuration

Parameter	Value
Input Window Size (L)	24
Input Window Size (L)	24
Features (p)	12
Activation (Hidden)	tanh
Optimization	Adam
Loss Function	Mean Squared Error (MSE)
Batch Size	32
Train/Test Split	80/20

3.4 Temporal Q-Based Subgroup Discovery

In order to incorporate a temporal aspect to subgroup discovery, the dataset is divided into M segments $\{d_1, d_2, \ldots, d_M\}$, representing fixed month intervals. For each subgroup S, the subset of data belonging to time segment d_M is denoted as $S_{d,M}$ The RMSE for the temporal segment M is represented as $RMSE_{S,d_M}$. This metric is calculated for all time segments that will enable us to track performance of various subgroups over time.

After this temporal segmentation, it is very likely that each subgroup S has multiple RMSE values that correspond to different time segments, therefore we calculate the mean and variance of the RMSE for a given subgroup across all time segments. The calculated mean captures a long-term behaviour of the LSTM model for a particular subgroup across all time segments. It is also important to take the variance into account since it gives the degree to which the model's performance within a subgroup varies over time.

$$\overline{\text{RMSE}}_S = \frac{1}{M}\sum_{m=1}^{M}\text{RMSE}_{S,d_m},$$

$$\text{Var}_d(S) = \frac{1}{M}\sum_{m=1}^{M}(\text{RMSE}_{S,d_m} - \overline{\text{RMSE}}_S)^2$$

Upon combining these two metrics, we propose a temporal quality measure into a single metric. For a subgroup S, the quality measure is defined as follows:

$$Q(S) = (\overline{\text{RMSE}}_S - \text{RMSE}_G) - \lambda \cdot \text{Var}_d(S)$$

where $RMSE_G$ is the global RMSE across the entire dataset and λ is a parameter that controls the penalty on temporal variance.

The first term in the expression $(\overline{\text{RMSE}}_S - \text{RMSE}_G)$ captures the average deviation in performance of a subgroup from the global baseline. The second term $\lambda \cdot \text{Var}_d(S)$ penalizes time based variance. A penalty factor is important here in order to favour those subgroups which have a stable performance over time and do not fluctuate significantly. Therefore, this technique improves interpretability for sequential prediction tasks by extending SCaPE's static concept into a time-based framework.

3.5 Search Strategy

In order to identify subgroups that have extremely well or poor performance, a beam search strategy was implemented in our experiments. The search proceeds in the following steps:

(1) **Feature Discretization:** The continuous variables (such as temperature and wind speed) are transformed into a limited number of categorical intervals by discretizing them into quantile-based bins.
(2) **Initial Subgroup Evaluation** The temporal quality measure $Q(S_j)$ is used to evaluate each single-feature subgroup S_j separately. This helps in identifying the specific intervals or features that are most responsible for performance deviations from the global model.
(3) **Subgroup Refinement:** The subgroups with a high $|Q(S)|$ are combined with additional features to form refined subgroups, denoted as $S_{refined}$. Every refined subgroup ($S_{refined}$) is evaluated again across all time segments using the same quality measure. This helps to identify the combination of features that significantly influence the LSTM model performance.
(4) **Subgroup Ranking and Selection:** The remaining subgroups are ranked according to their absolute quality score $|Q(S)|$. High $Q(S)$ value suggests that the model performs worse than average, while negative $Q(S)$ values correspond to conditions of stable or improved performance.

4 Experiments and Results

4.1 Data

This study leverages historical data offered by National Data Bouy Center, which captures 24-hour every 10 minutes information about meteorological condition in various part of the world. Since there are 377 buoys in total, a heuristic was used to select the buoys for analysis. The selection was inspired by previous studies, under the assumption that those works—focused on time series forecasting—had already identified buoys with high-quality data. Based on article [Leibundgut 2021], two Hawaii buoys 51001 and 51101 were selected. These stations are close to each other which was crucial, in creation of the final data set. This will be further discussed in data reprocessing. In this report, the data covers the period from September 19, 2018, at 04:40 to December 31, 2018, at 23:40. This represents the most recent time span, chosen to minimize the number of missing values.

4.1.1 Variable Design. In the final dataset, significant wave height was selected as the dependent variable. This choice influenced the dataset's structure, as all data were aggregated to an hourly frequency, matching the interval at which wave height was measured. Additionally, thirteen other variables were included in the final dataset, three of which were engineered using the available data.

Table 2: Variable Descriptions — Original Features

Variable	Very short description	Type
Significant Wave Height	Avg of top 1/3 waves (m)	Continuous
Wind Direction	Coming-from direction (°)	Discrete
Wind Speed	Avg wind over 8 min (m/s)	Continuous
Gust Speed	Peak 5–8 s gust in 8 min (m/s)	Continuous
Dominant Wave Period	Period at max energy (s)	Continuous
Average Wave Period	Mean wave period over 20 min (s)	Continuous
Wave Direction	Dominant-wave direction (°)	Discrete
Sea Level Pressure	Pressure at sea level (hPa)	Continuous
Air Temperature	Air temp (°C)	Continuous
Sea Surface Temperature	Sea surface temp (°C)	Continuous

Engineered features were created to capture additional information not directly available from the original variables. Previous studies have analyzed the effect of the air–water temperature difference on wave height [Kettle 2015], which motivated the inclusion of this feature. Furthermore, visual inspection of the data revealed clear seasonal patterns. To account for these, two additional features — *isSummer?* and *Hour* — were introduced to analyze variations in behavior across specific seasons and times of day.

Table 3: Variable Descriptions — Engineered Features

Variable	Very short description	Type
Air–Water Temperature Difference	$\lvert T_{\text{air}} - T_{\text{sea}} \rvert$ (°C)	Continuous
isSummer?	1 if May–Oct, else 0	Binary
Hour	Measurement hour (0–23)	Discrete

4.1.2 Data merging and preprocessing. As mentioned earlier in this chapter, two buoy stations were used in the analysis to address issues of missing data. Both stations experienced some downtime during the six-and-a-half-year observation period. However, buoy 51101 had significantly fewer missing values (only 696 hours) compared to buoy 51001 (only 2904 hours), and was therefore selected as the primary station for analysis.

Because the two stations are located close to each other (13.85 km), the missing values in buoy 51001's data were imputed using readings from buoy 51101, based on the distribution of data points and their low Wassertein distance (indicating high similarity) 4. That way 209 hours of records were borrowed from station 51001. There were 487 hours where data from both stations were unavailable , the median values were used as replacements.

In total, the final dataset contained 55,100 hourly observations used for the overall analysis.
After resolving the missing data issues, a set of selected columns was normalized using a Min–Max scaler to enhance the performance of the trained LSTM model.

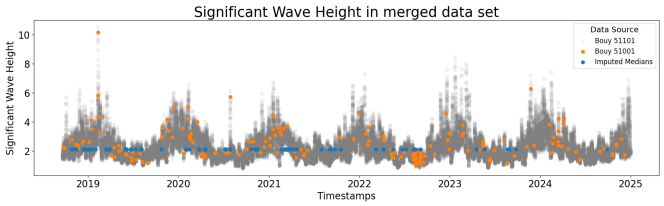

Figure 2: Merged data from buoy 51101 and 51001.

Table 4: Wasserstein Distance between bouys 51101 and 51001

Variable	Wasserstein Distance
Sig. Wave Height	0.0119
Wind Direction	2.5348
Wind Speed	0.0525
Gust Speed	0.0486
Dominant Wave Period	0.0577
Average Wave Period	0.0462
Wave Direction	5.0194
Sea Surface Pressure	0.0884
Air Temperature	0.0593
Water Temperature	0.2097
Air-Water Temp. Diff.	0.0823

4.2 Experimental setup

Prior to model training, feature selection was performed by correlation analysis to identify the set of predictors that associated most strongly with the target variable. Shuffling was disabled to maintain the temporal order in the dataset, and preserve the sequential nature of the data.

For sequence modeling, time series windows of length 24 were constructed to capture the temporal dependencies within the data. The predictive model comprised a single LSTM layer with 50 units followed by a dense output layer. Non linear dependencies were modeled using the tanh activation function.

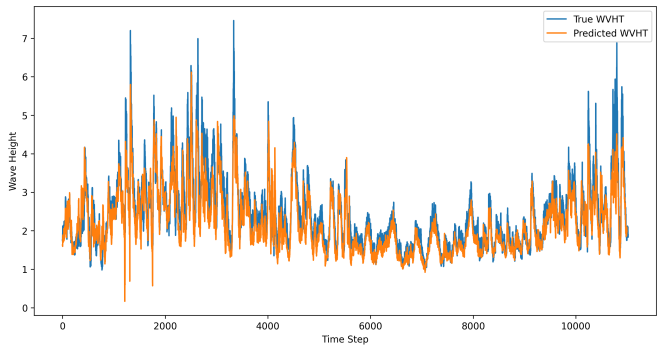

Figure 3: LSTM: Recorded vs Predicted Wave Height

The LSTM model is trained to predict the wave height based on the selected environmental predictors within each time window. To identify the subgroups with distinct model performance, a temporal Exceptional Model Mining approach was used. Due to computational constraints and exploratory scope of this study, the subgroup search was restricted to single and dual predicate subgroups, providing a balance between interpretability and tractability while still identifying key patterns in model performance.

Each input feature was first discretised into bins to convert continuous variables into categorical ranges, to stablize the subgroup search and improve interpretability. It allows patterns to be analyzed over specific interval rather that individual continuous values. For subgroups with single-predicate each binned feature was evaluated independently, and sampled by grouping its quality measure, that accounts for both deviations from the global RMSE and temporal variance.

For dual-predicate subgroups, all unique pairs of features were combined, and samples satisfying both condition simultaneously were evaluated using the same quality measure. To ensure statistical reliability, subgroups with frequency lesser than 100 samples were discarded and only the top ranking subgroups based on |Q| were retained.

4.3 Results

Having carefully described the data and methods used for this study, we now present the results of the analysis. Training the LSTM on the buoy data and subsequently applying the adapted *SCaPE*, TQ-SD system, yielded a few interesting subgroups.

Ranking the subgroups based on the absolute value of their quality measure, table: 5 presents the top five subgroups with a single attribute in their corresponding predicate, whereas table: 6 shows the ten highest ranking subgroups with predicates involving two descriptor.

Table 5: Top 5 Subgroups with a single attribute predicate

Subgroup	Quality measure Q	N_{subgroup}	μ_{RMSE}
WTMP ∈ [22.5, 23.5]	0.154	2221	0.601
AWTD ∈ [-0.0, 0.4]	-0.052	2347	0.389
AWTD ∈ [0.4, 0.7]	0.052	2320	0.388
AWTD ∈ [1.5, 7.1]	-0.045	1998	0.395
SEASON ∈ [-0.0, 1.0]	0.045	10997	0.395

Table 6: Top 10 Subgroups with a dual attribute predicate

Subgroup	Subgroup	Quality measure Q	N_{subgroup}	μ_{RMSE}
ATMP ∈ [26.2, 27.3]	AWTPDF ∈ [1.0, 1.5]	-0.269,	146	0.162
WSPD ∈ [4.5, 6.2]	WTMP ∈ [22.5, 23.5]	0.259,	300	0.717
WTMP ∈ [22.5, 23.5]	AWTPDF ∈ [1.5, 7.1]	0.251,	416	0.749
DPD ∈ [11.43, 13.79]	WTMP ∈ [22.5, 23.5]	0.249,	650	0.705
WSPD ∈ [-0.0, 4.5]	WTMP ∈ [22.5, 23.5]	0.232,	221	0.693
DPD ∈ [5.0, 8.33]	WTMP ∈ [25.4, 26.2]	-0.221,	442	0.215
WSPD ∈ [6.2, 7.4]	ATMP ∈ [26.2, 27.3]	-0.215,	533	0.218
DPD ∈ [5.0, 8.33]	ATMP ∈ [26.2, 27.3]	-0.210,	797	0.223
DPD ∈ [5.0, 8.33]	ATMP ∈ [25.2, 26.2]	-0.207,	696	0.225
ATMP ∈ [25.2, 26.2]	WTMP ∈ [25.4, 26.2]	-0.204,	692	0.236

The single predicate subgroups are not as significant as the dual predicate subgroups, except for one. The predicate "*WTMP ∈ [22.5, 23.5]*", has quality measure of *Q=0.154*, which might not

seem hihg compared to the more complex subgroups. However, it is important to notice that this predicate occurs in all of the four dual predicate subgroups for which the model performs worse than on the entire dataset. This suggest that the corresponding dual predicate amplifies that difference in peformance, meaning "*WTMP ∈ [22.5, 23.5]*" is most likely the dominant subgroup in terms of poor model performance.

The top subgroups in terms of significance also contain those subgroups for which the model is able to provide better predictions compared to the general population. One of the predicates well represented in these subgroups is "*ATMP ∈ [26.2, 27.3]*", furthermore another closely related predicate also occurs frequently in these subgroups namely, *ATMP ∈ [25.2, 26.2]*. The reoccurring presence of these two predicates in exceptionally well performing subgroups indicate that the air temperature is a crucial factor in model performance.

Overall the results show both subgroups for which the model performs poorly or exceptionally well, compared to global model performance. One specific predicate or subgroup regarding the water temperature, "*WTMP ∈ [22.5, 23.5]*" seems to be a particularly dominant factor in poor model performance, with more complex subgroups amplifying the effect. Whereas the air temperature seems to explain in part exceptionally good model performance.

5 Discussion

The global RMSE of the LSTM model (0.452) suggests that the model performs reasonably well; however, the model showed varying results under different environmental conditions. Traditional performance reporting, limited to calculating the global RMSE, would have posed the model as fair, but using EMM showed that this is far from truth. The subgroup mining of the prediction results highlights the fact that the model does not struggle randomly, but due to the presence of certain environmental conditions at the predictor level. It was accurate during calm periods but collapsed during turbulent conditions.

The frequent appearance of WTMP (water temperature) ∈ [22.5, 23.5] in nearly all subgroups that perform poorly strongly suggests that these water temperature bands correspond to a transitional thermal zone in the ocean. At these temperature bands the surface layer is neither fully stratified nor fully mixed with the layers below, so heat exchange between the air and water is unstable. When wind speed (WSPD) interacts with this partially mixed water column, the relationship between wind and wave height stops being predictable. Under normal conditions, stronger wind would lead to larger waves. But under such conditions, the wind energy gets lost in vertical mixing instead of forming consistent waves, resulting in chaotic waves, making it difficult for the LSTM to lock a stable temporal pattern.

The dual-predicate subgroups show that the combined effect of wind temperature (WTMP) with features like wind speed (WSPD) or dominant wave period (DPD) affects the prediction even more. In these cases, wind energy interacts with a thermally unstable water surface, causing inconsistent wave development. So, with the dual-predicate combination of WSPD ∈ [4.5, 6.2] and WTMP ∈ [22.5, 23.5] or DPD ∈ [11.43, 13.79] and WTMP ∈ [22.5, 23.5], noise and non-linearity starts to appear in the ocean and eventually the LSTM cannot find a repeatable temporal pattern between the the wind speed and wave height relationship, so prediction quality degrades.

On the other hand, one of the strongest high-performing subgroups is defined by air temperature (ATMP) ∈ [26.2, 27.3] and air–water temperature difference (AWTPDF) ∈ [1.0, 1.5]. A warm and stable air temperature, in combination with a minute heat difference between air and water, reflects a thermally balanced state. Under such conditions, vertical heat flux becomes predictable, stratification is stable, and the resultant sea state is smoother. The LSTM can finally lock onto consistent temporal patterns, with fewer fluctuations in wave formations, eventually leading to higher accuracy.

So, we can conclude that under stable atmospheric conditions, the model learns temporal patterns and can is trustworthy, but under transient thermal states, where noise and chaotic fluctuations make the ocean unpredictable, the model gets confused. From forecasting perspective, we can conclude that the model is reliable when the conditions are stable but becomes unreliable under dynamic and volatile weather patterns.

6 Conclusion

In our work, we combined LSTM for wave height prediction with a temporal extension of SCaPE to identify the conditions or subgroups of predictors where the model performs exceptionally well or poorly. The results show that the model struggles in thermally unstable ocean states, particularly when the water temperature lies within the range [22.5, 23.5]°C , which is validated by the fact that air–water heat exchange causes chaotic wave formation. In contrast, the model performs significantly better when air temperature is stable and heat difference between air and water is low, resulting in predictable sea states and lower RMSE values. We moved beyond conventional overall accuracy reporting and analyzed the subgroup behavior over time. Overall, this suggests that our approach identified meaningful environmental conditions that affect the prediction accuracy.

6.1 Limitations and Scope for Improvement

We used data from 2 buoys from the Hawaii region, which can lead to some region bias in the model. The model could generalize across different thermal regimes with data from more stations. The fixed window size is a bottleneck for the LSTM's ability to capture long-range temporal dependencies. Ocean dynamics often span longer windows, and models like transformers or hybrid-physics informed architectures capture it more effectively. In order to better access the utility of the Temporal Q-based Subgroup Discovery, it is necessary to validate the models performance on data from other domains, presenting varying predictive difficulties. If successful, these analyses would further prove the robustness and generalizability of the proposed framework. Furthermore, the method could be expanded by considering forecasting models different from an LSTM, further enhancing model utility.

The code used for this work, including the data engineering, LSTM, and Temporal Q-based Subgroup Discovery, can be found in the following GitHub repository: [Purohit 2025]

References

Vitor Cerqueira. 2023. Time Series for Climate Change: Forecasting Large Ocean Waves | Towards Data Science. https://towardsdatascience.com/time-series-for-climate-change-forecasting-large-ocean-waves-78484536be36/

Vitor Cerqueira. 2025. Time Series for Climate Change: Forecasting Large Ocean Waves. https://towardsdatascience.com/time-series-for-climate-change-forecasting-large-ocean-waves-78484536be36/

Manish Dakhore, R.Delecta Jenifer, Mohamed Dawood Shamout, Nilesh Anute, Wanda Gema Prasadio Akbar Hidayat, and Hendy Tannady. 2024. The Application of Time Series Forecasting to Financial Risk Management. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 1–7. doi:10.1109/ICCCNT61001.2024.10725122

Xin Du, Sikun Yang, Wouter Duivesteijn, and Mykola Pechenizkiy. 2025. Conformalized Exceptional Model Mining: Telling Where Your Model Performs (Not) Well. arXiv:2508.15569 [cs.LG] https://arxiv.org/abs/2508.15569

Wouter Duivesteijn and Julia Thaele. 2014. Understanding Where Your Classifier Does (Not) Work – The SCaPE Model Class for EMM. In *2014 IEEE International Conference on Data Mining*. 809–814. doi:10.1109/ICDM.2014.10

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. 2022. Forecast evaluation for Data Scientists: Common Pitfalls and Best Practices. *Data Mining and Knowledge Discovery* 37, 2 (Dec 2022), 788–832. doi:10.1007/s10618-022-00894-5

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (11 1997), 1735–1780. arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf doi:10.1162/neco.1997.9.8.1735

Anthony James Kettle. 2015. A Diagram of wind speed versus air-sea temperature difference to understand the marine atmospheric boundary layer. *Energy Procedia* 76 (2015), 138–147.

Billy Leibundgut. 2021. Wave Height Prediction Using ARIMA, Prophet, and XGBoost. (2021). https://billyleibundgut.medium.com/wave-height-prediction-using-arima-prophet-and-xgboost-b7ddc8a1bdd

Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, and Michael Weyrich. 2021. A survey on long short-term memory networks for time series prediction. *Procedia CIRP* 99 (2021), 650–655. doi:10.1016/j.procir.2021.03.088 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.

Divyansh Purohit. 2025. Divyansh-Purohit/data-mining—wave-height-prediction: Group Project for EMM Group 3. https://github.com/Divyansh-Purohit/Data-Mining---Wave-Height-Prediction

Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. doi:10.1145/3351095.3372827

Ravi Sha and Tapas Guha. 2023. Climate Time Series Prediction with Deep Learning and LSTM. In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*. 1631–1637. doi:10.1109/ICOSEC58147.2023.10276117

Lieke van den Biggelaar, Rianne M. Schouten, Ashley de Bie, R. Arthur Bouwman, and Wouter Duivesteijn. 2025. Characterizing the Risk of Atrial Fibrillation in Cardiac Patients with Exceptional Electrocardiogram Phenotypes. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 4925–4934. doi:10.1145/3711896.3737200

Mihaela van der Schaar and Nick Maxfield. 2023. Time series in Healthcare: Challenges and solutions // Van der Schaar Lab. https://www.vanderschaar-lab.com/time-series-in-healthcare/