Evaluating Teacher Forcing and Curriculum Learning in Recurrent Models for Oceanographic Time Series

Tiago H. Marum[1,2], Ronney Agra[1,2], Marcel Rodrigues de Barros[1], Anna Helena Reali Costa[1], Fábio Gagliardi Cozman[1], Fábio Cunha Lofrano[1], Fernando Akira Kurokawa[1]

[1]Polytechnic School of the University of São Paulo – São Paulo – SP, Brazil

[2]THM Statistical Consultancy – São Paulo – SP, Brazil.

marum@thmestatistica.com,agra@thmestatistica.com,
marcel.barros@usp.br, anna.reali@usp.br, fgcozman@usp.br,
fabio.lofrano@usp.br, fernando.kurokawa@usp.br

Abstract.: *Sea level prediction is vital for port and coastal operations, where short-term forecasting accuracy is critical for navigation and planning. Autoregressive neural models are powerful tools for this task, especially when combined with training strategies like Teacher Forcing and Curriculum Learning. This study analyzes their impact on RNN performance for short-term sea level forecasting using high-frequency data from the Port of Santos, Brazil. We compare standard autoregressive training, Teacher Forcing, and Curriculum Learning across MAE, RMSE, and R^2 metrics, focusing on long prediction windows. Results show that low levels of teacher forcing improve convergence and reduce error over long horizons, highlighting its value for precision and long-term stability.*

1. Introduction

Sea level forecasting plays a critical role in port and coastal operations, especially in areas affected by tides, weather disturbances, and oceanographic dynamics. Accurate short-term predictions — typically within a 24-hour horizon — are essential for vessel navigation, cargo planning, flood prevention, and coastal risk management. These needs have been emphasized in studies focusing on the Santos coastal area, where understanding and adapting to rising sea levels is a priority for local decision-making and resilience planning (MARENGO et al., 2019). Recurrent Neural Networks (RNNs), particularly in autoregressive configurations, have shown promise for such time-dependent tasks due to their ability to model sequential dependencies. However, the training strategy can significantly affect the performance of these models. Techniques such as teacher forcing and curriculum learning have been proposed to improve convergence and accuracy by managing the trade-off between learning from ground truth and relying on self-generated sequences during training (GOODFELLOW; BENGIO; COURVILLE, 2016).

In this study, we compare these training strategies — pure autoregressive learning, teacher forcing, and curriculum learning — on high-frequency sea level data from the Port of Santos, Brazil. Our goal is to evaluate the predictive performance and robustness

of each method in a real-world coastal setting, motivated by the observed difficulty and slow convergence of traditional RNN-based autoregressive models when applied to longer prediction windows (greater than 24 hours) for tidal forecasting. This work builds upon recent literature highlighting the importance of short-term sea level prediction using deep learning approaches in operational contexts (TUR et al., 2021), and aims to assess which training strategies can best mitigate the limitations associated with extended forecasting horizons.

2. Methodology

This study uses the Santos Port SSH dataset, which contains time series data of sea surface height (SSH) recorded at regular 10-minute intervals. This high-frequency dataset enables short-term sea level forecasting, a critical task for port operations, navigation safety, and coastal risk management. The original data was split into training and test sets using a fixed cutoff date (June 1st, 2020), ensuring a chronological separation between the past used for training and the unseen future used for evaluation.

We compare the performance of nine RNN-based models, all sharing the same architecture and data preparation pipeline but differing in their training strategies (Figure 1):

1. **Autoregressive Model (AR)**: This baseline model uses its own past predictions as input during decoding, without any access to ground truth during training.
2. **Teacher Forcing (TF)**: At each decoding step, the model receives the actual target (ground truth) from the previous step instead of its own prediction. This approach, introduced by Williams and Zipser (1989) and later consolidated in Goodfellow et al. (2016), stabilizes and accelerates training, especially in early epochs.
3. **Curriculum Learning (CL)**: A hybrid approach where the probability of using the ground truth decreases over training epochs, allowing a gradual transition from teacher forcing to autoregressive behavior. Both linear and exponential decay strategies were considered.

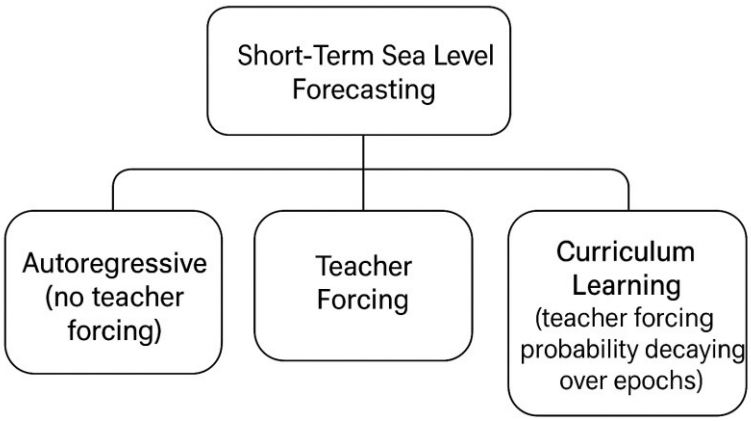

Figure 1. Train learning strategies adopted

2.1 Data Preparation and Parameter Choices

To prepare the data for modeling, a sliding window approach was employed, where each training sample consisted of a fixed-length sequence of past observations used to predict a subsequent sequence of future values.

Specifically, the context window of length p was set to 432 timesteps, corresponding to three days of historical data given the 10-minute sampling interval (6 observations per hour × 24 hours × 3 days). This choice is supported by prior research indicating that longer historical windows capture multiple tidal cycles and meteorological influences, leading to improved forecasting performance in coastal environments. Studies by Vicens-Miquel et al.(2024), Accarino et al. (2021), and Kartal et al. (2024) provide empirical evidence for using context windows ranging from 1 to 3 days in similar settings. Meanwhile, the forecast window, denoted as f, was defined as 144 timesteps, representing a 24-hour prediction horizon. This decision was motivated by the operational relevance of 24-hour forecasts in port and harbor settings, where such predictions are critical for navigation, mooring, flood prevention, and scheduling. Literature in the field confirms the frequent use of 1–24 hour horizons for short-term sea level prediction, as evidenced in works by Tur et al.(2021), Pang et al.(2023), Accarino et al.(2021), and Kartal et al. (2024).

Lastly, the sliding window step size was configured as 72 steps, which translates to 12 hours. This setting controls how much the window shifts forward after each training sample is generated. The selected value aims to reduce data redundancy while preserving sufficient diversity in training sequences, following recommendations from Li et al.(2018), who demonstrated that moderate step sizes improve model generalization and reduce overfitting risks in high-frequency tidal prediction contexts.

All models were trained under the same hyperparameters: 400 training epochs, batch size of 32, hidden layer size of 64, and a learning rate of 0.0001. The loss function used was Mean Squared Error (MSE), and model performance was evaluated using complementary metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

2.2 Experimental Design and Model Variants

In total, nine model configurations were trained and evaluated to investigate the impact of different decoding behaviors during training. All models shared the same architecture, input-output structure, and training hyperparameters. The variations arise solely from the probability of using ground truth values during decoding, and how that probability evolves across epochs. Each model is identified by its configuration file name and summarized in the Table 1.

For the models employing exponential curriculum learning, the probability of using the ground truth at each decoding step was progressively reduced using an exponential decay schedule. The probability p at epoch e was computed as $p(e) = p_0.\gamma^e$, where p_0 is the initial teacher forcing probability (e.g., 1.0 or 0.5), and γ is the decay rate. In this study, we adopted a decay factor of γ=0.97, meaning that the influence of the ground truth gradually diminishes across training epochs, allowing the model to transition smoothly from teacher forcing to autoregressive behavior. This approach is motivated

by the need to avoid abrupt changes in learning dynamics and to encourage stable convergence.

Each model's performance was compared using error-based metrics across the same prediction horizon. This setup allows for a systematic assessment of how varying levels of reliance on ground truth during training affect forecasting accuracy and stability.

Table 1. Model Variants and Decoding Strategies

Model ID	Strategy	Decay Tipe	Considerations
AR	Autoregressive	-	Baseline (no teacher forcing)
TF_p0.01	Teacher Forcing	Fixed	Near-autoregressive
TF_p0.05	Teacher Forcing	Fixed	Low teacher forcing
TF_p0.25	Teacher Forcing	Fixed	Moderate teacher forcing
TF_p0.50	Teacher Forcing	Fixed	Stronger teacher forcing
CL_linear_p1.00	Curriculum Learning	Linear	Linear decay from full TF
CL_linear_p0.50	Curriculum Learning	Linear	Linear decay from moderate TF
CL_ exponential _p1.00	Curriculum Learning	Exponential	Exponential decay from full TF
CL_ exponential _p0.50	Curriculum Learning	Exponential	Exponential decay from 50% TF

2.3 Statistical Analysis and Model Comparison

To evaluate and compare the forecasting performance of the different training strategies, we adopted a combination of quantitative metrics, per-timestep error analysis, and training dynamics monitoring. For each trained model, we computed Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) over the entire test set. These metrics allowed for a standardized comparison of accuracy and generalization capacity.

Additionally, we examined how error evolves across the forecast horizon (144 timesteps, or 24 hours) by plotting the MAE at each prediction step. These plots helped identify oscillatory patterns in forecast error, likely associated with semi-diurnal tidal cycles characteristic of sea surface height data. Models that presented smoother error trajectories and lower variability across the horizon were interpreted as more stable and capable of capturing these inherent dynamics.

We also visualized training and test loss curves over 400 epochs in order to monitor convergence behavior and detect potential signs of overfitting during model training. These plots assist in evaluating whether the models are learning effectively and generalizing appropriately over time.

Finally, all relevant outputs—such as per-timestep MAE values, summary statistics, and training metadata—were stored in structured CSV and JSON files to ensure transparency and reproducibility of the analysis pipeline.

3. Results and Discussion

3.1 Comparison Between AR and Fixed Teacher Forcing Variations

Table 2 presents a comparative evaluation of nine RNN-based forecasting models, each trained with distinct learning strategies. In this first segment, we focus on the comparison between the baseline AR model and the fixed teacher forcing variants (TF_0.25 and TF_0.50).

Although both fixed teacher forcing variants achieved very low training losses — with TF_0.50 reaching as low as 0.014 — this came at the cost of generalization. The fixed teacher forcing models exhibited signs of overfitting, yielding higher test losses, lower MAE scores, and weaker R^2 values when compared to the AR baseline. In fact, the AR model demonstrated more robust test performance despite its relatively higher training loss, highlighting the benefits of relying on a purely autoregressive approach over fixed, high-probability teacher forcing strategies.

Table 2. Comparative Evaluation of RNN-Based Forecasting Models

Model	Train Loss	Test Loss	MAE	RMSE	R^2
AR	0.6690	0.5548	0.6030	0.7436	0.3417
TF_p0.01	**0.5038**	**0.4878**	**0.5544**	**0.6940**	**0.4266**
TF_p0.05	0.3453	0.4981	0.5547	0.7082	0.4028
TF_p0.25	0.0445	0.7175	0.6613	0.8563	0.1271
TF_p0.50	0.0138	0.7050	0.6640	0.8396	0.1608
CL_exponential_p1.00	0.5475	0.4836	0.5526	0.6951	0.4248
CL_exponential_p0.50	**0.5445**	**0.4813**	**0.5520**	**0.6924**	**0.4293**
CL_linear_p1.00	0.6033	0.49172	0.5548	0.7003	0.4162
CL_linear_p0.50	0.6151	0.4843	0.5514	0.6930	0.4283

Figure 2 displays the test loss progression across 400 epochs for the autoregressive (AR) model and the fixed teacher forcing (TF) variants. The results clearly illustrate the impact of teacher forcing rate on generalization. Higher teacher forcing probabilities (TF_0.25 and TF_0.50) lead to greater test loss and more erratic behavior, indicating overfitting due to excessive reliance on ground-truth inputs. In contrast, low teacher forcing probabilities (TF_0.01 and TF_0.05) maintain smoother and more stable test loss curves, suggesting a better balance between ground-truth conditioning and the model's ability to learn long-term dynamics. The AR model, despite yielding relatively higher test loss overall, displays a steady and gradual trend across epochs, indicative of its purely self-regressive nature.

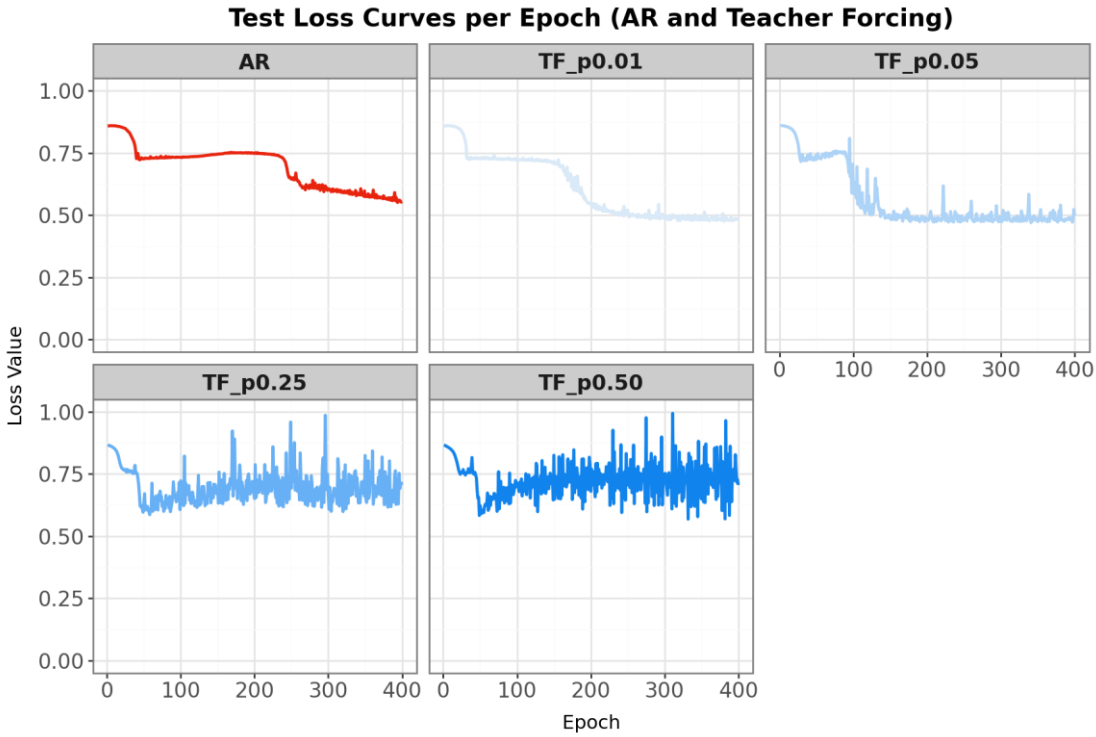

Figure 2. Test Loss Curves Across Epochs for Teacher Forcing Models

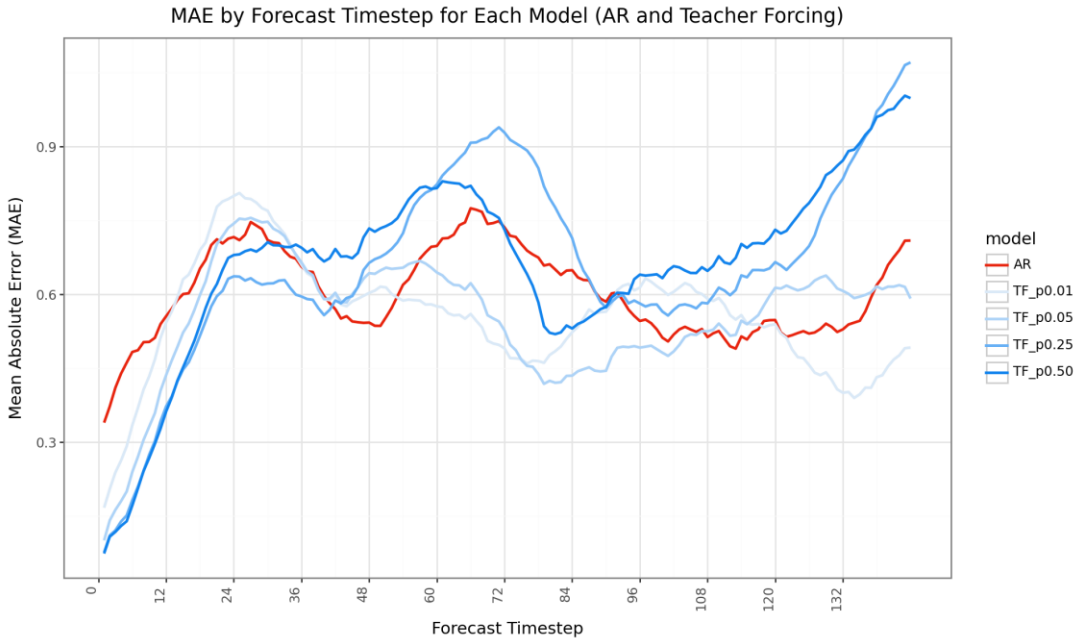

Figure 3. MAE Evolution across the Forecast Horizon for Teacher Forcing Models

Figure 3 illustrates the MAE across the forecast horizon for the AR and fixed teacher forcing models. The results reveal a characteristic oscillatory error pattern aligned with the semi-diurnal tidal dynamics of the target variable. Higher teacher forcing rates (TF_0.25 and TF_0.50) produce MAE curves that rise sharply as the prediction horizon progresses, highlighting weaker generalization beyond the initial timesteps. Meanwhile,

lower teacher forcing rates (TF_0.01 and TF_0.05) maintain flatter and more stable error profiles, allowing the model to adapt better across longer prediction windows. The AR model, despite capturing the periodic nature of the data, yields consistently higher MAE due to its limited forecasting precision.

These error patterns reflect the well-known phenomenon of exposure bias in autoregressive forecasting, where small prediction errors compound over time, creating error envelopes that mimic the original signal's frequency (GOYAL *et al.*, 2016). Techniques such as Professor Forcing have been proposed to mitigate this effect by aligning the dynamics of training and inference, leading to improved long-term forecasting performance.

3.2 Comparison Between AR and Curriculum Learning Variations

Figure 4 illustrates the test loss progression across 400 epochs for the AR and Curriculum Learning (CL) variants. The results reveal a distinct trend: I) the baseline AR model shows a slow and gradual decrease in test loss, maintaining relatively stable performance throughout training; II) The exponential decay CL models (CL_exp_0.50 and CL_exp_1.00) quickly reach lower and more stable test loss values, suggesting an effective balance between teacher forcing and autoregressive learning; III) The linear decay CL approaches (CL_lin_0.50 and CL_lin_1.00) also reduce test loss efficiently, but exhibit higher fluctuations, indicating less stable generalization compared to their exponential decay counterparts.

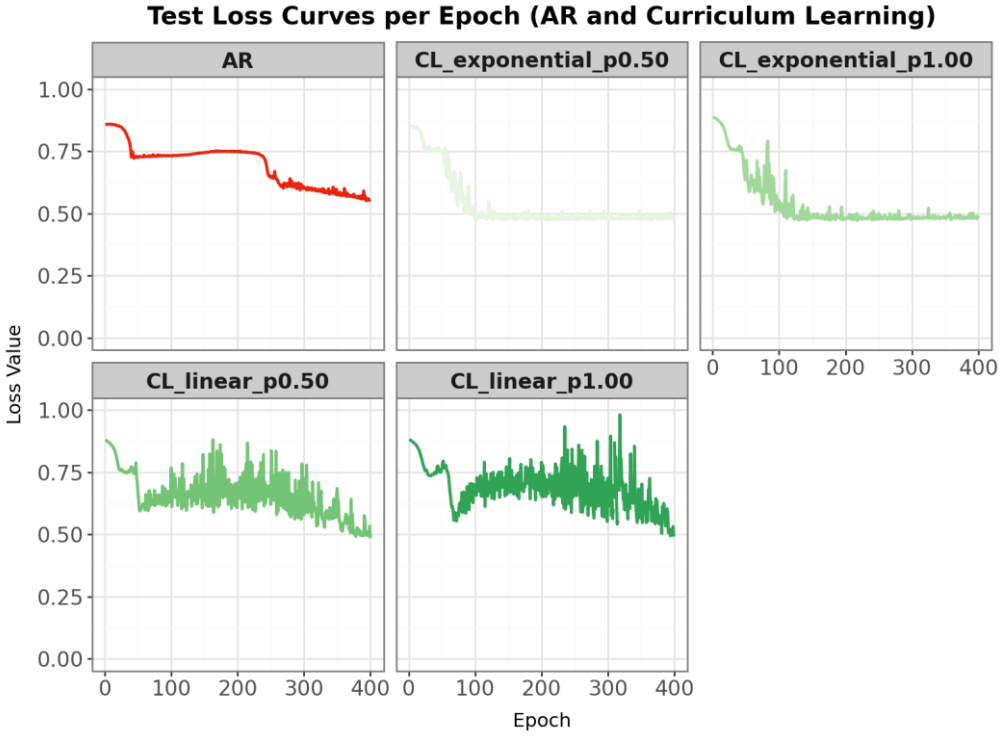

Figure 4. Test Loss Curves Across Epochs for Curriculum Learning Models

These observations align with the findings in Table 2, which confirms that CL strategies achieve lower test loss and improved overall performance compared to AR and fixed teacher forcing approaches.

Figure 5 shows the evolution of the Mean Absolute Error (MAE) across the forecast horizon for the AR and Curriculum Learning (CL) models. The results reveal a characteristic oscillatory error pattern, aligned with the semi-diurnal tidal dynamics of the target variable, with error peaks and troughs corresponding to specific phases of the tidal cycle. The AR model captures this behavior but exhibits larger error peaks and a generally higher MAE across the prediction window, indicating its limitations in long-term forecasting. In contrast, the CL models achieve consistently lower MAE values throughout the horizon, with smoother error profiles and better adaptation to longer prediction intervals. In particular, the CL variants with exponential decay demonstrate the best performance, yielding the lowest and most stable error across timesteps, reinforcing their effectiveness in transitioning from teacher forcing to autoregressive prediction. These findings are consistent with the results presented in Table 2, highlighting the advantage of curriculum learning over the baseline AR approach.

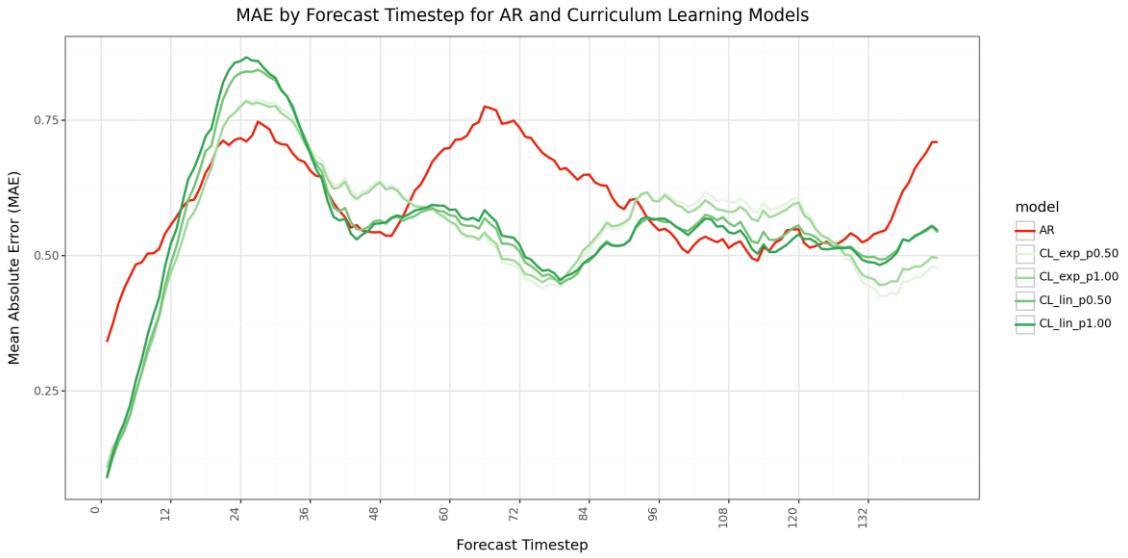

Figure 5. MAE Evolution across the Forecast Horizon for Curriculum Learning Models

3.3 Comparative Evaluation of AR, Teacher Forcing and Curriculum Learning

In this section, we compare the baseline Autoregressive (AR) model with the best-performing variants from the Teacher Forcing (TF_0.01) and Curriculum Learning (CL_exp_0.50) strategies. The goal is to assess their relative performance both in terms of test loss dynamics throughout training and their forecasting accuracy across the prediction horizon. To this end, we present a focused set of figures that overlay the test loss and MAE profiles for these three approaches, allowing a direct visual and quantitative comparison. This provides a clear understanding of the benefits of incorporating structured teacher forcing and curriculum learning into the forecasting framework when compared to the baseline AR approach.

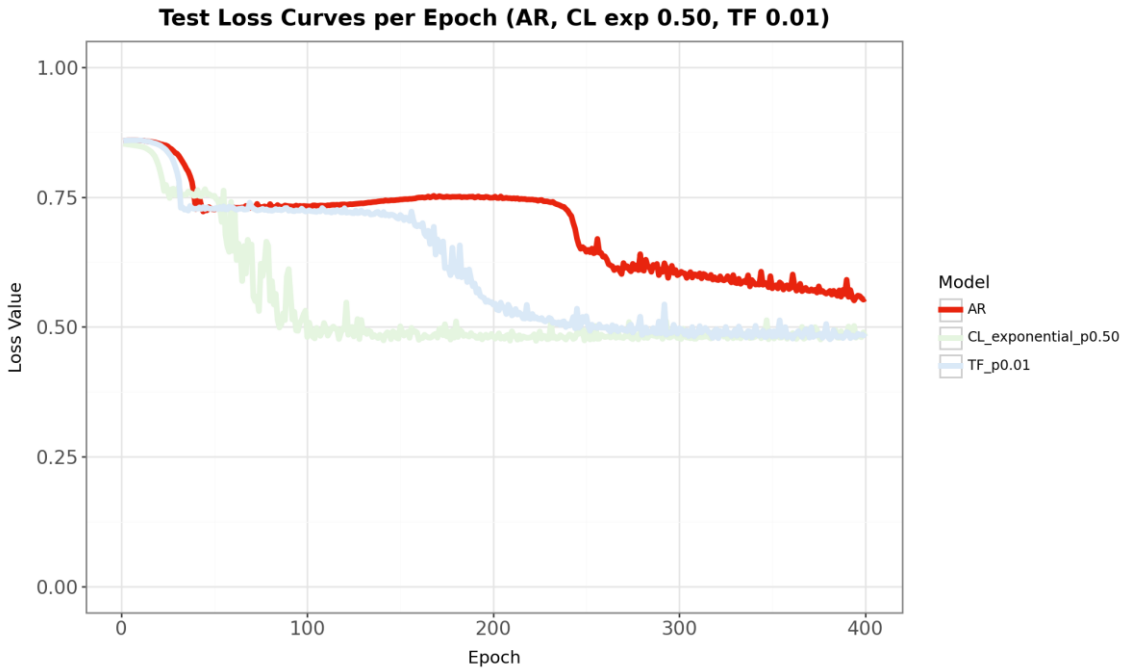

Figure 6. Test Loss Evolution across Training Epochs (AR, CL_exp_0.50 and TF_0.01)

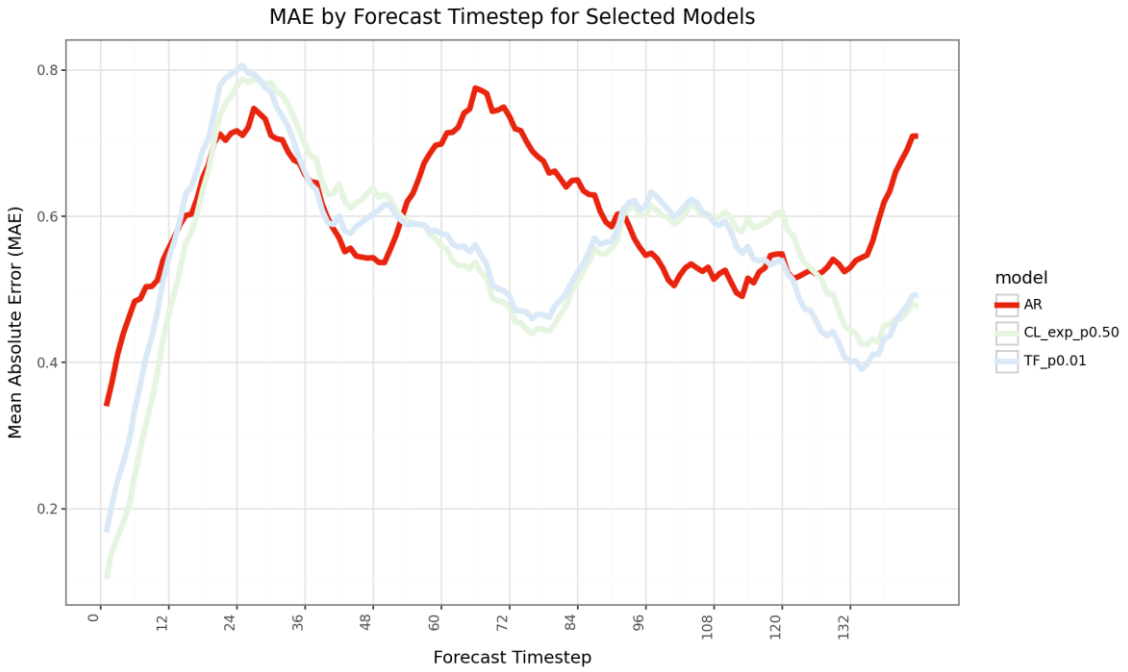

Figure 7. MAE across Forecast Timesteps (AR, CL_exp_0.50 and TF_0.01)

The results in Figure 6 and Figure 7 compare the performance of the baseline AR model, the best Curriculum Learning variant (CL_exp_0.50), and the best Teacher Forcing variant (TF_0.01) across both training dynamics and forecasting accuracy.

Figure 6 shows the test loss evolution across training epochs. The AR model presents a slow, steady decline with higher loss levels throughout training, indicating a more limited ability to fit complex temporal dynamics. In contrast, both CL_exp_0.50 and

TF_0.01 achieve lower test loss levels and more stable training behavior. The CL_exp_0.50 model, in particular, quickly settles into a low loss regime with less fluctuation across epochs, suggesting a better balance between guided learning and autonomous prediction.

Figure 7 analyzes the Mean Absolute Error (MAE) across the forecasting horizon. The AR model captures the semi-diurnal tidal cycle but suffers from larger error peaks and higher overall MAE. Meanwhile, CL_exp_0.50 and TF_0.01 maintain lower and more stable error profiles throughout the prediction window, highlighting their improved ability to generalize across timesteps. Notably, CL_exp_0.50 delivers the best performance, with a consistently dampened error amplitude that captures tidal dynamics more effectively and minimizes long-term prediction drift.

These results clearly demonstrate the benefits of combining autoregressive forecasting with structured learning strategies. In particular, curriculum learning with an exponential decay schedule (CL_exp_0.50) provides the best balance between capturing fine-scale dynamics and maintaining long-term forecasting accuracy, making it the preferred approach for this application.

4. Conclusions

This study conducted a comparative evaluation of recurrent neural network (RNN) models for short-term sea level forecasting, using high-frequency data from the Santos Port. We assessed the impact of three training strategies — autoregressive learning, teacher forcing, and curriculum learning — on model performance, convergence behavior, and forecasting accuracy.

The experimental design adhered to a consistent architecture and parameter setup, allowing a fair and direct comparison across approaches. Through systematic evaluation using MAE, RMSE, and R^2 metrics, as well as error-by-timestep and training loss analyses, we identified distinct learning dynamics associated with each regime. In particular, the curriculum learning approach with an exponential decay schedule (CL_exp_0.50) emerged as the best-performing method, offering a strong balance between convergence stability, generalization, and long-term forecasting precision.

Moreover, visual inspection of error patterns across the prediction horizon revealed characteristic oscillatory behavior aligned with the semi-diurnal tidal cycle, underscoring the role of temporal dynamics in the prediction of sea level fluctuations. The results also highlight the phenomenon of exposure bias in purely autoregressive and fixed teacher forcing setups, and demonstrate how curriculum learning can effectively mitigate its effects, yielding lower error envelopes across forecasting timesteps.

Overall, this work advances the understanding of how structured teacher-forcing strategies can benefit RNN-based forecasting for coastal and oceanographic applications. The framework and results presented here provide a foundation for future research into adaptive scheduling, data augmentation, and hybrid architectures, supporting the development of robust, high-precision prediction tools for port operations and maritime environments.

5. References

Accarino, G., Chiarelli, M., Fiore, S., Federico, I., Causio, S., Coppini, G. and Aloisio, G. (2021) "A multi-model architecture based on Long Short-Term Memory neural networks for multi-step sea level forecasting," Future Generation Computer Systems, vol. 124, pp. 1–9, November. https://linkinghub.elsevier.com/retrieve/pii/S0167739X21001588

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*, MIT Press, USA.

Goyal, A., Lamb, A., Zhang, Y., Zhang, S., Courville, A. and Bengio, Y. (2016) "Professor forcing: A new algorithm for training recurrent networks," In: *Advances in Neural Information Processing Systems*, vol. NIPS, pp. 4608–4616.

Kartal, E. and Altunkaynak, A. (2024) "Empirical-singular-wavelet based machine learning models for sea level forecasting in the Bosphorus Strait: A performance analysis," Ocean Modelling, vol. 188, p. 102324, April. https://linkinghub.elsevier.com/retrieve/pii/S1463500324000118

Li, B., Yin, J., Zhang, A. and Zhang, Z. (2018) "A precise tidal level prediction method using improved Extreme Learning Machine with sliding data window," In: *2018 37th Chinese Control Conference (CCC)*, IEEE, pp. 1787–1792.

Marengo, J. A., Muller-Karger, F., Pelling, M. and Reynolds, C. J. (2019) "The METROPOLE Project – An Integrated Framework to Analyse Local Decision Making and Adaptive Capacity to Large-Scale Environmental Change: Decision Making and Adaptation to Sea Level Rise in Santos, Brazil," In: *Climate Change in Santos Brazil: Projections, Impacts and Adaptation Options*, Cham: Springer International Publishing, pp. 3–15.

Pang, T. Y., Ding, B., Liu, L. and Sergiienko, N. (2023) "Short-Term Sea Surface Elevation Prediction Using Deep Learning Methods," In: *Volume 5: Ocean Engineering*, American Society of Mechanical Engineers.

Tur, R., Tas, E., Haghighi, A. T. and Mehr, A. D. (2021) "Sea level prediction using machine learning," *Water (Switzerland)*, vol. 13, no. 24.

Vicens-Miquel, M., Tissot, P. E. and Medrano, F. A. (2024) "Exploring Deep Learning Methods for Short-Term Tide Gauge Water Level Predictions," *Water*, vol. 16, no. 20, p. 2886, October 11. https://www.mdpi.com/2073-4441/16/20/2886

Williams, R. J. and Zipser, D. (1989) "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280.