

A LightGBM Framework for Operational Tide Forecasting and Predictive Imputation in the Río De La Plata

Diego Silva Piedra[1(✉)], Mónica Fossati[2], and Pablo Ezzatti[1]

[1] Institute of Computer Science, Facultad de Ingeniería, Universidad de la República, 11300 Montevideo, Uruguay
{diegos,pezzatti}@fing.edu.uy

[2] Institute of Fluid Mechanics and Environmental Engineering, Facultad de Ingeniería, Universidad de la República, 11300 Montevideo, Uruguay
mfossati@fing.edu.uy

Abstract. Accurate tide level records are essential for hydrodynamic modeling, coastal management, and the development of operational forecasting systems. This paper presents and evaluates a LightGBM-based framework for operational tide level forecasting and predictive imputation in the Río de la Plata (RdP) estuary, one of the world's most extensive estuaries situated between Argentina and Uruguay, whose complex hydrodynamics, driven by the interplay of meteorological and astronomical forcing, pose significant challenges for predictive modeling. Two forecasting model variants are proposed and evaluated across multiple tide gauge stations and forecast horizons up to 72 h: an autoregressive baseline trained exclusively on historical tide gauge records, and a physics-informed variant that incorporates WRF meteorological forecasts and PronUy_RPFM numerical sea level predictions as additional input features. Both models are assessed in terms of accuracy and feature relevance, in the context of the PronUy_RPFM operational forecasting system developed at the Facultad de Ingeniería, Universidad de la República. Additionally, a bidirectional predictive imputation scheme is developed, applying the pre-trained forecasting model directly to the reconstruction of missing tide gauge segments by combining forward and backward predictions through distance-weighted blending.

Keywords: tide level forecasting · LightGBM · estuarine hydrodynamics · time series prediction · Río de la Plata · predictive imputation

1 Introduction

Operational coastal forecasting systems depend critically on the availability of continuous, high-quality tide gauge records. In estuarine environments subject to strong meteorological forcing such as the RdP, one of the world's largest estuaries, situated between Argentina and Uruguay, accurate water level prediction is

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 532–549, 2027.
https://doi.org/10.1007/978-3-032-30488-9_33

essential for flood risk management, navigation safety, and the sustainable planning of coastal infrastructure. The RdP, especially along the Uruguayan coast, exhibits a predominantly meteorological tidal regime driven by the interaction of wind fields, atmospheric pressure gradients, river discharge, and ocean tides propagating from the South Atlantic shelf, making it a particularly challenging environment for predictive modeling [20].

PronUy_RPFM is the operational tide level forecasting system developed jointly by the Institute of Fluid Mechanics and Environmental Engineering (IMFIA) and the Institute of Computer Science (INCO) at the Facultad de Ingeniería, Universidad de la República (UdelaR), Uruguay. The system provides publicly available, hourly sea level forecasts for the next 72 h across the RdP and its maritime front, historically grounded in the TELEMAC2D finite element numerical model [7,21]. While physics-based numerical models offer physically consistent predictions [17], their accuracy is constrained by the quality of atmospheric forcing inputs and the inherent complexity of representing fine-scale estuarine dynamics. Over the years of operation, PronUy_RPFM has accumulated a substantial observational database, and that growing record opens the door to data-driven approaches that can learn directly from what the estuary has done, rather than from equations describing what it should do.

Data-driven forecasting is a natural complement to physics-based modelling in the RdP, where the meteorological component of sea level variability is difficult to parameterise but leaves a strong signature in tide gauge records that can be learned from historical data [5,6,24]. Gradient boosting models, and LightGBM in particular [10,13,25], handle lagged observations and external covariates natively and have shown competitive skill in estuarine water level prediction. Their histogram-based split-finding and leaf-wise growth strategy substantially reduce training time on decade-long hourly series while the ensemble structure provides a natural framework for incorporating meteorological forcing alongside autoregressive lags. While recurrent architectures such as LSTMs and Transformer-based models have attracted considerable attention for time series forecasting, their advantage over simpler approaches is not as clear as one might assume. Zeng et al. [27] demonstrated that a one-layer linear model outperforms several sophisticated Transformer-based solutions on standard long-term forecasting benchmarks, attributing this to the permutation-invariant self-attention mechanism's inherent difficulty in preserving temporal order. Likewise, the debate between neural networks and GBDTs on tabular data has been shown to be overemphasized: McElfresh et al. [15], in a large tabular data benchmark comparing 19 algorithms across 176 datasets, found that for a surprisingly high number of datasets either the performance difference is negligible, or light hyperparameter tuning on a GBDT matters more than the choice of algorithm family. The same study found that GBDTs are consistently superior on datasets with irregular or heavy-tailed feature distributions, a characteristic often present in estuarine water level records driven by intermittent storm surge events. This finding is consistent with Grinsztajn et al. [11], who showed that tree-based models remain state-of-the-art on medium-sized tabular datasets, as neural networks

tend to struggle with irregular target functions and uninformative features. From a practical forecasting perspective, Bojer and Meldgaard [2] found that gradient boosting methods with engineered temporal features consistently ranked at the top of large-scale forecasting competitions, with their strength particularly pronounced when informative exogenous variables, such as the meteorological forcing available in the RdP context, are available. LightGBM's native handling of heterogeneous features and its substantially lower computational requirements make it a well-motivated choice for operational tide level forecasting in the RdP.

Beyond forecasting, the integrity of historical tide gauge records poses an equally important operational challenge. Sensor failures, transmission errors, and maintenance interruptions give rise to data gaps that undermine the training of data-driven models and the calibration of numerical ones. Conventional gap-filling strategies like linear interpolation, harmonic reconstruction, or simple climatological substitution ignore the temporal structure of the series and the physical forcing that drives water level variability. Given the strong autoregressive signal inherent to tide gauge records and the availability of operational wind and sea level forecasts for RdP, a predictive imputation approach based on LightGBM and trained to reconstruct missing segments represents a natural extension of the forecasting framework developed here. Predicting bidirectionally (*i.e.*, from before and after the gap) and blending the two through distance weighting helps reduce the drift that accumulates in longer gaps.

This paper presents the development and evaluation of two LightGBM-based forecasting models integrated into the PronUy_RPFM system, together with a LightGBM-based predictive imputation scheme for tide gauge records. The main contributions of this work are as follows:

- We develop an autoregressive LightGBM baseline trained exclusively on historical tide gauge records that matches competitive forecasting performance across multiple RdP stations and horizons.
- We propose a hybrid physics-informed LightGBM framework that takes PronUy_RPFM numerical sea level predictions and WRF meteorological forecasts as input features, improving forecasting performance over the historical data-only baseline.
- We introduce a bidirectional predictive imputation scheme that reconstructs missing tide gauge records by combining forward and backward LightGBM predictions with distance-weighted blending.

The remainder of the paper is organized as follows. Section 2 describes the study area and the datasets employed. Section 3 details the LightGBM forecasting framework, including the recursive prediction strategy and feature engineering. Section 4 presents the predictive imputation approach. Section 5 reports and discusses results for both forecasting and imputation tasks. Section 6 summarizes the conclusions and outlines directions for future work.

2 Study Area and Data

2.1 The Río De La Plata Estuary and Tide Gauge Network

The RdP is a large estuarine system on the southeastern coast of South America, forming the boundary between Argentina and Uruguay (Fig. 1). Fed primarily by the Paraná and Uruguay rivers, it receives an average freshwater discharge of approximately 26,000 m^3/s. Its geometry spans roughly 290 km in length and up to 220 km in width at its mouth, transitioning from a broad shallow inner zone, where depths generally remain below 10 m west of the Montevideo–Punta Piedras cross-section, to a deeper outer region with depths exceeding 20 m [12]. This morphological configuration, combined with its large surface area, makes the estuary highly sensitive to surface wind stress and atmospheric pressure perturbations propagating from the South Atlantic shelf.



Fig. 1. Geographic setting of the study area. (a) Location of the Río de la Plata on the South American continent. (b) Location of the tide gauge stations used in this study – Montevideo (MVD), Buenos Aires (BsAs), and Isla de Flores (IdF).

The RdP operates under a microtidal regime in which the meteorological component of sea level variability dominates along the Uruguayan coast, while astronomical tidal forcing is more prominent along the Argentine coast [8,19]. Meteorological sea level anomalies arise from storm surges generated over the continental shelf and subsequently modified by estuarine geometry, wind friction, and bathymetric effects as they propagate inland. Sustained southeasterly winds are the primary driver of positive water level anomalies at the Uruguayan coast, whereas west-northwesterly winds are systematically associated with below-average levels [19].

Tide gauge records used in this study were obtained at three monitoring stations: Montevideo (MVD), located on the northern coast of the inner estuary; Buenos Aires (BsAs), on the southern Argentine coast; and Isla de Flores

(IdF), near the outer estuary mouth (Fig. 1). Records at MVD and IdF are maintained by the National Port Administration of Uruguay (Administración Nacional de Puertos, ANP), while the BsAs record is provided by the Argentine Naval Hydrography Service (Servicio de Hidrografía Naval, SHN). Sub-hourly raw measurements were aggregated to hourly averages; isolated outliers were removed and short discontinuities filled by linear interpolation. The resulting time series span 2017–2025, with varying coverage and continuity across stations (Fig. 2). The most recent year of available data at each location was withheld as an independent test set: 2025 for MVD and BsAs, and 2024 for IdF, with all preceding data used for training and validation.

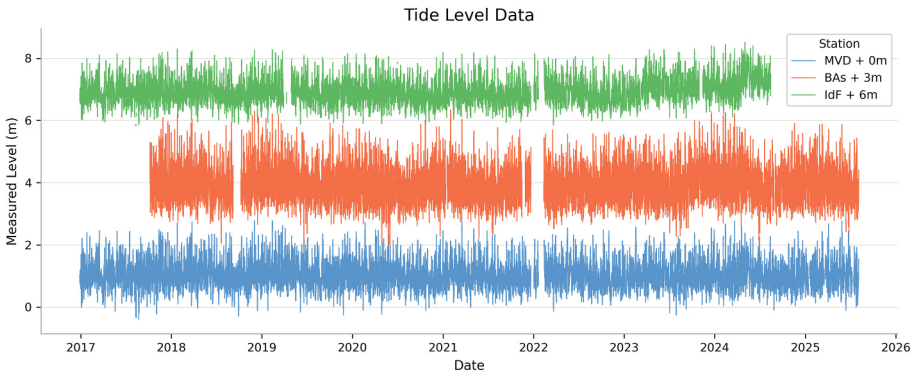

Fig. 2. Hourly tide level records at MVD (blue), BsAs (orange), and Isla de Flores (green) for the study period. An arbitrary vertical offset is applied at each station for visual clarity. Gaps in the series correspond to periods of missing data arising from sensor failures or data transmission errors. (Color figure online)

2.2 Meteorological and Numerical Forcing Data

Meteorological forcing data used as inputs to the PronUy_LightGBM2 model variant were obtained from numerical weather forecasts produced by IMFIA researchers at the Facultad de Ingeniería using the Advanced Research WRF model [22,23]. The WRF system provides hourly wind speed (m/s) and wind direction (degrees) at a grid point collocated with the Montevideo port tide gauge, with temporal coverage spanning the full tide level measurement period. As described in Sect. 3, wind speed and direction are subsequently decomposed into directional intensity projections along eight cardinal and intercardinal directions as part of the feature engineering process.

In addition, operational sea level forecasts from the PronUy_RPFM system [7] were incorporated as a physics-based predictor for the MVD station. These forecasts are generated by the TELEMAC2D finite element hydrodynamic model and provide a physically consistent signal encoding the large-scale tidal

and meteorological variability of the estuary, complementing the statistical patterns learned by the LightGBM model from historical tide gauge observations.

3 LightGBM-Based Forecasting Framework

3.1 Model Architecture and Recursive Multi-step Prediction

LightGBM [13] is an open-source gradient boosting framework developed by Microsoft that constructs a predictive model as an additive ensemble of decision trees. Each tree comprises a root node receiving the input data, internal nodes encoding learned splitting rules, and leaf nodes producing scalar predictions [26]. Building on the gradient boosting paradigm introduced by Friedman [9], each successive tree in the ensemble is fitted to the negative gradient of the loss with respect to the current predictions, progressively reducing the residual error. The histogram-based split finding, leaf-wise tree growth, and gradient-based one-sided sampling that distinguish LightGBM from conventional gradient boosting implementations translate into substantial gains in training throughput and memory efficiency, a practical advantage when retraining independently across stations and forecast horizons on the observational record accumulated by PronUy_RPFM.

Formally, given an input feature vector $\mathbf{x}$, the ensemble output $\hat{Y}_I$ after I boosting rounds is:

$$\hat{Y}_I = \sum_{i=1}^{I} T_i(\mathbf{x},\, \theta_i)\,, \tag{1}$$

where T_i denotes the i-th regression tree and θ_i its associated leaf weights. At each round, the new tree is obtained by minimizing an objective that combines a task loss $\mathcal{L}$ and a complexity regularizer Ω_i:

$$\mathcal{O}bj_i = \mathcal{L}(y,\, \hat{y}_i) + \Omega_i, \tag{2}$$

where y is the target observation and $\hat{y}_i$ the prediction at round i. The regularization term Ω_i penalizes the number of leaves and the magnitude of leaf weights, controlling model complexity and reducing the risk of overfitting at longer forecast horizons. Figure 3 illustrates this additive construction: each successive tree T_i is fitted to the residual of the previous ensemble and their outputs are summed to produce the final prediction.

Because LightGBM produces a single scalar output, a recursive multi-step prediction strategy is adopted to generate L-hour forecasts (Fig. 4). At step t, the model takes a feature vector $\mathbf{f}(t)$, constructed from the window of W most recent water level observations and auxiliary inputs, and outputs the predicted level $\hat{y}(t{+}1)$. This prediction is appended to the input window, the feature vector is recomputed for step $t{+}1$, and the process is iterated L times without retraining any additional model. The dependency of each step on its predecessors mirrors the operational setting in which only past and forecast-horizon inputs are available.

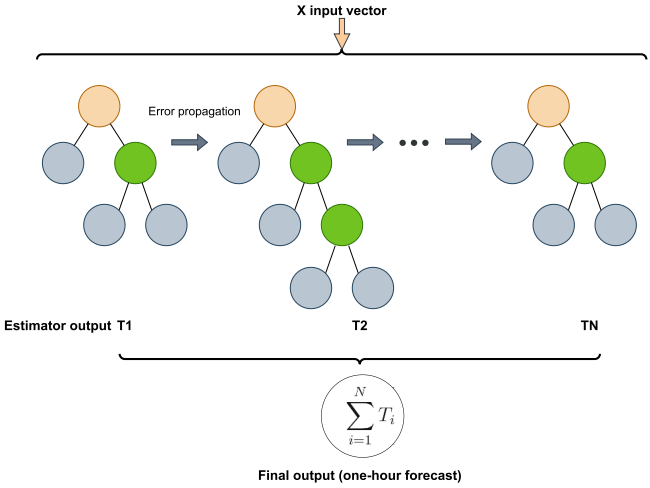

Fig. 3. Gradient boosting ensemble: each tree T_i receives the same input vector $\mathbf{x}$ and is trained on the residual error of its predecessors. The final prediction is the sum $\sum_{i=1}^{N} T_i(\mathbf{x})$.

3.2 Feature Engineering and Model Variants

The feature vector $\mathbf{f}(t)$ is constructed from three categories of inputs, designed to capture the temporal structure and physical drivers of tide level variability in the RdP.

Temporal Features. Calendar variables (hour of day, day of week, day of year, month, quarter, and week of year) are extracted from the timestamp index. To avoid artificial discontinuities at calendar boundaries (e.g., hour 23 followed by hour 0), each variable is encoded as a pair of sine and cosine projections, preserving the cyclic nature of tidal and meteorological periodicities.

Lagged and Rolling Features. A sliding window of W consecutive hourly observations of the target station provides the core autoregressive signal. Complementing these point-wise lags, rolling mean and rolling standard deviation computed over multiple window lengths summarize recent trend and variability, respectively, supplying the model with a compact representation of recent estuarine dynamics.

Meteorological and Numerical Forcing Features. Hourly wind speed and direction forecasts at the Montevideo port area are decomposed into eight directional intensity components aligned with the cardinal and intercardinal directions (N, NE, E, SE, S, SW, W, NW), enabling the model to associate specific wind regimes with the subsequent water level response at each gauge. Operational sea level forecasts from PronUy_RPFM serve as an additional physics-based predictor, providing a signal that encodes large-scale tidal forcing beyond the autoregressive window.

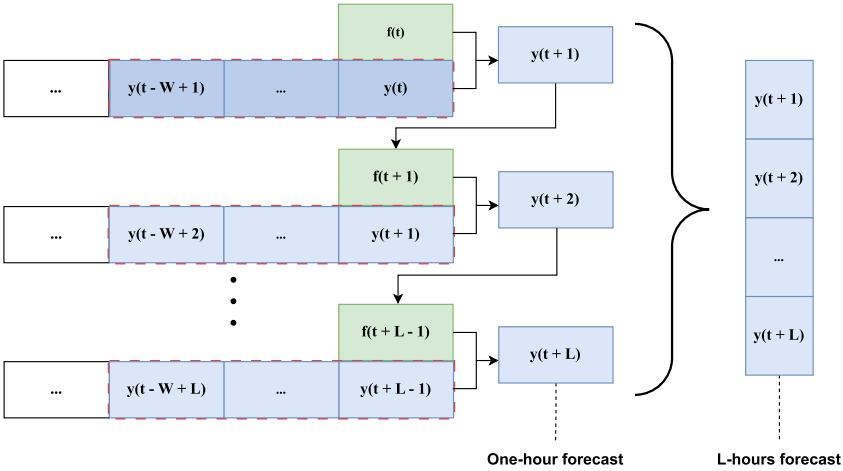

Fig. 4. Input–output diagram of the PronUy_LightGBM recursive forecasting framework. At each step the model predicts $\hat{y}(t+1)$ from the current feature vector $\mathbf{f}(t)$ and a window of W observed (or previously predicted) levels; the prediction is fed back as input for the subsequent step, yielding an L-step forecast through iteration.

These configurations define the two model variants evaluated in this study. PronUy_LightGBM1 uses only temporal and lagged features, relying exclusively on the tide gauge history of the target station. PronUy_LightGBM2 extends this baseline with the meteorological and numerical forcing features described above, currently implemented at MVD where both wind forecast and PronUy_RPFM data are available. Both variants employ the same recursive prediction strategy and are trained independently for each station.

To assess the individual and combined contributions of each forcing source, PronUy_LightGBM2 is evaluated under three configurations: (i) Wind Only, incorporating WRF directional wind features but not PronUy_RPFM sea level forecasts; (ii) Tide Only, incorporating PronUy_RPFM sea level forecasts but not wind features; and (iii) Wind + Tide, combining both sources. Performance is reported as RMSE improvement relative to PronUy_LightGBM1 at MVD.

3.3 Training, Hyperparameter Optimization, and Evaluation

PronUy_LightGBM1 was trained on hourly tide level records from MVD, BsAs, and IdF covering the 2017–2025 period. PronUy_LightGBM2 training was carried out at MVD, where the full set of forcing inputs is available. The window size W and all LightGBM hyperparameters were jointly optimized using the Optuna framework [1], which implements a Tree-structured Parzen Estimator (TPE) Bayesian search over the hyperparameter space. The search covered tree structure parameters (maximum depth, number of leaves), learning dynamics (learning rate, number of estimators), regularization weights, and subsampling

ratios. Each trial configuration was evaluated using recursive prediction on a validation set obtained by a temporal 80/20 split of the training data, with early stopping applied (patience of 100 iterations) to prevent overfitting. The primary optimization target was RMSE on the validation set.

Following optimization, the final model was retrained on the complete training set with the best configuration identified by Optuna and applied to the held-out test sets (2025 for MVD and BsAs; 2024 for IdF). For each day in the test period, a fresh L-step recursive forecast was initialized from the last observed window, and the RMSE between predicted and observed levels was recorded across horizons $L \in \{6, 12, 24, 48, 72\}$ hours. Feature importance was assessed using LightGBM's gain-based metric, which quantifies the average reduction in the objective function attributable to each input feature across all splits in the ensemble.

The computations for the optimization were performed on the high-performance computing infrastructure of the National Supercomputing Center (Cluster-UY) [16].

4 Predictive Imputation of Missing Tide Gauge Records

4.1 Problem Formulation and Approach

Gaps in tide gauge records arise regularly from sensor failures, power outages, and data transmission errors, and are particularly frequent during extreme meteorological surge events, conditions that are most relevant for coastal risk assessment. Conventional gap-filling strategies such as linear interpolation or harmonic reconstruction do not exploit the physical forcing that drives water level variability, while purely spatial approaches based on neighboring stations require simultaneous observations at multiple gauges, which are not always available. Gradient boosting methods have been shown to be effective general-purpose gap-filling tools for environmental time series, exploiting lagged and auxiliary features to reconstruct missing segments across a wide range of conditions [14]. The forecasting model developed in Sect. 3, on the other hand, is additionally informed by meteorological forcing and physics-based model output, a richer feature set that is particularly valuable during surge events, when autoregressive information alone is insufficient. This model transfers directly to the imputation setting, where the gap can be treated as an unknown future segment with respect to the pre-gap context.

A key distinction separates imputation from operational forecasting: when filling a historical gap, real observations from *after* the gap are available. This additional information can be exploited through a strategy inspired by BRITS [3], a bidirectional recurrent imputation model that anchors predictions to both temporal boundaries of the gap to reduce drift accumulation. In this work the same principle is applied to a gradient boosting model: two independent recursive predictions are computed and subsequently merged. The *forward* prediction propagates from the pre-gap context toward the end of the gap in the usual manner. The *backward* prediction is obtained by reversing the time

axis, initializing the recursive model from the post-gap context, and predicting in the direction of the gap start; the resulting sequence is then reversed to restore chronological order. Both predictions are combined via a distance-weighted blend:

$$\hat{y}_{imp}(t_i) = w_{fwd}(i)\,\hat{y}_{fwd}(t_i) + w_{bwd}(i)\,\hat{y}_{bwd}(t_i), \tag{3}$$

where $i = 0, \ldots, L-1$ indexes positions within a gap of length L, and the weights vary linearly:

$$w_{fwd}(i) = 1 - \frac{i}{L-1}, \qquad w_{bwd}(i) = \frac{i}{L-1}. \tag{4}$$

At the gap onset (i=0) full weight is assigned to the forward prediction; at the gap end (i=L−1) full weight goes to the backward prediction.

Unlike BRITS, which requires training a separate model on gap-annotated data, the proposed approach transfers the pre-trained PronUy_LightGBM2 Wind+Tide forecasting model directly to the imputation task without modification or retraining. This design avoids the need for supervised gap annotations and preserves the full meteorological forcing signal embedded during forecasting training. The linear distance-weighted blending in Eqs. (3)–(4) is intentionally parameter-free, requiring no additional training and carrying a natural physical interpretation, as predictions initialized closer to an observed boundary accumulate less recursive drift and are therefore assigned higher confidence.

4.2 Imputation Design and Evaluation

The imputation framework is evaluated following a protocol inspired by TSI-Bench [4], a standardized benchmark for time series imputation. Artificial gaps of lengths $L \in \{6, 12, 24, 48, 72\}$ hours are introduced into the held-out test set at MVD, and the model is tasked with reconstructing the missing segment. At least 30 independent gap instances are sampled per length to ensure statistically stable estimates, in accordance with TSI-Bench guidelines.

Two gap-generation mechanisms are considered, following the missing data taxonomy of Rubin [18]:

MCAR (Missing Completely At Random), in which gap start positions are drawn uniformly at random from the test set. This scenario represents an unbiased baseline and corresponds to random sensor dropouts.

MNAR (Missing Not At Random), in which start positions are sampled with probability proportional to the maximum water level within the candidate gap, biasing the selection toward high-level events. This scenario is more realistic for tide gauges, where sensor overload and power failures are more likely during extreme surge condition. In fact, some of the tide gauges—not included in this study—surveyed by Pronuy_RPFM exhibit this type of malfunction.

Performance is assessed using RMSE and MAE averaged over all sampled gaps at each length. Representative 24-hour reconstructions under both MCAR and MNAR conditions are included in Sect. 5 to provide visual insight into the bidirectional strategy.

5 Results and Discussion

5.1 Forecasting Performance and Feature Importance

PronUy_LightGBM1. For each day in the test period, a fresh L-step recursive forecast was initialized from the last observed window, and the RMSE between predicted and observed levels was computed. Results are reported as the mean RMSE and the 1st–99th percentile band across all daily initializations, for each prediction horizon $L \in \{6, 12, 24, 48, 72\}$ hours.

The relative performance across stations, visible in Fig. 5, persists across all horizons and reflects the contrasting hydrodynamic regimes along the estuary. At IdF, near the outer mouth of the estuary, the model achieves the lowest errors across all horizons, with 72-h RMSE remaining below 0.170 m. Buenos Aires shows the highest errors across all horizons (0.263 m at 6 h, 0.310 m at 48 h), consistent with the stronger astronomical tidal component at that station; normalised errors preserve the same ranking, suggesting that the relative difficulty of each station is not an artifact of differing tidal range but reflects a genuine difference in predictability under the autoregressive formulation.

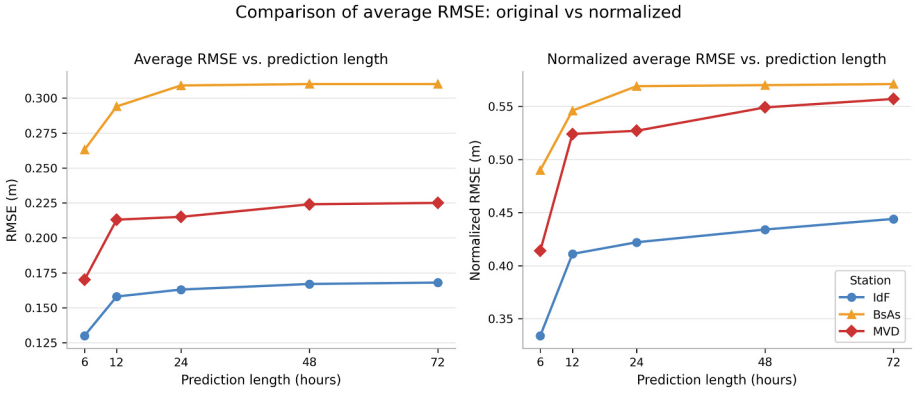

Fig. 5. PronUy_LightGBM1: mean RMSE as a function of forecast horizon L for Montevideo (MVD), Buenos Aires (BsAs), and Isla de Flores (IdF).

Figure 6 displays the five most influential features for each station, ranked by LightGBM's gain-based importance metric. The single-step lag $y(t)$ – the water level recorded at the immediately preceding hour – is the dominant predictor at all three stations, confirming the strong short-range autocorrelation of the tidal signal in the RdP. Beyond this leading lag, the importance patterns diverge across locations. At IdF and MVD, lags at 24, 27, and 48 h also rank highly, reflecting the role of tidal periodicities and estuarine propagation timescales in shaping the predictive signal. Buenos Aires presents a distinctly different profile: rolling standard deviation features appear prominently among the top five

predictors, indicating that local water level *variability* carries stronger predictive weight at this station. This can also be a result of the already mentioned dominance of astronomical tidal forcing on the Argentine coast.

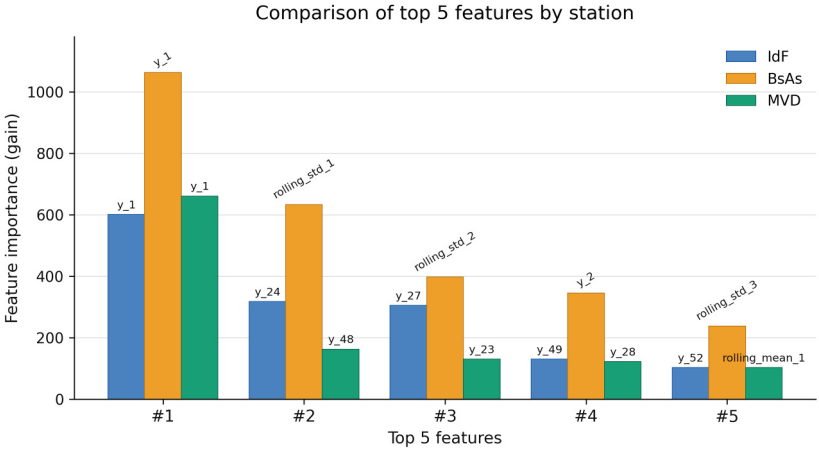

Fig. 6. PronUy_LightGBM1: five most influential input features ranked by LightGBM gain-based importance for each tide gauge station.

PronUy_LightGBM2. Table 1 reports the RMSE percentage reduction over PronUy_LightGBM1 at Montevideo for three forcing configurations: wind forecasts alone, PronUy_RPFM sea level forecasts alone, and both combined. All three configurations reduce RMSE relative to the autoregressive baseline at every horizon, and the magnitude of improvement grows consistently with L. The combined Wind + Tide configuration achieves the largest gains across all horizons (15.4% at 6 h, rising to 40.9% at 72 h), followed by the tide-only variant (10.2%–31.6%), with the wind-only configuration yielding the smallest improvements (10.5%–24.7%). PronUy_RPFM sea level forecasts capture tidal and storm-surge dynamics through the TELEMAC2D hydrodynamic model, encoding estuarine geometry and bathymetric propagation effects that directional wind components alone cannot represent. Wind forecasts contribute an independent component reflecting local surface stress that the hydrodynamic model, operating at coarser resolution, partially misses. The additive gains of the combined configuration confirm that the two sources address distinct components of the forecast error.

The rate at which RMSE grows with horizon also differs between the two models. For the combined PronUy_LightGBM2 configuration, the RMSE increment from 24 h to 48 h is approximately 0.025 m, dropping to 0.010 m from 48 h to 72 h. For PronUy_LightGBM1, the 48–72 h increment remains 0.021 m.

Table 1. RMSE improvement (%) of PronUy_LightGBM2 over PronUy_LightGBM1 at Montevideo for each forecast horizon and forcing configuration.

Horizon (h)	Wind Only	Tide Only	Wind + Tide
6	10.5	10.2	15.4
12	13.2	14.3	21.4
24	18.8	24.8	33.8
48	24.1	31.1	40.2
72	24.7	31.6	40.9

The deceleration in error growth for PronUy_LightGBM2 reflects the physics-based forcing sustaining predictive skill at long horizons beyond the range of the autoregressive window.

5.2 Imputation Performance

Table 2 summarises the imputation results for the Wind + Tide (bidirectional blend) variant across all gap lengths and both missing-data mechanisms. Figure 7 plots RMSE as a function of gap length.

Table 2. Imputation results for the Wind + Tide (bidirectional) variant at MVD.

Gap (h)	MCAR		MNAR	
	RMSE (m)	MAE (m)	RMSE (m)	MAE (m)
6	0.082	0.071	0.073	0.064
12	0.072	0.060	0.096	0.081
24	0.099	0.077	0.117	0.096
48	0.165	0.133	0.140	0.114
72	0.161	0.129	0.159	0.127

Under MCAR, the bidirectional imputer achieves RMSE of 8.2 cm for 6-hour gaps, rising to 16.5 cm at 48 h before a slight decrease to 16.1 cm at 72 h. Errors below 10 cm for gaps up to 24 h represent at most a 2× degradation relative to the typical 5 cm measurement uncertainty of RdP tide gauges, a tolerance compatible with retrospective data reconstruction and model reanalysis applications.

Under MNAR, where gap positions are biased toward high-water events, errors grow monotonically from 7.3 cm at 6 h to 15.9 cm at 72 h. For medium gaps (12–24 h), MNAR errors exceed MCAR errors, consistent with the model's limited ability to extrapolate surge peaks from pre-gap autoregressive context. At 72 h the two curves converge (MNAR: 15.9 cm vs MCAR: 16.1 cm), suggesting that for very long gaps anchor proximity (*i.e.*, the distance from the imputed point to the nearest known boundary) becomes the dominant source of error, independent of whether the gap coincides with an extreme event.

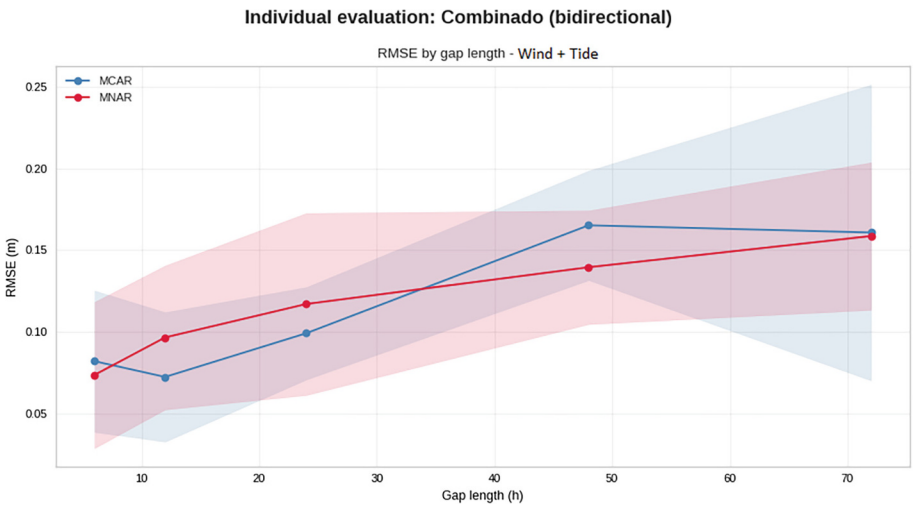

Fig. 7. RMSE as a function of gap length for the imputer under MCAR and MNAR conditions. Shaded bands indicate the interquartile range across sampled gap instances.

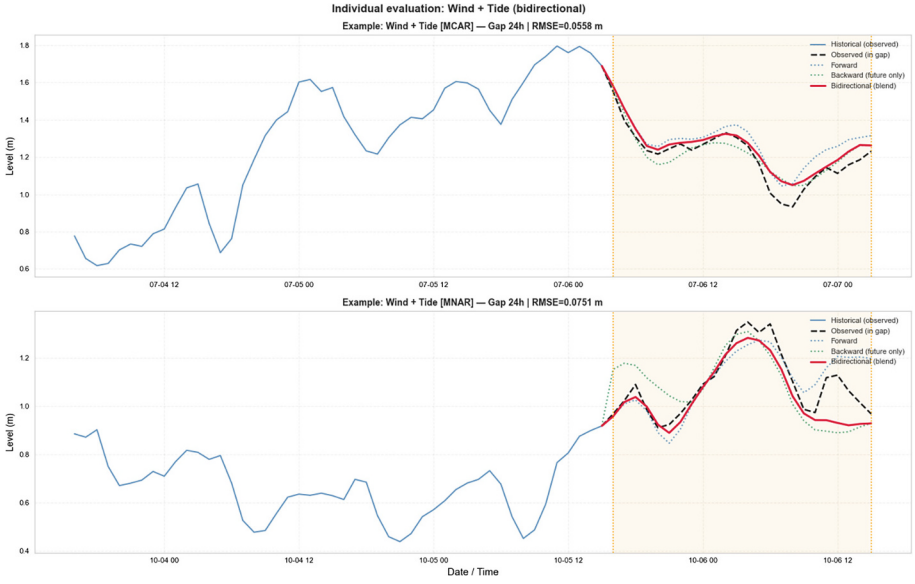

Fig. 8. Representative 24-hour gap reconstructions at MVD under MCAR (top) and MNAR (bottom) conditions. Panel shows historical observed levels (solid blue), observed levels within the gap (dashed black), forward prediction (blue dotted), backward prediction (green dotted), and bidirectional blend (solid red). (Color figure online)

Figure 8 illustrates representative 24-hour reconstructions under both mechanisms. In the MCAR example (RMSE = 10.9 cm), the forward prediction overshoots the observed tidal recession while the backward prediction, initialised from the post-gap context, tracks the descent more closely; the bidirectional blend combines both to reduce the error throughout the gap. In the MNAR example (RMSE = 11.8 cm), which captures a high-water surge event, neither unidirectional prediction alone reproduces the sharp peak and subsequent recession; the blend narrows the error by drawing on the pre-surge context from the forward pass and the post-surge boundary from the backward pass. Both cases confirm that post-gap observations, when available, are essential for accurate reconstruction of non-stationary tidal segments. Overall, bidirectional imputation with the Wind + Tide model is operationally viable for gaps up to 24 h, covering the large majority of interruptions recorded in the RdP tide gauge network.

6 Conclusions

This paper presented the development and evaluation of two LightGBM-based tide level forecasting models and a bidirectional predictive imputation scheme for the Río de la Plata estuary, integrated within the PronUy_RPFM operational forecasting system.

PronUy_LightGBM1 demonstrated that an autoregressive LightGBM model trained solely on tide gauge history yields competitive multi-step predictions across all three stations: mean RMSE ranges from 0.130 m (IdF, 6 h) to a peak of 0.310 m (BsAs, 48 h), with Isla de Flores consistently outperforming the other stations in both absolute and normalised terms. Feature importance analysis revealed that the immediate single-step lag dominates at all locations, but the structure of secondary predictors reflects the underlying forcing regime: periodic lags at 24 and 48 h are prominent at MVD and IdF (meteorologically dominated), while rolling variability features characterise BsAs (astronomically dominated). Incorporating external forcing in PronUy_LightGBM2 produced systematic RMSE reduction at all horizons at MVD. The combined Wind + Tide configuration proved the most effective, achieving up to 40.9% RMSE reduction at 72 h, with the physics-based PronUy_RPFM sea level forecasts being the more informative single source (up to 31.6% reduction at 72 h) relative to wind-only forcing (up to 24.7%). All variants exhibited an error plateau beyond 12 h: a qualitative shift from the monotonically growing error of the autoregressive baseline, confirming that external inputs effectively compensate for recursive drift accumulation at longer horizons.

The bidirectional Wind + Tide strategy achieved RMSE below 10 cm for gaps up to 24 h under MCAR conditions, and below 12 cm for the same lengths under the MNAR scenario, where gaps are biased toward high-water surge events. The convergence of MCAR and MNAR errors at 72 h (16.1 cm vs 15.9 cm) indicates a regime shift: for very long gaps the dominant error source is temporal distance from the nearest anchor rather than event severity. A representative

24-hour reconstruction ilustrated the advantage of the bidirectional blend over the forward-only prediction, particularly in segments with non-stationary tidal recession.

Several directions merit further investigation. In the first place, extending PronUy_LightGBM2 to BsAs and IdF, where wind and numerical model forcing data may be available, would test the generalisability of the improvement observed at MVD. The consistently superior performance of the combined Wind + Tide configuration suggests that both forcing sources contribute non-redundant information, and further analysis of their individual contributions at different meteorological regimes could help prioritise data acquisition for operational deployment. For imputation, incorporating the astronomical tide prediction as an additional input feature – or, alternatively, decomposing the signal into its astronomical and meteorological residual components prior to imputation – represents a natural extension of the current framework. Further investigation into the non-monotonic MCAR error behaviour at medium gap lengths (12–24 h) could also clarify whether it reflects genuine variability in tidal dynamics or a sampling effect of the test period. Finally, integrating both the forecasting and imputation modules into the live PronUy_RPFM pipeline, thus enabling real-time gap detection, automatic reconstruction, and seamless handoff to the forecasting engine constitutes the primary operational objective of this work.

Acknowledgments. This work is supported by funding from the National Agency for Research and Innovation (Agencia Nacional de Investigación e Innovación – ANII) FMV project "Sistema de pronóstico en tiempo real de niveles de marea en el Río de la Plata enriquecido con asimilación de datos".

The authors gratefully acknowledge the WRF model data provided by IMFIA researchers Victor Silveira and Alejandro Gutiérrez.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2623–2631. (2019)
2. Bojer, C.S., Meldgaard, J.P.: Kaggle forecasting competitions: an overlooked learning opportunity. Int. J. Forecast. **37**(2), 587–603 (2021)
3. Cao, W., Wang, D., Li, J., Zhou, H., Li, L., Li, Y.: BRITS: bidirectional recurrent imputation for time series. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
4. Du, W., et al.: TSI-Bench: benchmarking time series imputation (2024). https://arxiv.org/abs/2406.12747
5. Elshorbagy, A., Corzo, G., Srinivasulu, S., Solomatine, D.P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology. Hydrol. Earth Syst. Sci. **14**, 1931–1941 (2010)

6. Elshorbagy, A., Corzo, G., Srinivasulu, S., Solomatine, D.P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application. Hydrol. Earth Syst. Sci. **14**, 1943–1961 (2010)
7. Fossati, M., Balparda, D., Sellanes, L., Silva, D., Jackson, M., Ezzatti, P.: Desarrollo del sistema de pronóstico del Río de la Plata y su frente marítimo: PronUy_RPFM. Ribagua **9**(2), 25–40 (2022)
8. Fossati, M., et al.: Dinámica de flujo, del campo salino y de los sedimentos finos en el Río de la Plata. RIBAGUA - Revista Iberoamericana del Agua **1**, 48–63 (2014)
9. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
10. Gan, M., et al.: An improved machine learning-based model to predict estuarine water levels. Ocean Model. **190**, 102376 (2024)
11. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why tree-based models still outperform deep learning on tabular data. In: Advances in Neural Information Processing Systems. vol. 35 (2022)
12. Guerrero, R.A., Acha, E.M., Framinán, M.B., Lasta, C.A.: Physical oceanography of the Río de la Plata Estuary. Argentina. Continental Shelf Res. **17**(7), 727–742 (1997). https://doi.org/10.1016/S0278-4343(96)00061-1
13. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. vol. 30, pp. 3146–3154 (2017)
14. Körner, P., Kronenberg, R., Genzel, S., Bernhofer, C.: Introducing gradient boosting as a universal gap filling tool for meteorological time series. Meteorol. Z. **27**(5), 369–376 (2018). https://doi.org/10.1127/metz/2018/0908
15. McElfresh, D., et al.: When do neural nets outperform boosted trees on tabular data? In: Advances in Neural Information Processing Systems. vol. 36 (2023)
16. Torres, M., Klapp, J., Gitler, I., Tchernykh, A. (eds.): ISUM 2018. CCIS, vol. 948. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10448-1
17. Ocio, D., Beskeen, T., Smart, K.: Fully distributed hydrological modelling for catchment-wide hydrological data verification. Hydrol. Res. **50**, 1520–1534 (2019)
18. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976). https://doi.org/10.1093/biomet/63.3.581
19. Santoro, P., Fossati, M., Piedra Cueva, I.: Study of the meteorological tide in the Río de la Plata. Continental Shelf Res. **60**, 51–63 (2013). https://doi.org/10.1016/j.csr.2013.04.018
20. Savenije, H.H.G.: Prediction in ungauged estuaries: an integrated theory. Water Resour. Res. **51**, 2464–2476 (2015). https://doi.org/10.1002/2015WR016936
21. Sellanes, L., Fossati, M., Silva, D., Ezzatti, P.: Mejoras en la implementación del modelo numérico base del pronóstico de niveles del Río de la Plata y frente marítimo: PRONUY_RPFM. In: Mecánica Computacional. vol. XL, pp. 1681–1690 (2023)
22. Silveira, V., Sarazola, I., Gutiérrez, A., Fossati, M.: Offshore surface wind map in Río de la Plata and Atlantic Ocean shelf. J. Wind Eng. Industrial Aerodyn. **258** (2025)
23. Skamarock, W.C., et al.: A description of the Advanced Research WRF model version 4. Technical report. NCAR/TN-556+STR, National Center for Atmospheric Research, Boulder, CO, USA (2019)
24. Solomatine, D.P., Ostfeld, A.: Data-driven modelling: some past experiences and new approaches. J. Hydroinf. **10**, 3–22 (2008)
25. Sun, Y., Wang, R., Qi, C., Xu, J., Tu, Z., Yang, F.: An improved tidal prediction method using meteorological parameters and historical residual water levels. Appl. Ocean Res. **153**, 104289 (2024)

26. Wu, C.-C., Chen, Y.-L., Liu, Y.-H., Yang, X.-Y.: Decision tree induction with a constrained number of leaf nodes. Appl. Intell. **45**(3), 673–685 (2016). https://doi.org/10.1007/s10489-016-0785-z
27. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11121–11128 (2023)