TAL TECH

DOCTORAL THESIS

Development of Machine Learning Approaches for Geoid-referred Sea Level Forecasting

Saeed Rajabi Kiasari

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2026

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
20/2026

Development of Machine Learning Approaches for Geoid-referred Sea Level Forecasting

SAEED RAJABI KIASARI

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Engineering
Department of Civil Engineering and Architecture
This dissertation was accepted for the defence of the degree 20/03/2026

Supervisor: Prof. Artu Ellmann
School of Engineering
Department of Civil Engineering and Architecture
Tallinn University of Technology, Tallinn, Estonia

Co-supervisor: Ass. Prof. Nicole Delpeche-Ellmann
School of science
Department of Cybernetics, Laboratory of Wave Engineering
Tallinn University of Technology, Tallinn, Estonia

Pre-reviewer Dr. Sander Varbla
School of Engineering
Department of Civil Engineering and Architecture
Tallinn University of Technology, Tallinn, Estonia

Opponents: Dr. Svetlana Jevrejeva
Ocean-Shelf Processes Group
National Oceanography Centre, Liverpool, UK

Dr. Marcello Passaro
German Geodetic Research Institute
Department of Aerospace and Geodesy
Technical University of Munich (TUM), Munich, Germany

Defence of the thesis: 24/04/2026, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology has not been submitted for doctoral or equivalent academic degree.

Saeed Rajabi Kiasari

---
signature

Estonian Research Council

ISSN 2585-6898 (publication)
ISBN 978-9916-80-477-3 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-478-0 (PDF)
DOI https://doi.org/10.23658/taltech.20/2026
Printed by Koopia Niini & Rauam

Rajabi Kiasari, S. (2026). *Development of Machine Learning Approaches for Geoid-referred Sea Level Forecasting* [TalTech Press]. https://doi.org/10.23658/taltech.20/2026

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
20/2026

Masinõppe meetodite väljatöötamine geoidi suhtes määratletava merepinna taseme prognoosimiseks

SAEED RAJABI KIASARI

TAL TECH
KIRJASTUS

Contents

List of Publications

The dissertation has been prepared based on the following peer-reviewed journal articles (indexed by SCOPUS and WOS):

I **Rajabi-Kiasari, S.**, Delpeche-Ellmann, N., Ellmann, A. (2023). Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea. Computers & Geosciences, 178, 105406, doi: doi.org/10.1016/j.cageo.2023.105406.

II **Rajabi-Kiasari, S.**, Ellmann, A., Delpeche-Ellmann, N. (2025). Sea level forecasting using deep recurrent neural networks with high-resolution hydrodynamic model. Applied Ocean Research, 157, 104496, doi: doi.org/10.1016/j.apor.2025.104496.

III **Rajabi-Kiasari, S.**, Delpeche-Ellmann, N., Ellmann, A. and Soomere, T., (2026). Forecasting sea level maxima using Machine learning with explainability and extreme value analysis. International Journal of Applied Earth Observation and Geoinformation, 146, p.105064, doi: doi.org/10.1016/j.jag.2025.105064.

Author's Contribution to the Publications

Contributions to the **Publications I-III** in this dissertation are:

I Conceptualization of the idea in cooperation with all co-authors; developing the methodology, software and conducting analysis; validation of the results in cooperation with all co-authors; drafting the manuscript; reviewing and editing the manuscript in cooperation with all co-authors; visualizing the results.

II Developing the methodology, software and analysis; validation of the results in cooperation with all co-authors; drafting the manuscript; visualizing the results; reviewing and editing the manuscript in cooperation with all co-authors; conceptualization of the idea in cooperation with co-authors.

III Conceptualization of the idea in cooperation with all co-authors; developing the methodology, software and analysis and visualization; validation of the results in cooperation with all co-authors; drafting the manuscript; reviewing and editing the manuscript in cooperation with all co-authors.

Introduction

Increasing socio-economic pressures together with a changing climate require accurate sea-level forecasting across a broad range of spatial and temporal scales, encompassing both mean and extreme water levels conditions (cf. Figure 1). For instance, in marine engineering and navigation, international standards and guidelines (e.g., International Hydrographic Organization (IHO)) emphasize the need for high-accuracy short-term (hours to days) water level information to ensure safe navigation, reliable under-keel clearance (UKC) assessment, and effective port and dredging operations (IHO, 1987). This requirement is particularly critical during periods of elevated sea levels associated with storm surge, tides, and their non-linear interactions (Orseau et al., 2021). For longer timescales (years to decades), Intergovernmental Panel on Climate Change (IPCC) assessment highlight that changes in sea-level extremes and maxima, in addition to mean sea-level rise, represent major sources of uncertainty and risk for the design, performance, and service life of coastal and offshore infrastructure (IPCC, 2021).

Over the years, significant advances have been made in sea level observing systems, including in-situ measurements, numerical modelling, satellite remote sensing and airborne laser scanning observations (Varbla, 2023; Jahanmard, 2024). Despite this, current forecasting approaches continue to face challenges in consistently resolving sea-level variability and extremes with the accuracy, spatial coherence, and reliability required across scales relevant to both operational decision-making and long-term risk assessment.

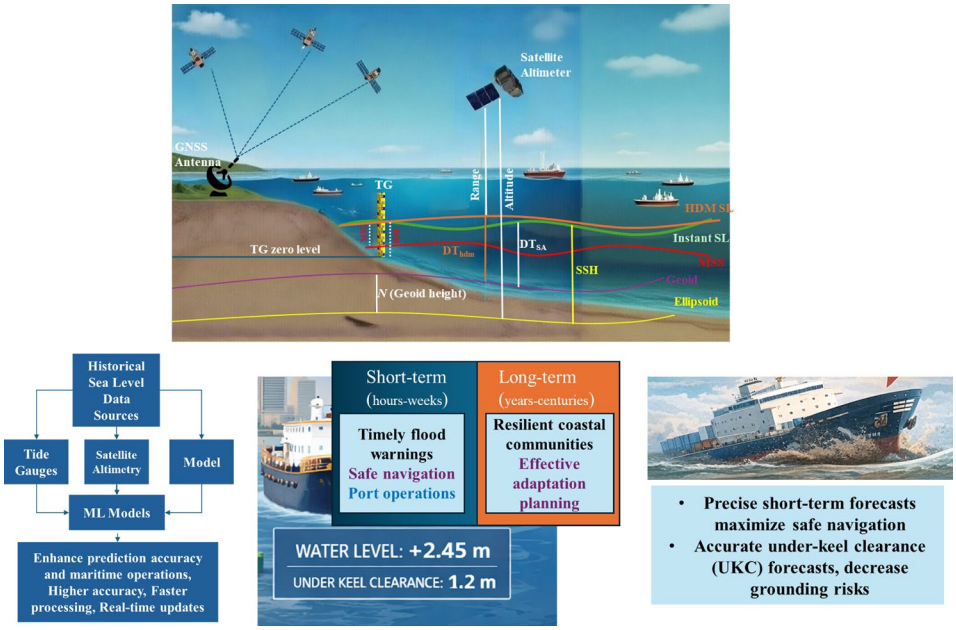


Figure 1. Sea level forecasting data sources, references and variables (top) along with the challenges and ML applications for sea level forecasting (bottom). Abbreviations: TG = Tide Gauge; DT = Datum Transformation; SSH = Sea Surface Height; HDM = Hydrodynamic Model; SL = Sea Level; GNSS = Global Navigation Satellite System; MSS = Mean Sea Surface.

Additionally, the growing number and diversity of available sea-level forecasting models and products complicates model selection, and limited guidance exist on which approaches are most suitable for specific applications, accuracy requirements, and temporal scales. Thus, the overall goal of this dissertation is to evaluate, improve, and provide guidance on sea-level forecasting approaches for marine engineering and navigation applications, on the accurate representation of sea-level variations and extremes across relevant spatial and temporal scales.

Historically, numerical simulations based on hydrodynamic models (HDMs) have been the primary tool for regional and global sea-level prediction (Ayinde et al., 2024). While these models are physically interpretable and widely used in both hindcast and operational forecasting systems, they have inherent limitations. For example, their computational cost can be substantial, and generating forecasts for different temporal horizons often requires repeated re-initialization and re-execution, particularly when applied to complex coastal and ocean environment (Calkoen et al., 2021; El Aouni et al., 2025). Also, their predictive skill depends strongly on the accuracy of initial and boundary conditions and on model parameterizations and discretization, which can introduce systematic biases (Alemseged & Rientjes, 2007; Ganju et al., 2016).

Such limitations in recent years have motivated growing interest in data-driven approaches for sea-level forecasting, particularly machine learning (ML) methods. By learning statistical relationships directly from observations and model outputs, ML techniques can offer computationally efficient alternatives that are well suited to rapid forecasting, scenario exploration, and the integration of various data sources. ML models are typically trained offline, thus once trained they can be deployed efficiently for near-real-time prediction. This reduces computational demands during short-term forecasting where latency is a constraint. Also, recent studies have demonstrated the potential of ML to capture complex, non-linear interactions among atmospheric forcing, tides and storm surges (Jahanmard et al., 2023; Hordoir et al., 2026), which are critical for representing sea-level variability and extremes.

Despite these advances, ML-based approaches also face several important challenges. Model performance depends strongly on input predictors used and their temporal scales. Furthermore, questions remain on their generalization, interpretability, resolutions required and robustness, particularly when models are applied outside the range of conditions represented in the training data and or when extremes are poorly sampled due to their limited occurrence. These limitations highlight four main knowledge gaps that this dissertation explores.

Firstly, a key challenge in ML-based sea level forecasting concerns data availability, quality and integration. Sea level observational datasets including tide gauges (TGs), satellite altimetry (SA), HDMs, Airborne Laser Scanners (ALS) and Global Navigation Satellite System (GNSS) offer complementary strength but differ substantially in spatial coverage, temporal resolutions, and associated uncertainties (Varbla et al., 2021; Mostafavi et al., 2024; Jia et al., 2022; Varbla et al., 2022). While using these data individually often introduces biases (Jahanmard et al., 2023), their synergistic integration by utilizing the geoid (the shape of the equipotential ocean surface under the influence of the gravity and rotation of Earth alone) as the unified vertical datum, ensures vertical consistency between the various data sources. This in effect allows the determination of dynamic topography (DT) which is critical for applications requiring half-decimetre accuracy in navigation and UK (Varbla et al., 2020, Jahanmard et al., 2021). This dissertation investigates and demonstrates such integrative approaches in **Publications I** and **II**,

showing that the combined use of these data sources leads to improved prediction accuracy, increased reliability, and opportunities for independent validation of sea-level forecasts. An approach that so far has not been utilized in such a manner in sea level forecasting.

Secondly, developing high-resolution spatiotemporal forecasting models is essential to capture rapid, localized sea-level changes that drive coastal risks, especially in dynamic semi-enclosed basins like the Baltic Sea. Spatial resolutions of 0.5–1 nautical mile resolve small-scale coastal features, eddies, and fronts (Soomere et al., 2008; Kowalewski & Kowalewska-Kalkowska, 2017), while high temporal resolution (e.g., hourly) tracks short-term atmospheric forcing such as wind and air pressure more accurately (Medvedev et al., 2021; Kulikov & Medvedev, 2013). This dissertation also addresses the multivariate nature of sea-level forecasting, as effective predictions must account for tides, storm surges, wind waves, and mean sea-level rise. For example, Nieves et al. (2021) showed temperature is the dominant driver of thermosteric sea-level change, but their model underperformed in regions influenced by other factors, highlighting the need for multiple predictors in multivariate forecasting. Multi-step ahead forecasting is also critical, as it predicts not only peak water levels but also the trajectory, rate of change, and duration of hazardous conditions. This detailed foresight supports safety-critical applications such as maritime navigation, storm surge forecasting, and coastal management, enabling timely and precise warnings to reduce damage and save lives.

Thirdly, an important challenge of ML based methodologies, is the tendency to misrepresent extremes and maxima, for they are nonlinear in space and time. Although relatively infrequent, such events can have severe consequences, underscoring the importance of accurately representing them in forecasting applications. This observation was highlighted in **Publications I** and **II**. Understanding that extremes are often site-specific and primarily driven by intense storms in combination with multiple compound factors (described in detail in Section 1.5) led to further investigation in **Publications III**. The rarity of extreme events in training sets is well known (Ramos-Valle et al., 2021; Qin et al., 2023). Even process-based physical models have also faced such challenges for extreme sea level (ESL) forecasting (Lorenz et al., 2025). To mitigate this, we introduced a novel strategy in **Publications III**: training ML models on sea level maxima (SLM) with long historical timespan of TG data and optimized feature set using Bayesian optimization (BO) coupled with an extreme-value analysis.

Fourthly, ML approaches are often characterized as “black-box” methods. This dissertation directly addresses the issue of explainability by adhering to a core principle: model performance and reliability are fundamentally governed by the quality and relevance of input data. Accordingly, in **Publications I**–**III**, a statistical analysis is initially performed to identify physically meaningful predictors and their appropriate spatiotemporal scales for sea-level variability. This domain-informed feature selection is then rigorously validated using explainability analyses, specifically SHAP (SHapley Additive exPlanations), to quantify individual feature contributions, and ensure the model’s logic aligns with prior physical knowledge. This integrated methodology—combining physics-based input curation with post-hoc algorithmic transparency—constitutes a novel framework for developing trustworthy ML models in sea-level forecasting. By producing more transparent and physically credible projections, this work provides a foundation for more confident, actionable and resilient coastal planning and risk management (Ayinde et al., 2024).

The Baltic Sea serves as an excellent natural laboratory for addressing these challenges. It is a semi-enclosed, micro-tidal basin in northern Europe. Surrounded by nine countries, it is highly active in maritime and recreational uses such as navigation, shipping, ports, beaches, offshore wind parks, pipelines and cables, which connects the surrounding nations to one another and to the rest of the world. The hydrodynamic is such that the Danish Straits connect the Baltic Sea to the Atlantic Ocean, allowing intermittent saline water inflow, while the Baltic receives large amounts of freshwater input from rivers and land runoff (Leppäranta & Myrberg, 2009). Thus, there is often persistent outflow. Short-term sea-level variations are mainly controlled by the wind's regime and atmospheric pressure. Wind can be anisotropic, complicating the sea level predictability. The region also benefits from extensive observational and modelled datasets, including long historical TG records, a well-established geodetic infrastructure, high-resolution HDM simulations, a range of advanced SA products such as Baltic+SEAL (Passaro et al., 2021) that specifically considers the complexities of the study area. This abundance of high-quality, complementary data provides a solid foundation for training and validating advanced ML frameworks.

Studies on forecasting sea level anomalies in the Baltic Sea combine HDM and ML to support real-time predictions and long-term risk assessments. Research has progressed from statistical analyses showing that wind and pressure dominate sea-level variability, to operational forecasting systems for short-term predictions and long-term risk, including global sea-level rise and land uplift effects. ML now seems to complement traditional models by supporting the short-term forecasts. Across studies, wind forcing, basin-scale preconditioning, cyclonic effects and vertical land motion remain the main drivers of variability and extremes. More details on the Baltic Sea dynamics are discussed in Section 1.5.

This dissertation aims to advance the state of the art in ML-based sea-level forecasting by bridging four challenging key areas: the integration of multi-source observational and modelled data, the development of high-resolution and multivariate multi-step ahead spatiotemporal prediction models, the improved treatment of extreme events, and providing explainable physically consistent forecasts. An additional component of this work concerns the treatment of sea-level reference systems. While many operational or ML-based sea-level forecasts are expressed in relative reference frames—typically tied to local TG datum—this dissertation focuses on forecasting sea levels in an absolute reference system (i.e., DT as geoid-referred sea level). This approach enables consistent comparisons across observation types, facilitates the integration of SA and GNSS data, and supports broader geophysical interpretation.

To the best of our knowledge, this study is the first to apply ML-based spatiotemporal forecasting of regional sea level in the Baltic Sea region. It achieves an unprecedented resolution for ML-based sea level forecasting. Furthermore, the results are referenced to the geoid, enabling direct comparison with observational datasets such as SA—an aspect not explored in previous research. This work is also among the few ML-based studies focused on forecasting sea level maxima and extremes within an explainable framework.

Scope and Objectives

In this Section, we summarize our main objectives for addressing the existing challenges in ML-based sea-level forecasting in the Baltic Sea, as outlined in the introduction.

This dissertation tackles key challenges in ML-based sea-level forecasting by developing models that produce forecasts in an absolute reference system (the geoid), enabling the seamless integration of multiple data sources. Rather than relying on models trained solely on TG data—which limits performance to nearshore areas—this study advances spatiotemporal sea-level forecasting, an approach that has gained increasing attention in recent years. This is specifically addressed in **Publications I** and **II**, with **Publication II** achieving an unprecedented hourly resolution at one-nautical-mile resolution coupled with high-performance computing (HPC).

The models are designed to support both day-ahead **(Publications I** and **III**) and multiple-hours ahead (**Publication II**) short-term forecasting. To improve predictive skill, they incorporate multiple physical input features selected based on domain knowledge and explanatory analysis. Although more complex, such multivariate approaches better capture the nonlinear interactions between oceanic and atmospheric processes (Tur et al., 2021; Ayinde et al., 2024). This multivariate framework is employed throughout **Publications I–III**.

External validation of ML-based results is another critical concern. Many current ML-based sea-level studies use the same datasets for both training and evaluation, risking overfitting and reduced generalizability. Independent validation against multiple observational sources, such as SA, is essential to ensure model robustness across different data types. This practice, rarely addressed in previous studies, is specifically implemented in **Publications I** and **II**.

Additionally, model explainability is incorporated to make outputs more understandable for decision-making. This is particularly important when dealing with extreme events studies, addressed in **Publication III**.

These considerations define the scope of this research, which aims to develop sea-level forecasting using a geoid-referenced datum and the integration of multiple data sources. The methodology addresses spatiotemporal coverage, high-resolution modelling, multivariate and multi-step-ahead predictions, explainable frameworks, and extreme sea-level analysis. By tackling these challenges, the work provides a comprehensive, robust, and interpretable approach to high-resolution sea-level forecasting in complex coastal environments.

In summary, this dissertation applies ML-based sea-level forecasting through two complementary approaches:

- **Method I** develops deep neural networks combined with geoid-referenced hydrodynamic model (HDM) outputs for spatiotemporal forecasting, with validation against independent SA observations.
- **Method II** develops ML models to forecast daily sea-level maxima within an explainable framework, coupled with extreme value analysis for robust assessment of extreme events.

Accordingly, the main objectives of this dissertation are summarized as follows:

- Demonstrate geoid-referenced sea-level forecasting that resolves both coastal and offshore dynamics throughout the Baltic Sea basin (**Publication I–II**).
- Develop a regional and Multivariate ML framework based on convolutional neural networks (CNNs) for sea-level forecasting across the Baltic Sea (**Publications I**).
- Develop a high-resolution (one-nautical-mile, hourly) and multi-step ahead ML-based forecasting system via deep recurrent neural networks (Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)) for dynamic Baltic sub-basins (e.g., the Gulf of Finland) (**Publication II**).
- Validate ML-based sea level forecasts using independent SA observations (**Publication I–II**).
- Investigate sea-level maxima and extreme events in ML forecasting frameworks to improve extreme-event prediction (**Publication III**).
- Integrate explainability techniques to identify key physical drivers of extreme sea-level events and enhance model transparency (**Publication III**).
- Coupling of short-term ML-based SLM forecasting with extreme value analysis to quantify uncertainties and support long-term coastal risk assessment (**Publication III**).

Limitations

This study found following limitations related to data, design of networks and computational constraints:

- In **Publication I**, we studied the performance of a 2D CNN (Conv2D) model for daily sea level forecasting in the Baltic Sea region. A key limitation of this study was the reduced spatiotemporal resolution of the input data. Instead of using the original high-resolution (hourly, one-nautical-mile) data from the corrected HDM, we employed a lower-resolution scheme. This compromise was necessary to ensure computational efficiency while maintaining the model's ability to perform daily forecasts at regional scales.
- The resolution limitation of **Publication I** was addressed in **Publication II**, where we utilized high-resolution (hourly, one-nautical-mile) data exploiting LSTM and GRU modelling approaches. However, this study had different constraints:
    - Due to the high-resolution framework, the forecasting horizon could not be extended beyond 24 hours. Instead, we fixed the multi-step predictions at 3, 6, 9, 12, and 24 hours ahead.
    - Despite using HPC resources, computational limitations restricted the study to a sub-basin of the Baltic Sea—the Gulf of Finland.
    - The multivariate forecasting framework did not evaluate the impact of additional meteorological metrics, and the look back parameter for historical timesteps was limited to 12 hours. Despite these limitations, the models performed well in forecasting spatially distributed grid points across the Gulf of Finland.
- For **Publications I and II**, a common limitation was the lack of model explainability, stemming from the nature of sequence-to-sequence DL approaches having a large number of model weights. This was specifically addressed **Publication III** which introduced explainability techniques, which are essential for sea-level–related decision-making.
- In **Publication III**, the proposed strategy reduced the underestimation level of ML models (from 100 cm in **Publications I** and **II** to around 150 cm). However, challenges remain in improving rare extreme-event forecasting. Future work should test the methodology on TGs with longer time spans (e.g., century-long records). Additionally, while the study focused on daily SLM forecasting, expanding to longer horizons was not feasible due to the complexity of applying explainability techniques to sequence-to-sequence models.

Structure

The dissertation is organized as follows. Section 1 provides the general background on sea-level forecasting, including the primary contributors to sea-level variability, the fundamental challenges associated with forecasting, available sea-level data sources, and issues related to vertical reference datum inconsistencies. This Section also contains relevant characteristics of the study area. Section 2 focuses on ML–based sea-level forecasting, reviewing the current state of the literature and presenting the theoretical framework and methodological background of the ML models developed in this study, along with descriptions of the evaluation metrics utilized.

Section 3 presents the results of the spatiotemporal DL–based sea-level forecasting implementation. Section 4 examines extreme sea-level patterns and provides a comparative analysis of different ML approaches and their performance for SLM forecasting. Finally, Section 5 summarizes the key findings of the dissertation, discusses the implications of the results, and outlines directions for future research.

Abbreviations

Adam	Adaptive Moment Estimation
ADCIRC	ADvanced CIRCulation Model
ADT	Absolute Dynamic Topography
AI	Artificial Intelligence
ALS	Airborne Laser Scanners
ANN	Artificial Neural Networks
ASL	Absolute Sea Level
BIC	Bayesian Information Criterion
BO	Bayesian Optimization
BSCD	Baltic Sea Chart Datum
BSHC	Baltic Sea Hydrographic Commission
CNN	Convolutional Neural Network
Conv2d	Two-dimensional Convolutional Neural Network
DAC	Dynamic Atmospheric Correction
DL	Deep Learning
DOT	Dynamic Ocean Topography
DOY	Day of the Year
DT	Dynamic Topography
EDA	Exploratory Data Analysis
ELU	Exponential Linear Unit
ENSO	El Niño–Southern Oscillation
ERA5	ECMWF Reanalysis V5
ESL	Extreme Sea Level
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FVCOM	Finite-Volume Coastal Ocean Model
GETM	General Estuarine Transport Model
GEV	Generalized Extreme Value Distribution
GIA	Glacial Isostatic Adjustment
GLDAS	Global Land Data Assimilation System
GPR	Gaussian Process Regression
GNSS	Global Navigation Satellite System
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HDM	Hydrodynamic Model
HIROMB-BOOS	High-Resolution Operational Model for the Baltic and the Baltic Operational Oceanographic System
HPC	High Performance Computing
HYCOM	HYbrid Coordinate Ocean Model
IPCC	Intergovernmental Panel on Climate Change

KNN	K-Nearest Neighbors
LightGBM	Light Gradient-Boosting Machine
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
MERRA	Modern Era Retrospective-Analysis for Research and Applications
MI	Mutual Information Index
ML	Machine Learning
MLP	Multilayer Perceptron
MODIS	Moderate Resolution Imaging Spectroradiometer
MSLP	Mean Sea Level Pressure
MSS	Mean Sea Surface
NAO	North Atlantic Oscillation
Narva	Narva-Jõesuu Tide Gauge Station
NEMO	Nucleus for European Modelling of the Ocean
NKG2015	Nordic Geodetic Commission developed geoid model, year 2015
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficients
PICP	Prediction Interval Coverage Probability
PT	Pole Tides
ReLU	Rectified Linear Unit
RF	Random Forest
RMSE	Root Mean Squared Error
RMSProp	Root Mean Square Propagation
RNN	Recurrent Neural Networks
ROMS	Regional Ocean Modeling System
RSL	Relative Sea Level
SA	Satellite Altimetry
SAR	Synthetic-Aperture Radar
SET	Solid Earth Tides
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive Explanations
SLA	Sea Level Anomaly
SLM	Sea Level Maxima
SLP	Sea Level Pressure
SMHI	Swedish Meteorological and Hydrological Institute
SSB	Sea State Bias
SSH	Sea Surface Height
SSHA	Sea Surface Height Anomaly
SSS	Sea Surface Salinity
SST	Sea Surface Temperature
SSTA	Sea Surface Temperature Anomalies

SVM	Support Vector Machine
SVR	Support Vector Regression
SWH	Significant Wave Height
SWOT	Surface Water Ocean Topography
Tanh	Hyperbolic Tangent
TG	Tide Gauge
UKC	Under Keel Clearance
VLM	Vertical Land Movement
XAI	Explainable Artificial Intelligence
XGB	Extreme Gradient Boosting

Symbols

SL	Sea level
φ	Latitude
λ	Longitude
δ	Lag window
t	Time
K	Forecast horizon
H	Orthometric height
h	Ellipsoidal height
N	Geoid height
$DT_{(\varphi,\lambda)}$	Dynamic topography
$SSH_{(\varphi,\lambda)}$	Sea surface height
H_{orbit}	Satellite altitude
R	Satellite range
$iono$	Ionospheric correction
wtc	Wet tropospheric correction
dtc	Dry tropospheric correction
SSB	Sea state bias correction
PT	Pole tide correction
SET	Solid earth tide correction
$H_{mean-tide}$	Height in the mean-tide system
$H_{zero-tide}$	Height in the zero-tide system
φ_B	Geodetic latitude of the desired benchmark
φ_{NAP}	Geodetic latitude of Normaal Amsterdams Peil (NAP)
MSE	Mean squared error
$RMSE$	Root mean squared error
R	Correlation coefficient
R^2	Coefficient of determination
$PICP$	Prediction Interval Coverage Probability
MI	Mutual information
BIC	Bayesian information criterion
TP	True Positive
FP	False Positive
FN	False Negative
y_i	Observed sea level
$\hat{y}_i$	Forecasted sea level
$\bar{y}$	Average of observed sea level
n	Number of observations
$H_{i.j}^{K}$	Output feature maps in convolutional layers
$W_{m,n}^{(k)}$	Weights of the convolutional layers

$X_{i+m-1,j+n-1}^{(m,n)}$	Input feature map in convolutional layers
b_k	Bias term for each output channel in convolutional layers
$Z_{i,j}$	Pooled feature map in convolutional layer
$W^{(l)}$	Weights in the fully connected layer
$h^{(l)}$	Hidden neurons in the fully connected layers
$b^{(l)}$	Bias term in the fully connected layers
Λ	Controlling factor for length of regularization term
f_t	Forget gate in LSTM
i_t	Input gate IN LSTM
$\hat{C}_t$	Cell candidate in LSTM
o_t	Output gate in LSTM
h_t	Hidden state
W_i	Weights of input gate in LSTM
W_f	Weights of forget gate in LSTM
h_{t-1}	Previous hidden state in LSTM and GRU models
x_t	Input signal for LSTM and GRU models
b_f	Bias term for the forget gate in LSTM
b_i	Bias term for the input gate in LSTM
b_c	Bias term for the
b_o	Bias term for the output gate in LSTM
P	Activation function
z_t	Update gate in GRU model
r_t	Reset gate in GRU model
b_z	Bias term in GRU model
b_r	Bias term of reset gate for GRU model
W_z	Weights of update gate in GRU
W_r	Weights of reset gate in GRU
h_t	Hidden state of GRU model
$L(\emptyset)$	Loss function
$\Omega(f_t)$	Regularization term
η	Learning rate
ϕ_i	SHAP explainability values
ξ	Shape parameter of GEV fit
μ	Location parameter of GEV fit
σ	Scale parameter of GEV fit
L_i	Lower prediction band in PICP formula
U_i	Upper prediction band in PICP formula

1 Background on Sea Level Dynamics and Forecasting

1.1 Components of Sea Levels

Sea-level change is a complex, multi-scale phenomenon driven by interactions among climatic, oceanic, and geophysical processes (cf. Figure 2). Understanding it, requires a framework that moves from detecting the global signal and attributing its causes, to downscaling projections for regional impacts, and applying this knowledge for coastal management. A major achievement has been the closure of the 20th-century sea-level budget, showing that observed global mean sea-level rise (GMSL) is fully explained by thermal expansion of warming oceans, glacier and ice sheet melt, and changes in land water storage, confirming that modern acceleration is largely anthropogenic (Church et al., 2001; Frederikse et al., 2020). Short-term trends, however, can be misleading due to natural variability, such as El Niño–Southern Oscillation (ENSO) cycles, which temporarily redistribute water and mask the long-term rise (Cazenave et al., 2014). Historical records show acceleration began in the late 18th century, with natural variability dominating early trends and human influence rising post-industrialization (Jevrejeva et al., 2008).

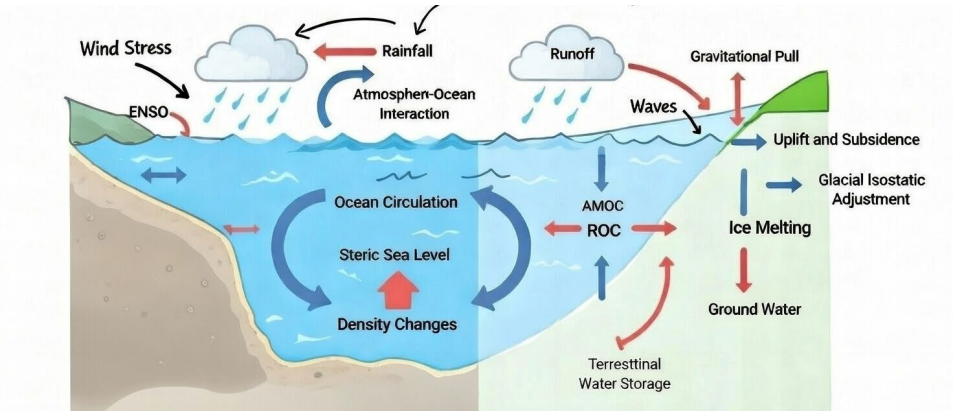


Figure 2. Sea level changes components. Abbreviations: ROC = Receiver Operating Characteristic; AMOC = Atlantic Meridional Overturning Circulation; ENSO = El Niño–Southern Oscillation.

While global mean sea level is a key indicator, coastal risk depends on regional relative sea level (RSL), which can deviate significantly due to steric (caused by variations in seawater density due to temperature and salinity changes) and dynamic ocean effects, ice-mass gravitational and rotational fingerprints, and vertical land motion from glacial isostatic adjustment (GIA) and local factors (Milne et al., 2009; Horton et al., 2018). Future risk reflects the compound effect of regionalized sea-level rise, natural variability, and short-term extremes. Accurate projections require integrating geological, instrumental, and model-based data to map changes in time, space, and probability (Muis et al., 2016; Horton et al., 2018). This evolution—from early societal impact studies (Devoy, 1987) to sophisticated, integrated assessments—now supports location-specific adaptation and resilience strategies.

Sea-level fluctuations occur across multiple timescales. On short timescales (hours to weeks), they are driven by meteorological forcing, including storm surges, wind waves,

and swell from distant storms, which can cause rapid extreme changes (Wunsch & Stammer, 1997; Muis et al., 2016). Accurate hazard modelling depends on precise boundary conditions and inputs (Alemseged & Rientjes, 2007).

On intermediate timescales (months to years), seasonal steric changes and interannual climate models, such as ENSO, North Atlantic Oscillation (NAO), and the Indian Ocean Dipole, redistribute ocean mass and heat, imprinting regional patterns and sometimes obscuring the long-term trend (Stammer et al., 2013; Cazenave et al., 2014; Dangendorf et al., 2014). On long timescales (decades to centuries), persistent GMSL rise is dominated by thermal expansion and ice melt, with glaciers driving most of the 20th century before thermal expansion became dominant post-1970 (Church et al., 2013; Frederikse et al., 2020). Regional deviations arise from ocean dynamics, ice-mass gravitational and rotational effects, vertical land motion, and human impacts on the terrestrial water cycle (Church et al., 2013; Horton et al., 2018; Frederikse et al., 2020; IPCC, 2021).

Coastal extremes are also shaped by interactions between local and large-scale ocean processes, with seasonal to decadal variability influencing their frequency and magnitude (Jevrejeva et al., 2024). For the Baltic Sea, sea level dynamics are particularly complex due to pronounced differences between subbasins. Its semi-enclosed nature and weak tidal range make variability primarily driven by meteorological and thermodynamic forces (Weisse et al., 2021; Wulff et al., 2001). Short-term extremes are mostly storm surges caused by wind stress, waves, and atmospheric pressure, with anomalies exceeding one meter (Wolski & Wiśniewski, 2021). On seasonal to decadal scales, steric changes, freshwater fluxes from rivers and precipitation, and large-scale climate patterns such as NAO modulate variability. Persistent vertical land motion from GIA also affects long-term relative sea-level trends. ESLs often result from the compound effect of these drivers, such as storm surges coinciding with high background water levels. More details are provided in Section 1.5.

1.2 From Physical Modelling to Data-Driven Sea Level Forecasting

Hydrodynamic models are physics-based computer models used to simulate and predict water levels and flows. They are typically based on the Navier–Stokes equations (or their simplified forms) and are forced by inputs from atmospheric, river, and sea-ice dynamics. These models apply fundamental physical laws—mass, momentum, and energy conservation—to represent water movement realistically.

Sea-level modeling systems now span a hierarchy from global climate models, which provide long-term projections of steric and circulation-driven changes, to regional and coastal ocean models (e.g., ROMS[1], HYCOM[2], and unstructured-grid systems), that resolve fine-scale variability (Ponte et al., 2019). While global models are essential for boundary conditions and climate projections, they lack the resolution and coastal process detail needed for accurate local forecasts. Coastal HDMs better capture shelf dynamics, tides, and storm surge processes, yet their accuracy depends on sparse observations and often struggles with river inflows, wave–current interactions, and other ancillary drivers. Data assimilation has emerged as a key mechanism to improve forecast skill by merging observations with model physics, but coastal-specific assimilation

[1] Regional Ocean Modeling System
[2] Hybrid Coordinate Ocean Model

techniques remain underdeveloped due to strong nonlinearities and limited nearshore data (Ponte et al., 2019).

To overcome these limitations, sea level forecasting evolved toward reduced-complexity and data-driven approaches that retain physical interpretability, while improving computational efficiency. Physically informed statistical models showed that most coastal sea level variability can be explained by a few local predictors, such as alongshore wind stress and atmospheric pressure, achieving predictive skill comparable to operational surge models (Tilburg & Garvine, 2004). Persistent regional biases in numerical weather prediction, especially in complex coastal settings, highlighted the need for complementary methods capable of adapting to nonlinear forcing (Wedam et al., 2009). This evolution culminated in hybrid and ML frameworks in which neural networks emulate computationally expensive parameterizations or directly learn nonlinear forcing–response relationships from data, delivering substantial speedups without compromising physical consistency (Krasnopolsky & Fox-Rabinovitz, 2006).

Forecasting Sea-Level variations in the Baltic Sea is a mature operational field that integrates physics-based HDMs, like NEMO[3] or HIROMB-BOOS[4], with modern data-driven methods including ML and neural networks—an approach pioneered in early storm surge research (Sztobryn, 2003)—to provide real-time predictions and long-term climate risk assessments (Buch, 2002; Meier et al., 2006; Lagemaa et al., 2011; Barzandeh et al., 2025; Bellinghausen et al., 2025). Despite advances in forecasting, several key limitations remain. Hydrodynamic models systematically misrepresent ESLs due to inaccuracies in atmospheric forcing and insufficient coastal resolution (Lagemaa et al., 2011), and there are inherent predictability limits in capturing the timing and magnitude of extreme events (Bellinghausen et al., 2025). Additionally, the Baltic Sea's shallow bathymetry and complex coastal morphology increase the difficulty of accurately simulating sea-level dynamics, as they lead to strong local variability and require very high spatial resolution (Kowalewski & Kowalewska-Kalkowska, 2017). These aspects were specifically addressed in this dissertation in **Publications I–III**, which will be discussed in more details in Section 2.2.

Recently, data-driven methods have emerged as powerful complements. ML and neural networks now predict storm surges and extreme events, from Random Forest (RF) classifiers providing short-term warnings to DL ensembles like HIDRA2 (Rus et al., 2023), delivering 72-hour forecasts with high skill (Barzandeh et al., 2025; Bellinghausen et al., 2025). While these approaches improve efficiency and pre-warning capability, they can smooth high-frequency dynamics, struggle with extreme events outside the training data, and are best viewed as supplements rather than replacements for physics-based models (Barzandeh et al., 2025). Accurately modeling the Baltic Sea's complex coastal zones remains challenging due to shallow, highly intricate geometries, which constrain local forecast precision. Predicting extremes is especially critical for flood warning systems. While ML approaches have proven useful for classifying extreme events, they currently do not predict the timing or magnitude of these extremes (Bellinghausen et al., 2025). Section 2.2 provides a detailed discussion of the ML models developed for sea level forecasting and their relevance to our study.

[3] Nucleus for European Modelling of the Ocean

[4] High Resolution Operational Model for the Baltic Sea- Baltic Operational Oceanographic System

1.3 Sea Level Sources

1.3.1 Bias-Corrected Nemo-Nordic HDM Data

This dissertation employs multiple sea level data sources to develop ML models for forecasting and validating predictions. These include TGs, SA, and HDM outputs, each offering distinct spatial and temporal coverage, resolutions, and advantages (Figure 1). We will discuss these observational tools in detail, along with their technical specifications and data characteristics, in Section 2.

Hydrodynamic models are powerful computational tools used to simulate and predict ocean behaviour by solving the fundamental equations of fluid dynamics, primarily the Navier-Stokes equations (Jahanmard et al., 2021). These models generate valuable data on essential oceanographic variables such as SSH, Sea Surface Temperature (SST), Sea Surface Salinity (SSS), and current velocities. By integrating physical laws with initial and boundary conditions—such as atmospheric forcing, tidal inputs, and bathymetric data—HDMs provide a continuous, three-dimensional representation of ocean processes across space and time. One of their key advantages is the ability to offer consistent spatiotemporal coverage, filling gaps left by other observational methods like TGs and SA. HDMs deliver seamless data from shallow coastal waters to the open ocean. This makes them particularly useful for applications such as spatiotemporal sea level forecasting.

However, HDMs are not without challenges. A significant issue is the lack of a standardized vertical reference datum, which complicates comparisons with measurements from other sources (Slobbe et al., 2014). Additionally, HDMs rely heavily on the accuracy of their input data, including bathymetry, atmospheric forcing, and boundary conditions, all of which can introduce uncertainties (Jahanmard et al., 2021). Discretization errors due to finite spatial resolution and time-stepping approximations, along with imperfect parameterizations of sub-grid-scale processes like turbulence and mixing, can further lead to deviations from real-world conditions (Mardani et al., 2020).

In the Baltic Sea region, several HDMs are available, each designed to capture the unique characteristics of the area. Among these, the NEMO-Nordic model, developed by Swedish Meteorological and Hydrological Institute (SMHI), is a three-dimensional coupled ocean–sea ice model covering both the Baltic and North Seas (Hordoir et al., 2019). It provides hourly outputs of sea levels and other oceanographic variables at a horizontal resolution of one nautical mile, enabling detailed regional simulations and analyses. The bias-corrected model employed in this study was constructed using an extensive network of geoid-referred TGs (Jahanmard et al., 2022). This correction resulted in a substantial reduction in Root Mean Squared Error (RMSE) compared with the original NEMO-HDM sea level dataset. Validation was also performed against TGs, providing the most realistic ground truth, as well as SA data, which allows coverage of both coastal and offshore regions (e.g., average RMSE of 4 cm and correlation of 0.98 with TGs). Other models, like HIROMB-BOOS[5] (She et al., 2007), ROMS (Shchepetkin & McWilliams, 2005) and GETM[6] (Burchard & Bolding, 2002), are also used for regional forecasting and research. In this dissertation, we utilized the NEMO-Nordic model but with adjusted

[5] High-Resolution Operational Model for the Baltic and the Baltic Operational Oceanographic System

[6] General Estuarine Transport Model

vertical reference datum. This bias-corrected version, which involves the integration of Geoid, TG, and SA data is employed for **Publications I** and **II** and scrutinized in Jahanmard et al. (2022).

1.3.2 Satellite Altimetry-Based Observations

Satellite altimetry has served as a fundamental tool for ocean observation since 1973, when the SkyLab mission first demonstrated spaceborne radar altimeter capabilities. The technology has evolved significantly from early missions like GEOS-3 (1975–1978), Seasat (1978), Geosat (1985–1990), ERS-1 (1991–2000), Topex/Poseidon (1992–2005), and GFO (1998–2008) to modern innovations incorporating delay-Doppler Synthetic Aperture Radar (SAR) and interferometric SAR (InSAR) techniques, as implemented on CryoSat-2, Jason (1–3), Sentinel (3A,3B,6) and SWOT[7] satellites (Abdalla et al., 2021). The continuity of the reference altimetry record is being extended by the Sentinel-6 series, with Sentinel-6B already launched and the planned Sentinel-6C mission are expected to further ensure long-term high-precision sea level monitoring into the 2030s (Ablain et al., 2025). These advancements have expanded applications across glaciology, coastal studies, and physical oceanography, with SAR-mode now employed globally. The state-of-the-art SWOT mission, launched in December 2022 shows further breakthroughs through wide-swath altimetry technology.

The core measurement principle behind SA involves precisely timing a radar pulse's journey from satellite to sea surface and back. By knowing the signal velocity and satellite altitude with sufficient precision, SSH can be determined relative to a reference ellipsoid with centimetre-level accuracy. Achieving this precision requires applying numerous corrections to the raw measurements, as shown in Equation 1:

$$SSH = H_{orbit} - (R + iono + wtc + dtc + SSB + PT + SET), \quad (1)$$

where R represents the measured range, H_{orbit} the satellite altitude, and the correction terms account for ionospheric effects ($iono$), wet and dry tropospheric delays (wtc, dtc), sea state bias (SSB), pole tides (PT), and solid earth tides (SET). The resulting SSH value represents sea surface height (SSH) relative to the ellipsoid reference frame. It should be noted that Equation (1) does not include Dynamic Atmospheric Correction (DAC). In this study, DAC was intentionally not applied when comparing SA with hydrodynamic model (HDM) outputs. The reason is that the HDM explicitly incorporates atmospheric pressure forcing and wind-driven barotropic responses. Since DAC removes high-frequency atmospheric loading effects from altimetry observations, applying it would lead to inconsistency when comparing instantaneous satellite measurements with model outputs that already account for these processes. Given that SA provides near-instantaneous snapshots (e.g., ~27-day revisit for Sentinel-3 along a given track), maintaining consistency in the treatment of atmospheric effects is essential. Therefore, DAC was excluded to ensure that both observational and modeled sea-level fields represent comparable physical quantities, following the methodology established by Jahanmard et al. (2022).

Multiple organizations provide processed altimetry products, including AVISO, European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT[8]), and NASA Earth Data. For the Baltic Sea region, we utilized the specially tuned Baltic+SEAL dataset

[7] Surface Water Ocean Topography
[8] https://www.eumetsat.int/

(Passaro et al., 2021), which offers 0.5–1 cm improved accuracy over standard products through regional optimization (Mostafavi et al., 2021). These altimetry products proved invaluable for evaluating the developed ML predictions, providing independent, observation-based benchmarks for model performance evaluations. These datasets served as the primary validation source for ML results in **Publication I**. For **Publication II**, we employed EUMETSAT altimetry data instead, as the study period fell outside the Baltic+SEAL temporal coverage. In this dissertation, since the focus was on geoid-referenced sea levels (DT), we employed Equations 2 and 4 to compute DT from altimetry SSH.

The Sentinel-3A and Sentinel-3B satellites, forming the Copernicus program's altimetry constellation, provide complementary observations with near-identical payloads, enhancing the temporal and spatial coverage of SSH measurements. Both satellites carry Synthetic Aperture Radar (SAR) altimeters with a 300-meter along-track resolution, enabling high-resolution monitoring of coastal and open-ocean dynamics. Their coordinated orbits improve the temporal sampling of rapidly evolving ocean features, with Sentinel-3A (launched in 2016) initially demonstrating the capability to resolve mesoscale variability, while Sentinel-3B (2018) further enhanced observational density through reduced revisit times. In this dissertation, we utilized both the original high-rate (20 Hz) data and smoothed (1 Hz) data, applying a moving average filter with a 5% sampling rate to reduce noise. Such nadir altimeters have large footprints and remain noisy even after heavy pulse averaging. This limits their ability to resolve fine-scale SSH features. Consequently, SSH signals smaller than ~100 km are effectively unobservable.

A critical challenge in SA is land contamination near coastlines, where radar echoes are distorted due to mixed land-water signatures in the footprint. To mitigate this, we implemented an outlier detection method in **Publication II**, where we employed the Median Absolute Deviation (MAD) filter to exclude erroneous coastal measurements from Sentinel-3 A/B data.

1.3.3 Tide Gauge Measurements

Tide gauge observations have long served as the most reliable and continuous source of sea level measurements, providing invaluable records for coastal monitoring and climate studies. Installed along coastlines worldwide (cf. Figure 1), these instruments deliver high-frequency water level data essential for understanding both short-term variability and long-term trends. However, being land-based installations, TGs are inherently affected by vertical land movements (VLMs), which can range from several millimetres to several centimetres annually depending on regional geological factors such as GIA, tectonic activity, and anthropogenic subsidence. To account for these displacements, GIA models are typically applied to separate true sea level change from land motion effects.

An additional challenge in TG data usage stems from their reference to national vertical datums, which vary between countries and can complicate regional sea level comparisons. In the Baltic Sea region—the focus of this study—significant progress has been made to harmonize these discrepancies through the recent establishment of the Baltic Sea Chart Datum 2000 (BSCD2000), a unified geoid-based reference system (Liebsch et al., 2023). This standardization enables more consistent cross-border analysis of sea level trends and extremes. Table 1 describes the main characteristics of the TG stations used in this study.

In this research, TG data played a central role in model development. **Publication III** relied on TG observations as the primary input for training SLM forecasting methods, using RSL measurements to focus on extreme events.

Regarding the utilization of TG data, an additional complexity arises from differences in permanent tide systems across Baltic countries. The German height system follows the mean-tide concept, whereas Finnish, Swedish, Polish, and Estonian systems are referenced to the zero-tide concept (Jahanmard et al., 2021). These variations were considered in these analyses to ensure datum consistency when comparing multi-national TG data. We will discuss the differences between these systems in Section 2.3.

Table 1. Characteristics of the TG stations utilized in the dissertation.

TG station	Latitude	Longitude	Country	Datum
Narva-Jõesuu (Narva)	59.4691	28.0421	Estonia	EH2000
Ristna	58.9212	22.0552	Estonia	EH2000
Oulu	65.0403	25.4182	Finland	N2000
Kungsholmsfort	56.1053	15.5894	Sweden	RH2000
Władysławowo	54.7968	18.4187	Poland	PL-EVRF2007-NH
Greifswald	54.0928	13.446	Germany	DHHN92
Landsort Norra	58.7689	17.8589	Sweden	RH2000
Nynas Fiskehamn	58.9006	17.9536	Sweden	RH2000
Visby	57.6392	18.2844	Sweden	RH2000

1.3.4 Other Sources

In addition to TG, SA, and HDM, also GNSS and Airborne Laser Scanning (ALS) provide valuable complementary sea level–related data. GNSS stations co-located with TG measure vertical land motion, enabling separation of true sea level change from land uplift or subsidence and providing ellipsoidal height control within a global reference frame. ALS, on the other hand, delivers high-resolution coastal topography and nearshore elevation data, which are essential for accurate coastal flood modeling, shoreline change analysis, and linking marine water levels to terrestrial height systems. Together, GNSS and ALS strengthen vertical datum consistency and improve the integration of sea level observations across different reference surfaces. GNSS-based methods, including shipborne surveys and ALS, enhance SSH determination accuracy in near-coastal regions where conventional techniques are limited (Varbla, 2023).

1.4 Vertical Datum Representations of Sea Levels

For sea level studies, heights are measured relative to different reference surfaces, each serving distinct scientific and practical purposes. Hence it is necessary to exploit these terms cautionary (Ayinde et al., 2024). Mean Sea Level (MSL) is a traditional geodetic reference defined as the long-term average sea levels observed at TG stations, approximating the geoid and commonly used for coastal and engineering applications. The ellipsoid is a mathematically defined, smooth reference surface (e.g., WGS84) used in SA, where heights are expressed as ellipsoidal heights, representing purely geometric distances that are independent of Earth's gravity field.

While TGs refer to the national vertical datum, SA data refer to the reference ellipsoid. Sea Level Anomaly (SLA) defines the short-term deviations of SSH from a long-term mean, typically derived from SA to monitor variability such as tides, storm surges, and climate-related changes (cf. Figure 1).

Hydrodynamic model-based sea levels may operate with an unknown or implicit vertical reference datum (Slobbe et al., 2013), meaning its simulated water levels are relative to an internal model baseline rather than a clearly defined geodetic surface. This lack of a well-defined datum can complicate comparisons with observational data unless appropriate vertical transformations are applied.

Geoid, as the gravitational equipotential surface of the earth, effectively represents the Earth's "zero-height" level and coinciding with average sea level in the absence of currents, air pressure variations, and other perturbing phenomena. Geoid is being computed based on gravity data and can be approximated with higher resolution nowadays using the coefficients of spherical harmonic expansions and other approximate techniques. SA provides SSH relative to the reference ellipsoid. To obtain geoid-referenced heights, which are physically meaningful for oceanographic applications, the geoid undulation (N) must be subtracted from the ellipsoidal height. In practice, Baltic Sea models such as NKG2015 and BSCD2000 are quasigeoid models, meaning that the resulting heights correspond formally to normal heights. However, offshore, the difference between normal and orthometric heights is negligible. The conversion from ellipsoidal to geoid-referenced height is given by:

$$H = h - N, \tag{2}$$

where h is the ellipsoidal height and H is the geoid-referenced height, and N is the geoid (or quasigeoid) undulation at the same location. Figure 3 displays the pattern for the geoid undulations for the latest Baltic geoid model (BSCD2000).

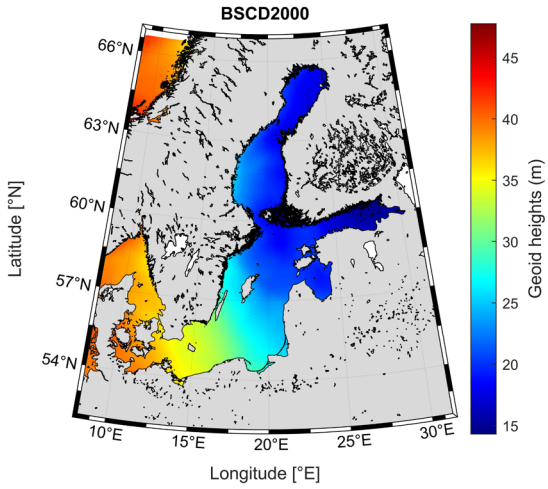


*Figure 3. The BSCD2000 geoid model (Utilized as a vertical reference datum in **Publications III**).*

The geoid offers several advantages over traditional mean sea level (MSL)-based reference systems. It provides a globally consistent vertical reference, avoids the spatial variability and temporal instability of the actual sea surface, and is directly compatible

with GNSS-based height determination, enabling straightforward transformation between ellipsoidal heights and physically meaningful orthometric heights. In this study, we utilized two main geoid models that were developed for the Baltic Sea region. For the **Publications I** and **II**, we utilized the NKG2015 geoid model (Ågren et al., 2016), optimized for the Baltic–Nordic region and offering 1–2 cm accuracy over land in gravimetrically well studied regions. Originally defined in a zero-tide system, it was converted to the mean-tide system (explained the conversions in below). The NKG2015 geoid model was used in **Publications I** and **II** to convert SSH values from reference ellipsoid to the geoid surface. BSCD2000 (Liebsch et al., 2023) is also the latest version of Baltic geoid model that showed better accuracy compared to NKG2015. The Baltic Sea Hydrographic Commission (BSHC) has recommended the BSCD2000 as a new common chart for hydrographic surveying, hydrographic engineering, nautical charts, and publications in the Baltic Sea. This geoid model was utilized for **Publication III** for consistent comparison of the measurements.

Dynamic topography (DT), also commonly referred to as absolute dynamic topography (ADT), represents the difference between the instantaneous sea surface and the geoid. It describes spatial variations in SSH associated with ocean circulation. These variations are primarily driven by ocean currents, wind forcing, and thermohaline processes, rather than tidal motions, which are typically removed during altimetry processing. Studying DT is essential for understanding ocean circulation, heat transport, climate variability, and large-scale ocean dynamics.

If we apply the Equation 2 for the sea surface, we can express the DT as below:

$$DT_{(\varphi,\lambda)} = SSH_{(\varphi,\lambda)} - N, \tag{3}$$

where $SSH_{(\varphi,\lambda)}$ is the sea surface height at geographical point with coordinates (φ, λ) (e.g., measured by satellite altimeter) referenced to ellipsoid, $DT_{(\varphi,\lambda)}$ is the geoid-referenced sea level (DT) and N is the geoid undulation at the same location.

Another important aspect when integrating different sea-level datasets is the treatment of permanent tide systems. Permanent tides refer to the long-term deformation of the Earth caused by the static component of the lunar and solar gravitational potential (Mäkinen, 2001). To ensure consistency in height and gravity measurements, three permanent tide concepts are commonly defined: tide-free, zero-tide, and mean-tide (Mäkinen, 2021). In the tide-free system, all permanent tidal effects are removed from both the Earth's crust and gravity field. In the mean-tide system, permanent tidal deformation of the Earth is retained. The zero-tide system removes permanent crust deformation while retaining indirect tidal effects in the gravity field. Modern geoid models are typically expressed in the zero-tide system. Therefore, when combining datasets referenced to different permanent tide systems, appropriate transformations are required to ensure consistency. For example, geoid models such as NKG2015 are provided in the zero-tide system and must be converted when necessary using established correction formulas (Ekman 1989; Varbla et al., 2022):

$$\begin{aligned} H_{mean-tide} = H_{zero-tide} &+ 0.29541(sin\,\varphi_B{}^2 - sin\,\varphi_{NAP}{}^2) \\ &+ 0.00042(sin\,\varphi_B{}^4 - sin\,\varphi_{NAP}{}^4), \end{aligned} \tag{4}$$

where φ_B and φ_{NAP} refer to the geodetic latitude of the desired benchmark and Normaal Amsterdams Peil (NAP) (52°22'53'' N), respectively. In the Baltic Sea region,

TG data are provided in different permanent tide systems: German, Finnish, Estonian, and Polish TGs use the zero-tide system, Swedish TGs use the mean-tide system, and Danish TGs use the tide-free system. Since Danish TGs were not used in this dissertation, the corresponding conversion to the mean-tide system is not shown here; readers are referred to Varbla et al. (2022) for details on that conversion. It should also be noted that the SA observations were originally provided in the mean-tide system.

1.5 Study Area

The Baltic Sea, a semi-enclosed basin in Northern Europe spanning 53°N–66°N and 13°E–30°E, covering approximately 377,000 km². It is relatively shallow, with an average depth of 55 meters, though the Landsort Deep in the Gotland Basin reaches 459 meters. The sea comprises several sub-basins, including the Gulf of Finland, Gulf of Riga, Bothnian Bay, Bothnian Sea, Gotland Basin, Bornholm Basin, and Arkona Basin, each with distinct stratification and circulation patterns (cf. Figure 4). The Baltic Sea is micro-tidal, with tidal amplitudes generally below 5 cm due to limited connection to the open ocean and hydraulic restrictions at the Belts and Øresund (Wulff et al., 2001; Kulikov et al., 2015). As a result, most sea level variability is driven by meteorological forces, particularly storm surges caused by wind, wave setup, and atmospheric pressure gradients. In some areas, such as the eastern Gulf of Finland, local resonance can amplify tides to about 20 cm (Särkkä et al., 2023), establishing a clear hierarchy of forcing mechanisms across timescales (Weisse & Hünicke, 2019).

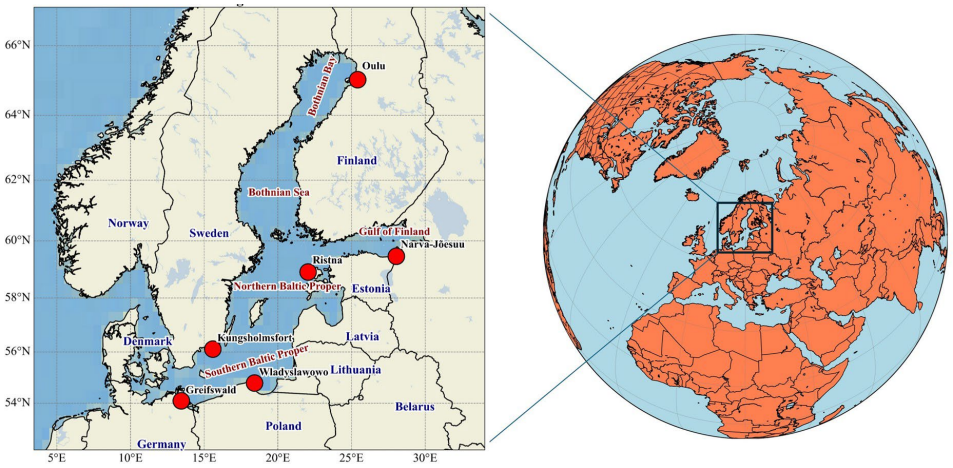


*Figure 4. Baltic Sea is the study area (Modified from **Publication III**), located in the northern Europe. For the **publication I**, the study area is the whole Baltic Sea; for the **Publication II** the study area is the eastern Baltic Sea, Gulf of Finland, and for the **Publication III**, the red markers are the studied TG stations.*

Sea level variations in the Baltic Sea occur across seconds to millennia and are amplified in different sub-basins, complicating interpretation (Weisse & Hünicke, 2019; Weisse et al., 2021). Short-term fluctuations (hours to weeks) are mainly driven by meteorological forcing, including wave and wind setup, atmospheric pressure changes, and seiches, while low-pressure systems can elevate sea levels by 10–30 cm via the inverse barometer effect (Medvedev et al., 2021). Regional wind patterns, particularly

westerly components linked to high NAO phases, are the dominant drivers of elevated sea levels and extremes, unlike the tidally dominated North Sea (Hieronymus et al., 2017). On monthly to seasonal scales, external forcing from the North Sea and freshwater input dominates, while internal atmospheric forcing drives sub-monthly variability and excites natural seiche modes (Samuelsson & Stigebrandt, 1996).

Long-term changes are influenced by steric effects, GIA, and climate-driven shifts in precipitation, river runoff, and atmospheric circulation like NAO. The Baltic region continues to adjust following the last ice age, with uplift rates of ~10 mm/year in the north and subsidence in the south (Weisse et al., 2021). Centennial tide gauge records indicate a statistically significant increase in the annual sea level amplitude, likely driven by changes in seasonal precipitation (Hünicke & Zorita, 2007).

SA (1995–2019) shows a significant rise in absolute geocentric sea level throughout the Baltic, with a gradient exceeding 3 mm/yr from southwest to northeast, causing northern basins like the Bay of Bothnia to rise faster than southwestern areas, partly due to wind forcing associated with NAO phases (Passaro et al., 2021). This absolute rise, combined with VLM, produces highly heterogeneous RSL changes, critical for assessing coastal impacts (Weisse et al., 2021).

The Baltic Sea is brackish, with a permanent halocline separating fresher surface waters from saltier deep waters (Lehmann et al., 2022). Its oceanography is shaped by river runoff, limited North Sea exchange, and seasonal ice cover in northern basins. Coastal upwelling in summer–autumn elevates cooler, nutrient-rich bottom waters, while most basins exhibit stratified, multi-layer structures (Delpeche-Ellmann et al., 2017, 2021). These features make the Baltic Sea highly dynamic and sensitive, providing an ideal environment for studying sea level variability.

Long-term observational infrastructure includes some of the world's oldest tide gauge records, exceeding 200 years at several stations (Weisse et al., 2021). Analyses along the Finnish coast provide baselines for assessing century-scale changes in sea level statistics and extremes (Johansson et al., 2001). Vertical referencing is standardized using the Baltic Sea Chart Datum and geoid models such as NKG2015 and BSCD2000, enabling consistent sea level measurement. Coastal altimetry is further supported by specialized algorithms like Baltic+SEAL (Passaro et al., 2021).

Wind is a major driver of Baltic sea level variability. Passaro et al. (2015) demonstrated that wind stress dominates the annual sea level cycle across the basin, while steric effects account for spatial differences both within and between sub-basins. Prevailing westerly and southwesterly winds average 5–10 m/s, with storms exceeding 25 m/s, particularly in autumn and winter. Strong winds associated with low-pressure systems can cause rapid pressure drops of 20–30 hPa in 24 hours, producing storm surges of 1–2 meters on southern coasts. The Kattegat and Danish Straits can funnel winds, amplifying surge effects. Future wind pattern changes are uncertain, with some projections indicating increased winter storm surges in the eastern Gulf of Finland and Gulf of Riga (Meier et al., 2004). Wave heights are generally 0.5–2 m but can exceed 8 m during severe storms in the central and northern basins, influencing coastal erosion, sediment transport, and mixing (Soomere, 2003). Seasonal ice cover suppresses wave activity, but climate change is reducing ice, extending the wave season and exposing coasts to winter storm energy.

Extreme sea levels have been extensively investigated due to their growing societal relevance. Projections indicate that by 2100, ESL could average 4.2 m globally and reach

9–10 m in some regions, primarily driven by long-term sea level rise rather than changes in storm intensity, thereby making rare events increasingly frequent (Jevrejeva et al., 2023).

The Baltic Sea is particularly susceptible to ESLs, with historical surges exceeding three meters include the 1872 Lübeck Bay flood and the 2005 Gulf of Finland event (Hofstede & Hamann, 2022). ESLs result from storm surges, wind waves, pre-existing high-water levels, resonance, and seiches. Analyses of up to 160 years of Estonian tide gauge data, corrected for land uplift, indicate RSL rises of 1.5–2.7 mm/yr and faster increases in extremes (3.5–11.2 mm/yr), driven by combinations of storm surges, seiches, and rare meteorological tsunamis (Suursaar & Sooäär, 2007). Future ESL risk is spatially heterogeneous: northern regions may continue to experience relative fall due to GIA, while increasing absolute sea level and changing storminess will dominate elsewhere (Meier et al., 2004; Weisse et al., 2021). Upwelling, most frequent along eastern and southern coasts, displaces surface waters, revealing colder, nutrient-rich depths and causing SST drops of 5–10°C (Lehmann & Myrberg, 2008).

Thermodynamic processes, including thermal expansion from SST gradients and salinity variations driven by freshwater input, influence sea levels regionally. Freshwater fluxes from precipitation and river runoff further modulate water mass and salinity, creating decimeter-scale differences, while glacier melting and ice sheet loss contribute mass and gravitational effects. Large-scale climate oscillations, such as ENSO and NAO modulate sea level through teleconnections, producing coherent anomalies across the basin.

Given its complex oceanography, long-term observational records, and well-defined geodetic infrastructure, the Baltic Sea is an excellent testbed for studying sea level dynamics, coastal processes, and climate-related changes in semi-enclosed basins. Its multi-country coastline, spanning nine nations, underscores the importance of regional collaboration for consistent and accurate sea level monitoring.

To capture the complex and multiscale sea-level dynamics of the Baltic Sea, this study integrates multiple datasets, including hydrodynamic model outputs, SA, TG observations, meteorological reanalysis data, and large-scale climate indices. Ensuring vertical consistency across these diverse sources through geoid referencing is a key component of the proposed framework. Table 2 summarizes the datasets and variables employed in this dissertation alongside their application and sources.

Table 2. Overview of datasets used in this thesis.

Dataset Category	Main Variables Used	Data Sources	Role in Study
Hydrodynamic Model Outputs	dynamic topography (DT), sea surface temperature (SST), sea surface salinity (SSS), wind forcing	NEMO Nordic, MUR, CCMP	Primary predictors for spatiotemporal sea-level forecasting
Tide Gauge Observations	Relative sea level (RSL)	Baltic national tide-gauge networks	Training and evaluation of sea-level maxima forecasting
Satellite Altimetry (SA)	Sea surface height (SSH)	Sentinel-3 missions (Baltic+SEAL+Eumetsat)	Independent validation of model forecasts
Meteorological features	Atmospheric pressure, precipitation, evaporation, River runoff	ERA5 reanalysis	Atmospheric forcing predictors
Wave Data	Significant wave height (SWH)	WAM / SWAN wave models	Coastal extreme-event drivers
Large-Scale Climate Indices	Baltic Sea Index (BSI)	Climate data archives	Representation of basin-scale atmospheric forcing for extreme sea levels
Vertical Reference Models	Geoid undulation	NKG2015 and BSCD2000 geoid models	Ensuring vertical consistency across datasets

2 Machine Learning Frameworks for Sea Level Forecasting

2.1 Foundations of ML models

Machine Learning is a subfield of artificial intelligence (AI) that focuses on creating systems that learn patterns from data and make predictions or decisions without being explicitly programmed (Figure 5). It is useful for tasks that involve repetitive operations, situations where rules are hard for humans to solve, or problems that would be time-consuming to solve manually. They are inspired by how the human brain processes information.

The main components of an ML system include data, features, models, and algorithms. Data form the foundation of any ML approach and may consist of structured or unstructured observations used for training and evaluation. From data, relevant features are extracted or engineered to represent the most informative characteristics of the problem. A model (e.g., linear regression, decision trees, or neural networks) defines the mathematical relationship between input features and the desired output. Then, the model's performance is quantified using a loss function, which measures the difference between predicted (or forecast) and true values. An optimizer (such as gradient descent or its variants which will be discussed below) iteratively adjusts the model parameters to minimize the loss function, enabling the system to learn from data and improve its predictive capability.

Typical ML models include linear and logistic regression, decision trees, RF, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and gradient boosting models such as Extreme Gradient Boosting (XGB), Light Gradient-Boosting Machine (LightGBM), and CatBoost. In most ML workflows, a separate feature selection or feature engineering step is needed before training the model because the algorithms do not automatically extract all relevant patterns from raw data.

DL is a specialized branch of ML (cf. Figure 5) that uses artificial neural networks (ANN) with many interconnected layers and designed to handle complex, high-dimensional, and nonlinear problems such as image recognition, audio processing, natural language understanding, and large unstructured datasets. Since DL models consist of many layers and neurons, they can automatically learn useful features from raw data (also called feature extraction), which reduces or eliminates the need for separate feature selection step. These intermediate layers are responsible for getting the best representations from input features but simply using them still make the problem linear, so activation functions are being exploited for adding nonlinearity term.

In a typical ML/DL workflow, data is split into training, validation, and test sets—commonly—70/15/15, though ratios may vary depending on dataset size—for robust model development and unbiased evaluation. The training set is used to estimate the model parameters, validation set tunes hyperparameters, and test set assesses final generalization performance.

Feature selection identifies the most influential predictors, keeping only variables that provide meaningful, non-redundant information. This improves accuracy, reduces overfitting, and lowers computational cost, while unnecessary or collinear features can harm generalization. Feature importance adds explainability and can be measured statistically (e.g., correlation), via model-based scores (e.g., tree-based algorithms), or with model-agnostic tools like permutation importance and SHAP. Choosing key, non-collinear features make models more robust, interpretable, and generalizable.

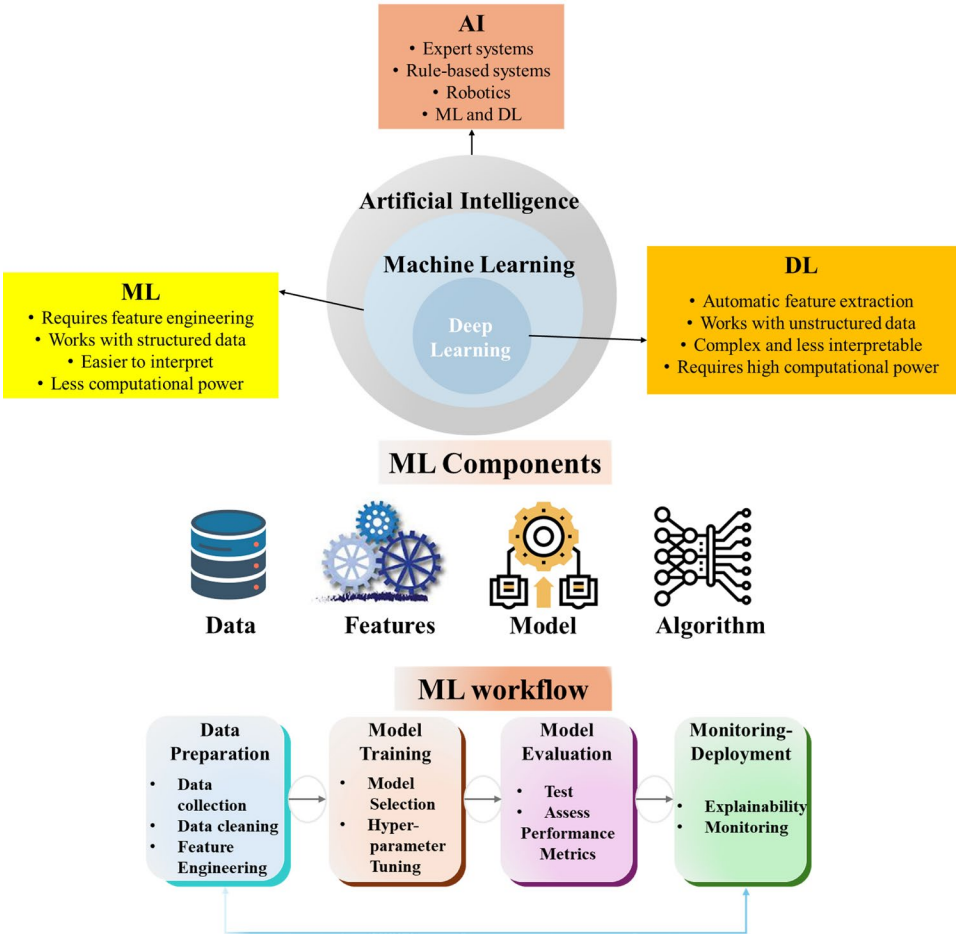


Figure 5. Comparison between ML and DL, along with the main componensts and the general workflow.

Common DL model architecture types include feedforward neural networks, CNNs for image data, and recurrent neural networks (RNNs) for sequential data, as shown in Figure 6. There are also other network types like transformers, useful for language and multimodal tasks, generative models for super-resolution tasks and autoencoders for unsupervised representation learning.

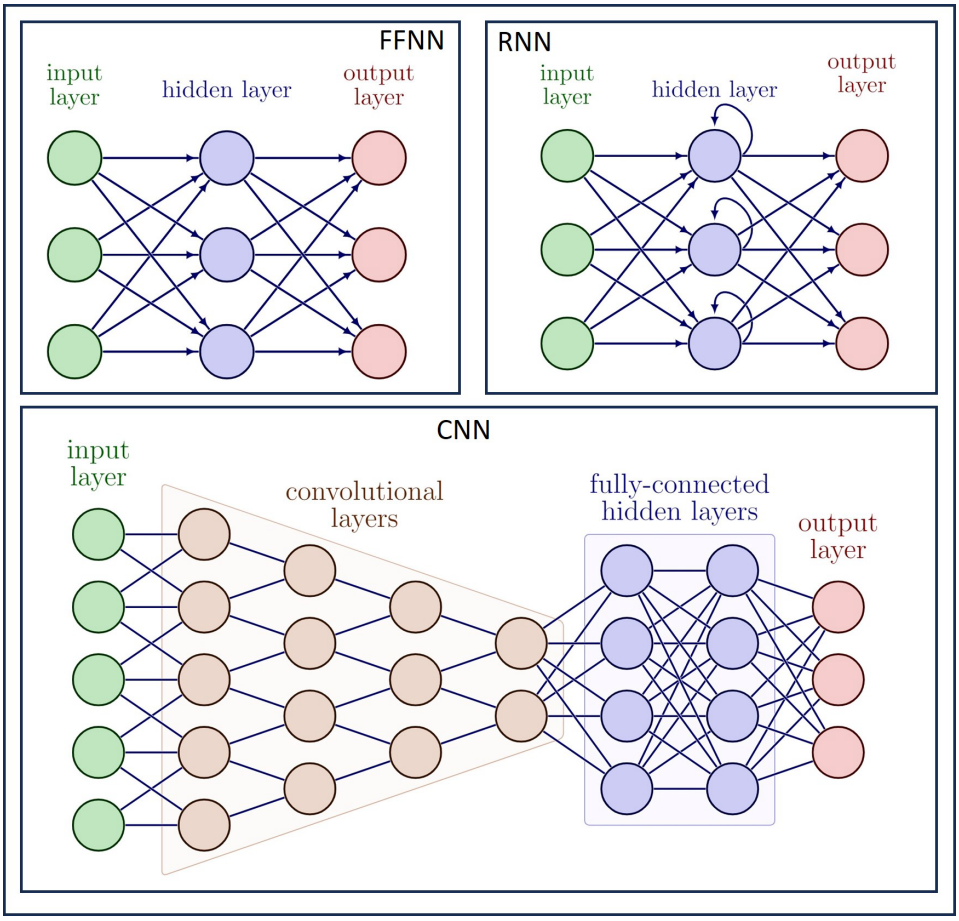


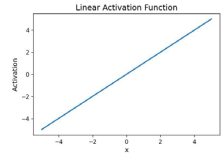
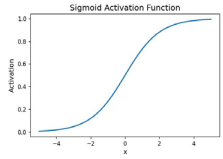
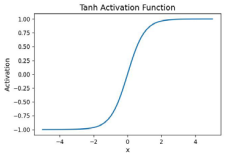
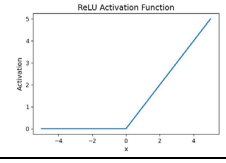
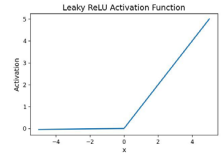
*Figure 6. Different DL models architectures (from **Publication II**).*

The main differences between ML and DL come from data needs, complexity, and the degree of automation. ML models usually require manual feature engineering, work well with smaller datasets, train faster, and are easier to interpret. DL models (while recognized as a specialized ML models) typically require large amounts of data, are computationally expensive to train, rely on Graphics Processing Units (GPUs) or other specialized hardware, and are harder to interpret because of their large number of parameters. ML is commonly used for structured data with simpler patterns, while DL is favored for images, videos, speech, natural language, and other complex tasks. In below, we describe the main DL models utilized for the task of spatiotemporal sea level forecasting (**Publications I and II**). It is also important to note that some of these DL models have been introduced before 2000 but due to the hardware change, parallel computing capabilities and increased datasets, they got more attention in recent years. In the subsections below, we discuss the main aspects of ML models in more detail.

2.1.1 Activation Functions

Activation functions are needed in ML, especially in neural networks, to introduce non-linearity, enabling the model to learn and represent complex patterns beyond simple linear relationships. There are different types of activation functions. Table 3 displays different activation functions with the related equations and representations. The Rectified Linear Unit (ReLU) activation function is one of the most widely used because it offers good convergence properties, is computationally efficient, and reduces the problem that often occurs with sigmoid or Hyperbolic Tangent (Tanh) functions. Sigmoid and Tanh tend to be saturated for large positive or negative inputs, which causes gradients to become very small and slows down weight updating process.

Table 3. Different activation functions commonly used in ML designs.

Activation Function	Equation	Output Range	Shape
Linear (Identity)	$f(x) = x$	*(−∞, ∞)*	
Sigmoid	$\sigma(x) = \frac{1}{1+e^{-x}}$	*(0, 1)*	
Tanh	$tanh\,(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	*(−1, 1)*	
ReLU	$f(x) = max\,(0, x)$	*[0, ∞)*	
Leaky ReLU	$f(x) = max\,(\alpha x, x)$, typically $\alpha = 0.01$	*(−∞, ∞)*	

Despite its simplicity and efficiency, the ReLU activation function may also face challenges such as dying neurons and slower convergence in certain scenarios. To address ReLU's drawback of producing zero gradients for negative inputs, several improved variants have been developed, such as Leaky ReLU, Parametric ReLU, and exponential linear unit (ELU), which maintain small but non-zero gradients in the negative region and help the model learn more effectively.

The selection of activation functions in this study was guided by both theoretical considerations and empirical validation. In deep neural networks, activation functions should preserve gradient flow and enable efficient optimization. ReLU was adopted in hidden layers of the CNN models due to its computational efficiency and ability to mitigate vanishing-gradient issues. Since the target variable (e.g., DT) can take both positive and negative values, a linear activation function was used in the output layer. For recurrent architectures such as LSTM and GRU, the primary activation functions are embedded within the model structure (sigmoid gates and tanh state functions), ensuring stable learning of temporal dependencies. Final selection was confirmed through validation performance rather than purely heuristic choice.

2.1.2 Optimizers

In DL models, the main task is to learn a function that maps inputs to desired outputs. For this purpose, an objective function or loss function which measures the difference between the model's predictions and the true values is defined, and we try to minimize it. To achieve this, the model's parameters are adjusted iteratively using gradient-based optimization, so that the loss is reduced and the model's predictions become more accurate. There are two general ways of computing gradients: numerical gradients and analytical gradients. Numerical gradients approximate the derivative using small finite differences. They are simple to implement but slow and less accurate. Analytical gradients, on the other hand, are computed directly from the derivative mathematical formulas of the model. They are faster, more precise, and are the standard approach used in the DL frameworks.

As mentioned earlier; to optimize any DL model, a loss function should be defined to minimize the error term, and this function must be differentiable so that we can compute its gradient and update the model's weights. For gradient computations and model weights updating, different optimization algorithms use gradient information in multiple ways. Basic gradient descent computes the gradient using the entire training dataset, which is accurate but slow and often impractical for large datasets. Stochastic gradient descent (SGD) computes the gradient using only a single training example at each step. It is fast but can be noisy and may get stuck in local minimum. Classic SGD may struggle when the loss surface has directions that require faster movement and other directions that require slower movement. To address this, momentum was introduced. Momentum accelerates updates in consistent directions and dampens oscillations. Nesterov momentum is a later improvement that looks ahead before computing the gradient, often providing more stable convergence.

More advanced optimizers like Adaptive optimizers were also developed to adjust the learning rate for each parameter. The Adagrad optimizer introduced the idea of adapting the learning rate based on the accumulated squared gradients. When gradients change a lot, the effective step size decreases; when gradients are small, the step size increases. However, Adagrad's accumulated squared gradient term can grow too large, causing the learning rate to become overly small. Root Mean Square Propagation (RMSProp), introduced in 2012, solves this limitation by using an exponential moving average of squared gradients instead of the cumulative sum. This prevents the effective learning rate from shrinking too much over time. Adaptive Moment Estimation (Adam), introduced in 2015, combines the benefits of RMSProp and momentum by keeping moving averages of both the gradients and the squared gradients, making it one of the most widely used optimizers in DL field. All of these optimizers include a hyperparameter called the learning

rate, which controls how large each update step is. The learning rate can be fixed or scheduled to change during training. Other hyperparameters include factors, minimum delta, patience, and minimum learning rate.

The choice of optimizer in this research was also based on convergence stability, robustness to noisy gradients, and computational efficiency. For geophysical regression tasks involving high-dimensional inputs and time-series data, adaptive optimizers generally provide more stable convergence than classical stochastic gradient descent. Adam was selected here because it combines adaptive learning rates with momentum, enabling efficient and reliable training across different architectures. Although alternative optimizers were explored during preliminary testing, Adam consistently demonstrated stable convergence and strong validation performance.

2.1.3 Overfitting and Generalization

Overfitting happens when a model learns patterns that are too specific to the training data and fails to generalize to new unseen test data. DL models often have high variance and low bias, while traditional ML models tend to have lower variance and higher bias. In practice, the goal is to achieve both low bias and low variance. Regularization is also important component in DL model designs and are applied to prevent overfitting. They are necessary when training accuracy is high, but test performance is low. This is different from optimization, which focuses on adjusting model parameters to minimize the loss during training, with the regularization focusing on improving the generalization.

There are different ways to improve generalization. For example, reducing the number of layers or neurons is a basic approach, but doing so excessively may lead to underfitting. More robust regularization methods include L1 and L2 weight penalties, dropout, early stopping, batch normalization, and data augmentation. These techniques limit model complexity, prevent reliance on specific features, stabilize learning dynamics, or increase the diversity of training samples. Choosing the appropriate method depends on the architecture and the nature of the problem. Proper dataset splitting is also essential. In literature, usually 70–30 or 80–20 training-test ratio is recommended (Ayinde et al., 2024). The training set is used to estimate model parameters, the validation set is used for hyperparameter tuning and model selection, and the test set is used only at the end to provide an unbiased evaluation of the final model's generalization performance.

2.2 Review of Current Machine Learning Approaches for Sea Level Forecasting

The problem of sea level forecasting is regarded as a multivariate forecasting problem. This means that along with the previous values of sea levels, there are also other external factors impacting the sea level variations. These factors include tides, zonal and meridional wind speed (u- and v-wind), surface atmospheric pressure (SLP), significant wave height (SWH), SST, SSS, among others. The physical processes governing sea-level variability are highly nonlinear and interconnected, so the inclusion of these drivers is essential for improving forecast accuracy. A more detailed discussion behind these external features is provided in Section 1.3. Given the impacts of external features and the main purpose of spatiotemporal sea level (SL) forecasting (forecasting at different grid points), we can formulize the problem as below:

$$SL_{(\varphi,\lambda,t+1:t+k)} = F(SL_{(\varphi,\lambda,t-\delta:t)} + influential\ features_{(\varphi,\lambda,t-\delta:t)}), \quad (5)$$

where, *F* represents the function learned by ML algorithm to map past observations and external inputs to future sea-level values. The parameter *K* denotes the forecast horizon: when *K*=1, the problem will be the single step forecasting (**Publications I** and **III**), otherwise it shows the forecast horizons, and the problem will be a multi-step ahead forecasting (**Publication II**). Parameter δ shows the lag window (also called look-back parameter), which is a predefined in time series forecasting, showing the number of previous timesteps that the model will look to forecast the next horizons.

ML-based model development starts with the problem definition, where the objective is clearly outlined, whether it is classification, regression, clustering, or another task, along with defining key performance metrics such as R^2, RMSE, Mean Squared Error (MSE), Mean Absolute Error (MAE) and correlation coefficient for a regression task or precision, accuracy, confusion matrix, recall, and F1-score for a classification task. The next step involves related input data collection and integration. Since data often comes from various spatiotemporal sources with different formats and structures, a resampling is usually needed to make all data types consistent. Once this is done, data preprocessing is performed, including outlier detection and removal, normalization or standardization to bring features to a similar scale, and missing data imputation using techniques like mean/median substitution or advanced imputation methods.

Next step is the feature engineering and selection, where the most influential set of input combinations is identified, often guided by domain knowledge, further supported by exploratory data analysis (EDA) techniques such as correlation analysis, mutual information (MI) values, principal component analysis (PCA) or any other feature selection techniques. In a forecasting problem, an additional step is to select the number of previous timesteps (window of lagged features) and number of future timesteps (also called horizon) to be forecast.

After setting these up, the next step is selecting the right model types for each problem, either it is a regression or classification, forecasting or prediction, the complexity of the problem and also computational resources available. These models range from traditional ML models such as ANNs, SVMs, Gradient boosting machines (GBMs), decision trees, or ensemble methods to more advanced DL architecture like CNNs, RNNs, hybrid structures and generative models depending on the data characteristics and task requirements. The model is then designed for training, which involves splitting the data into training, validation, and test sets, and selecting the most appropriate hyperparameters (using an optimizations technique e.g., trial and error, grid search, random search, genetic algorithm, particle swarm optimization and BO). Consequently, a training process will be done on the training set by minimizing a loss function (e.g., MSE) by iteratively updating the model weights during training using an optimizer (e.g., Adam, SGD, and RMSprop).

Once trained, the model undergoes testing using a different set of data, which was not used during the model training to assess its performance and generalization capabilities. Different models may be compared based on metrics such as accuracy, generalization ability, computational efficiency, and interpretability. Finally, the model with the best performance in terms of both accuracy and generalization is selected as the best-performing approach. Depending on the application, further steps may include sensitivity analysis or model explainability to find the most influential inputs, external validation, deployment, and continuous monitoring.

Machine learning, a subfield of AI, enables computers to learn from data and perform specific tasks without explicit programming. In oceanography, the growing adoption of ML stems from the convergence of three critical factors: advanced computing power, sophisticated algorithms, and the availability of large-scale ocean data (Qin et al., 2023). Before ML-based approaches became prevalent, sea level prediction primarily relied on numerical modelling techniques that solve fluid dynamics equations, such as ADCIRC[9] and FVCOM[10]. These physics-based models remain effective but face several challenges: they are computationally intensive, contain inherent uncertainties, and require precise boundary conditions (Qin et al., 2023).

Machine learning-based sea level forecasting has emerged as a powerful alternative for real-time sea level forecasting, offering significant advantages like effectively handling nonlinear relationships between variables and identifying underlying patterns with more limited data. The development of ML applications for sea level studies has progressed through multiple stages. Initial efforts employed simple linear regression models (Sertel et al., 2008), followed by more sophisticated techniques. ANNs became widely used (Ghorbani et al., 2010), along with other algorithms including RF (Ayinde et al., 2023; Passaro and Juhl, 2023; Bellinghausen et al., 2025), SVM (Sithara et al., 2020; Balogun and Adebisi, 2021; Hazrin et al., 2023), Gradient Boosting Machines (Ayinde et al., 2023), Adaptive Neuro-Fuzzy Inference Systems (ANFIS) (Karimi et al., 2013; Wang et al., 2020) and XGB (Nguyen et al., 2021), demonstrating particular success in modelling sea level dynamics.

Comparative studies have also yielded important insights about model performance. Imani et al. (2018) successfully applied Support Vector Regression (SVR) for daily sea level predictions along Taiwan's Chiayi coast, while Khaledian et al. (2020) found SVR outperforming ANNs in simulating Caspian Sea levels. Hazrin et al. (2023) conducted an extensive comparison of multiple models, including Linear Regression, SVM, Ensemble Regression, Regression Trees, ANN, and Gaussian Process Regression for daily sea level forecasting, identifying optimal performance with a 7-day lag window across four TG stations.

The superiority of multivariate approaches over univariate methods has been well documented (Balogun and Adebisi, 2021; Tur et al., 2021). Notably, Tur et al. (2021) demonstrated that incorporating meteorological features such as wind speed and pressure could improve short-term sea level forecasting accuracy by 33% in Antalya Harbor, Turkey, with ANFIS consistently outperforming Multiple Linear Regression in both univariate and multivariate scenarios. Ayinde et al. (2023) applied different supervised ML models including neural networks e.g., LSTM, Multilayer Perceptron (MLP), along with MLR, RF and GBM for mean sea level anomaly (SLA) prediction where different input features in the Gulf of Guinea. Nguyen et al. (2021) also proposed hybrid evolutionary algorithms coupled with XGB method to forecast hourly sea levels in Jungrang urban basin, South Korea and compared the model performance with classification and registration tree (CART) and RF models and genetic algorithm coupled XGB for 1–5 hours ahead forecasting.

Among the conventional ML models utilized for sea level application, ANNs have been extensively employed in the previous studies (Altunkaynak, 2007; Pashova & Popova, 2011; Filippo et al., 2012). Also, MLP, a specific type of ANN architecture, has proven

[9] ADvanced CIRCulation model

[10] Finite-Volume Coastal Ocean Model

particularly valuable in numerous sea level studies (Sztobryn, 2003; Ghorbani et al., 2018; Guillou and Chapalain, 2021; Ayinde et al., 2023; Raj et al., 2023).

Recent developments in DL have significantly advanced sea level modelling capabilities for both forecasting and prediction through specialized neural network architectures designed to handle sequential and spatial data patterns. These DL approaches primarily include CNNs, which excel at processing grid-structured data, and recurrent-based neural networks such as LSTMs, and GRUs that are particularly effective for time series analysis. A notable application by Ishida et al. (2020) demonstrated the effectiveness of LSTM models for hourly sea level forecasting at Osaka's TG station in Japan. Their model incorporated multiple meteorological inputs including wind speed, mean sea level pressure (MSLP), air temperature, annual global mean air, and solar position data, achieving strong performance in forecasting sea levels at both hourly and monthly timescales, though it tended to underestimate peak values. LSTM model has also been applied in other studies (Liu et al., 2020; Accarino et al., 2021; Balogun and Adebisi, 2021; Ayinde et al., 2023). Raj et al. (2022) applied empirical mode decomposition coupled with the CNN and GRU models to predict sea level trends in three TG stations in south pacific region using 53 predictor variables from MODIS[11]-Terra, GLDAS[12] 2.0 and MERRA[13]-2 models.

While these studies have demonstrated the effectiveness of DL methods over traditional ML models, most have focused on individual or limited numbers of TG stations, leaving their performance for basin-wide or regional forecasting largely unassessed. This limitation has prompted the development of comprehensive spatiotemporal forecasting frameworks that enhance the practical utility of ML-based approaches for operational forecasting systems. The advancement of spatiotemporal forecasting has been made possible through convolutional neural networks (CNNs) and hybrid DL approaches, such as CNN-LSTM models, which benefit from capturing both spatial and temporal dependencies. These architectures have demonstrated superior performance in handling high-dimensional and nonlinear sea level dynamics that vary across both space and time.

A landmark study was conducted by Tiggeloven et al. (2021) who applied four distinct models—CNN, LSTM, ANN, and Convolutional LSTM (ConvLSTM)—to predict sea levels across a global network of 1,276 TG stations. Their results showed the superior performance of the LSTM and CNN models when using predictors including mean sea level pressure (MSLP), hourly MSLP gradients (ΔMSLP), 10-meter meridional and zonal wind components (U and V), and wind speed magnitude. Braakmann-Folgmann et al. (2017) demonstrated the effectiveness of CNN-ConvLSTM architectures in predicting sea level anomalies (SLA) using 23 years of merged SA data (1993–2015) in the northern and central Pacific Ocean, highlighting the model's ability to capture large-scale oceanic variability.

Recent advances in transformer-based architectures have further improved spatiotemporal forecasting. Fang et al. (2024) introduced a two-stage spatiotemporal transformer model for SLA prediction, leveraging self-attention mechanisms to better capture long-range dependencies in both spatial and temporal dimensions. Similarly, Shahabi and Tahvildari (2024) developed a CNN-LSTM hybrid model to predict coastal water levels across a wide geographical range using wind and tidal data as inputs,

[11] Moderate Resolution Imaging Spectroradiometer

[12] Global Land Data Assimilation System

[13] Modern Era Retrospective-Analysis for Research and Applications

demonstrating robust performance in capturing localized and regional sea level fluctuations. For regional applications, Wang et al. (2022) proposed HMnet3, a hybrid multivariate deep neural network designed for multistep-ahead SLA forecasting in the South China Sea. Their model incorporated sea surface temperature anomalies (SSTA), wind speed anomalies, and historical SLA data, achieving high accuracy in capturing complex ocean-atmosphere interactions. Zhang et al. (2024) utilized a Multi-head Attention Residual-Unet (ResUnet) architecture for daily sea surface height anomaly (SSHA) field forecasts in the North Atlantic Ocean. Their model, trained on gridded multi-mission SA data, demonstrated strong performance in multistep-ahead predictions, highlighting the potential of attention mechanisms in improving spatial feature extraction.

A significant limitation of current spatiotemporal sea level forecasting approaches using ML/DL models is their tendency to underestimate peak events (Qin et al., 2023). Al Kajbaf & Bensi (2020) evaluated the performance of ANN, Gaussian Process Regression (GPR), and SVR models in predicting storm surge peaks along U.S. coastlines. While ANNs outperformed other methods, they still systematically underestimated large surge events. Similar findings were reported by Bruneau et al. (2020), who demonstrated that ANN models globally struggle to accurately predict rare, high-intensity extremes. This pattern of underprediction has been consistently observed across multiple studies (Sztobryn, 2003; Žust et al., 2021; Pachev et al., 2023; Sun & Pan, 2023). The primary cause of this limitation lies in model bias toward normal sea level conditions, coupled with insufficient representation of extreme events in training datasets. As noted by Ramos-Valle et al. (2021), this skewness toward ordinary conditions rather than extremes fundamentally limits the models' ability to capture exceptional sea level variations. Given this, there is a limited number of studies assessing the ML model performance in ESL forecasting. In the Baltic Sea region recent studies have focused on classifying extreme and non-extreme events, for example, Bellinghausen et al. (2025) used a RF model to classify extremes and non-extreme events for next days' sea levels across Baltic stations. However, this approach does not quantify the real performance of the model and the underestimation level. Harter et al. (2024) highlighted that adding wave properties (like SWH) and wind stress can reduce the underestimation levels of ESL prediction.

Another important aspect is the need for high-resolution sea-level forecasting models, as they are essential for resolving fine-scale ocean features such as small eddies, filaments, and vortices. These high-resolution mappings are especially vital in dynamic regions such as the Baltic Sea, where sea levels exhibit rapid variations across both space and time (Weisse et al., 2021). Utilizing high-resolution datasets allows models to produce more precise eddy maps, capturing sub(mesoscale) structures that govern ocean circulation patterns. Enhancing model resolution and incorporating such detailed data is therefore essential for improving forecast accuracy in active marine zones and supporting informed coastal management decisions (Ghosh et al., 2024).

The literature review analysis leads to the following key conclusions:

- Recent sea level forecasting studies have increasingly prioritized spatiotemporal field forecasting, requiring advanced DL frameworks to capture complex oceanographic patterns across both space and time (**Publications I and II**).
- An absolute sea level (ASL) reference (geoid) is required for consistency across observational tools and for inter-comparison validation (**Publications I–III**).

- Multivariate prediction methods have become dominant in contemporary research due to their demonstrated ability to identify and model nonlinear relationships between sea level dynamics and multiple environmental drivers (**Publications I–III**).
- High-resolution schemes are needed to capture the sub mesoscale ocean features, specifically for dynamic regions (**Publications II**).
- Current spatiotemporal ML models exhibit a systematic limitation by underestimating peak sea level events, as they tend to optimize normal conditions rather than extremes. This highlights the need for specialized modeling strategies targeting extremes forecasting (**Publication III**).
- There remains a critical need to evaluate and improve the explainability of DL models used in sequence-to-sequence sea level forecasting through Explainable Artificial Intelligence (XAI) techniques (**Publication III**). This is particularly important for operational forecasting systems requiring explainable results.

2.3 Developed ML Approaches

2.3.1 Convolutional Neural Networks (CNN)

Convolutional neural networks, introduced by Yann LeCun in the 1980s (LeCun et al., 2015), are a class of DL models particularly well-suited for processing grid-structured or array data (images, audio, video) and spatiotemporal datasets. For the sea level forecasting, we frame the problem as an image-to-image regression task where input fields (past sea level maps and associated atmospheric/oceanic variables) are transformed into forecast outputs (future sea level maps). This formulation makes CNNs particularly suitable for capturing the complex spatiotemporal variability of dynamic SSH in the Baltic Sea.

In **Publication I**, we leveraged CNNs for daily sea level forecasting in the Baltic Sea, treating the problem as a spatiotemporal regression task. CNN architecture excels at capturing local and large-scale spatial patterns through its hierarchical feature extraction layers (e.g., convolutional and pooling operations). Additionally, its ability to handle multivariate input data (e.g., sea levels, wind speed, pressure, SST) makes it highly effective for predictive modeling in oceanographic applications. By structuring the forecasting problem in this way, the CNN model efficiently learns the underlying spatiotemporal dependencies in sea level variability, providing robust short-term predictions while maintaining computational efficiency.

As displayed in Figure 7, CNN models consist of several key components that work together to process spatial and spatiotemporal data. The first fundamental component is the convolutional layer, which applies learnable filters to extract local features through discrete convolution operations. The convolution operation at position (i, j) in layer l is mathematically expressed as:

$$H_{i,j}^{k} = f\left(\sum_{m=1}^{M}\sum_{n=1}^{N} W_{m,n}^{(k)}\, X_{i+m-1,j+n-1}^{(m,n)}\right) + b_k, \tag{6}$$

where $W^{(k)} \in \mathbb{R}^{k\times k\times C_{\text{in}}\times C_{\text{out}}}$ are the filter weights with kernel size k, $X^{(l-1)} \in \mathbb{R}^{H\times W\times C_{\text{in}}}$ is the input feature map from the previous layer with height H, width W, and C_{in} input channels, $b_k \in \mathbb{R}^{C_{\text{out}}}$ is the bias term for each output channel, and f is a nonlinear activation function (e.g., ReLU, Sigmoid, Tanh, as displayed in Table 3). The convolution operation also slides filters (with different filter or kernel sizes) across the input with a specified stride s, producing output feature maps that capture local spatial patterns.

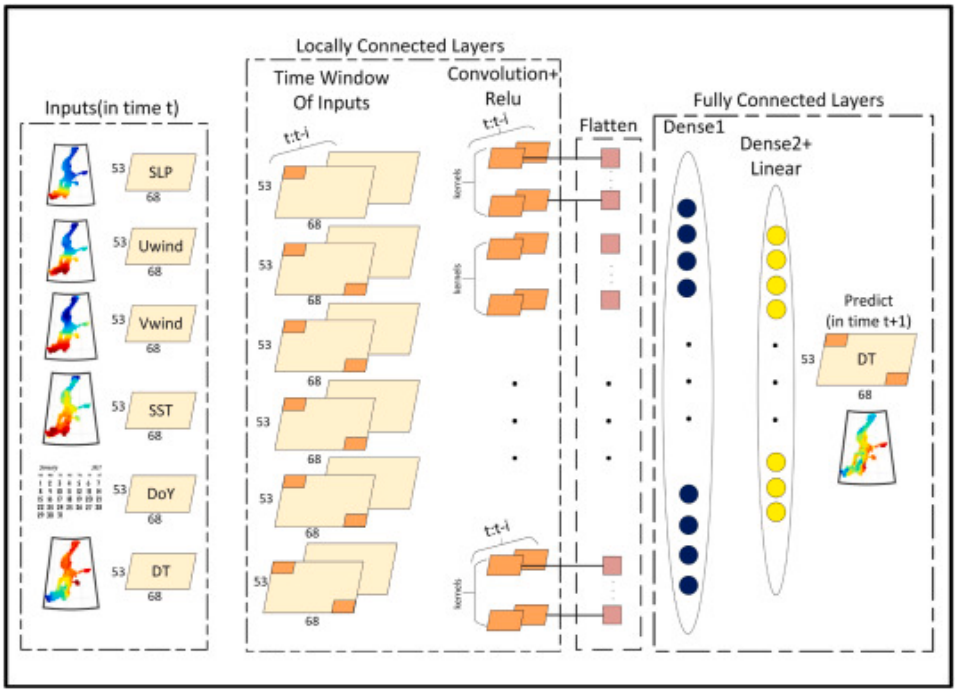


*Figure 7. CNN model for multivariate daily sea level forecasting in the Baltic Sea (from **Publication I**).*

Following convolutional layers, CNNs typically include pooling layers that reduce spatial dimensions and integrating similar features while retaining important features through subsampling operations. The most common pooling operation is the max pooling operation selects the maximum value in each $k \times k$ window:

$$Z_{i,j} = \max_{(m,n) \in N(i,j)} A_{i+m,j+n}, \tag{7}$$

where $N(i,j)$ defines the neighborhood around position (i,j). An alternative approach is the average pooling. These pooling operations are typically applied with strides equal to the kernel size *k* to achieve downsampling.

Convolutional blocks are being used to first extract low-level and then higher-level features from inputs before the final flattening layer that will convert the learnt feature maps into a vector $h \in \mathbb{R}^d$ through the flattening operation:

$$h = flatten(Z). \tag{8}$$

This is then fed into fully connected (dense) layers that perform the final prediction mapping through matrix multiplication and nonlinear transformations:

$$h^{(l+1)} = f\left(W^{(l)} h^{(l)} + b^{(l)}\right), \tag{9}$$

where $W^{(l)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ contains the weights connecting all input features to output neurons, and $b^{(l)}$ is the bias vector.

For regression tasks like sea level forecasting, the output layer typically uses a linear activation:

$$\hat{y} = W^{(L)} h^{(L-1)} + b^{(L)}, \tag{10}$$

where $W^{(L)} \in \mathbb{R}^{\text{output\_dim} \times d}$ and $b^{(L)} \in \mathbb{R}^{\text{output\_dim}}$. The network is trained to minimize a loss function, commonly the MSE with optional L2 regularization:

$$L = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 + \Lambda \|W\|_F^2, \tag{11}$$

where Λ controls the strength of L2 regularization and $\|\cdot\|_F$ denotes the Frobenius norm.

Altogether, CNNs offer several advantages, including local connections (or local receptive fields), which allow the network to focus on small regions of an image. This is important because nearby pixels tend to have stronger correlations. CNNs also benefit from weight sharing, meaning that once a filter has learned to detect a particular feature in one part of the image, it can detect the same pattern in any other part. This contrasts with fully connected neural networks, which require separate weights for each location. Together, these properties make CNNs highly efficient for image processing.

As displayed in Figure 7, for daily sea level forecasting of Baltic Sea in **Publication I**, the CNN architecture processes multivariate input grids (with channels (features) including past sea levels, wind speeds, SLP, SST, and DOY) through convolutional block to extract hierarchical spatial features, followed by fully connected layers that integrate these learnt features to produce the final predictions. The complete forward pass transforms input fields X through these operations to produce forecasts $\hat{Y}$, with all parameters optimized end-to-end via backpropagation to minimize prediction error on the training data. For the CNN model, we used the ReLU activation function in the intermediate layer to add nonlinearity term and linear for the final layer, to make it possible for all range of values.

2.3.2 Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)

In many problems, it is not possible to have fixed-size inputs and outputs. RNNs are types of DL models designed for sequential data processing, making them particularly effective for sequence-to-sequence processing like speech recognition, and time series forecasting applications including multiple future horizons predictions using the same architecture. RNNs have been started with standard (also called vanilla) RNNs that process sequential information through hidden states by capturing temporal dependencies. However, they suffer from the vanishing or exploding gradient problem. This was shown in Bengio et al. (1993) that learning long-term dependencies in recurrent networks using gradient descent is extremely difficult limiting their ability to learn long-term patterns (like 12–24 hours ahead). This is because when computing the derivatives for a back propagation in time, for loss optimization, this will result in multiplying many small numbers that leads to gradient shrinking exponentially and very low updates in the weights and the model cannot learn anymore. Exploding gradient problem can be fixed by gradient clipping and the main challenge is for gradient vanishing.

To address these challenges, more advanced or gating RNNs—LSTM and GRU—were introduced. Both LSTM and GRU utilize gating mechanisms to not forget the important information and to regulate the flow of information through time and how much past

information should be forwarded without repeated multiplication, effectively keeping long-range dependencies that standard RNNs struggle to capture.

The LSTM architecture (see Figure 8), introduced by Hochreiter & Schmidhuber (1997), employs three specialized gates to control information flow. The forget gate determines what information to discard from the cell state, where P is the sigmoid activation function. The input gate regulates which new information to store, while a candidate cell state is created. These components update the cell state through Tanh, where . denotes element-wise multiplication.

$$\begin{aligned} f_t &= \mathrm{P}\big(W_f[h_{t-1}, x_t] + b_f\big), \\ i_t &= \mathrm{P}(W_i[h_{t-1}, x_t] + b_i), \\ \hat{C}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c), \\ C_t &= f_t \,.\, C_{t-1} + i_t \,.\, \hat{C}_t, \\ o_t &= \mathrm{P}(W_i[h_{t-1}, x_t] + b_o), \\ h_t &= o_t.\tanh(C_t). \end{aligned} \tag{12}$$

Finally, the output gate controls the hidden state output. This elaborate mechanism allows LSTMs to maintain relevant information over extended sequences, making them particularly powerful for modeling the complex temporal patterns in sea level data where historical trends and external forcings influence future states.

The GRU architecture (Figure 8), proposed by Cho et al. (2014), offers a more streamlined alternative with two gates. The update gate decides how much of the previous information to retain, while the reset gate determines how much past information to ignore when computing the new candidate state $\hat{h}_t$.

$$\begin{aligned} z_t &= \mathrm{P}(W_z[h_{t-1}, x_t] + b_z), \\ r_t &= \mathrm{P}(W_r[h_{t-1}, x_t] + b_r), \\ z_t &= \mathrm{P}(W_z[h_{t-1}, x_t] + b_z), \\ r_t &= \mathrm{P}(W_r[h_{t-1}, x_t] + b_r). \end{aligned} \tag{13}$$

The final hidden state is then calculated as

$$\begin{aligned} \hat{h}_t &= tanh(W[h_{t-1}, x_t] + b), \\ h_t &= \big(1 - z_t\big).h_{t-1} + z_t.\hat{h}_t\big). \end{aligned} \tag{14}$$

For Equations (12)–(14), x_t denotes the input at time step t (e.g., the current observation or feature vector), and h_{t-1} represents the hidden state from the previous time step, which carries short-term memory information. The term $[h_{t-1}, x_t]$ indicates the concatenation of the previous hidden state and the current input, forming the standard input to the gating mechanisms. The matrices W_* denote learnable weight parameters associated with each gate or candidate computation, while b_* represent the corresponding learnable bias vectors.

The function $\sigma(\cdot)$ denotes the sigmoid activation function, which maps values to the range (0, 1) and is used within the gates to regulate the flow of information. The function $\tanh(\cdot)$ represents the hyperbolic tangent activation, mapping values to (−1, 1), and is used to generate candidate memory or hidden-state values.

For the LSTM formulation, f_t denotes the forget gate, which controls the fraction of the previous cell state retained; i_t represents the input gate, determining how much new

candidate information is written to the cell state; $\hat{C}_t$ is the candidate cell state generated from the current input and previous hidden state; C_t denotes the updated cell state, combining retained and newly added information; o_t represents the output gate, controlling the exposure of the cell state to the hidden state; and C_{t-1} is the previous cell state (long-term memory). The hidden state at time step t, denoted h_t, serves as both the output of the current step and the input to the next time step. For the GRU formulation, z_t denotes the update gate, which controls the balance between retaining the previous hidden state and incorporating new information; r_t is the reset gate, determining how much past information is ignored when computing the candidate state; $\hat{h}_t$ represents the candidate hidden state computed from the current input and the reset-modified previous hidden state; and h_t is the final hidden state output, formed by combining the previous hidden state and the candidate state weighted by the update gate.

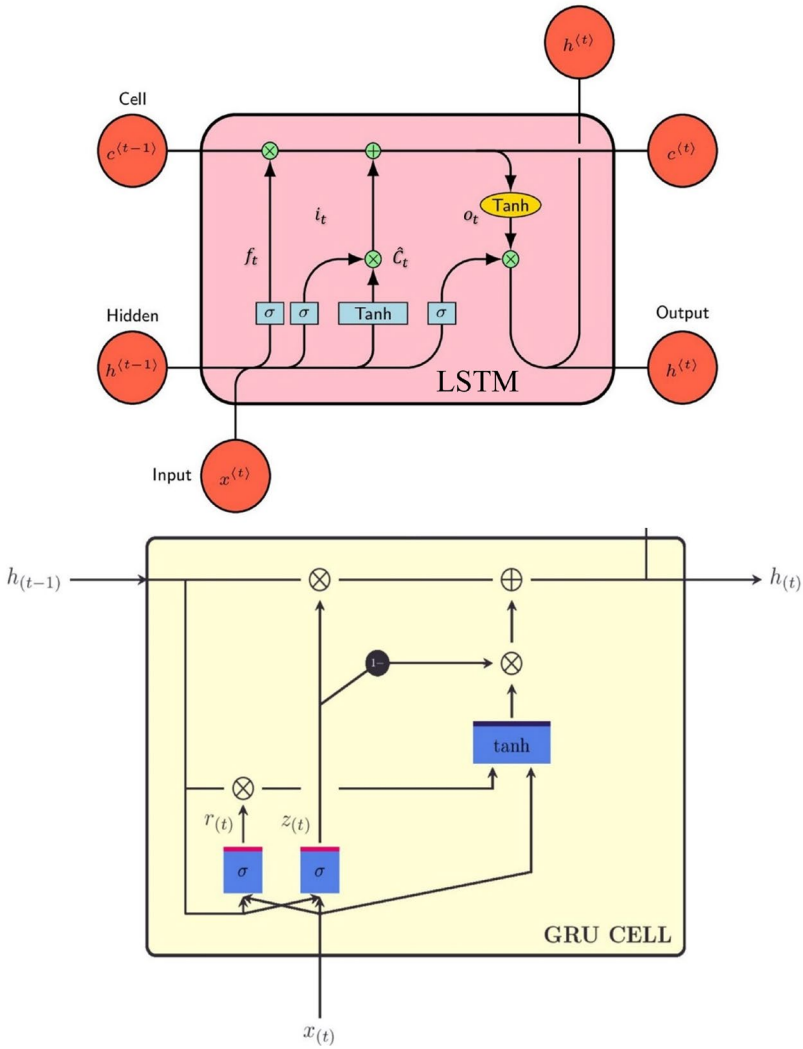


*Figure 8. LSTM (left) and GRU (right) cells employed for hourly sea level forecasting in the Gulf of Finland (Modified from **Publication II**).*

Although computationally more efficient than LSTMs due to this simplified structure, GRUs still maintain robust performance in capturing sequential dependencies, making them well-suited for high-resolution spatiotemporal forecasting tasks.

In this study, we applied both LSTM and GRU models to multistep ahead sea level forecasting in the eastern Baltic Sea, Gulf of Finland. These models were trained on multivariate input sequences (e.g., historical sea levels, zonal wind speed, SLP, and SST) to forecast future sea level states. Their ability to handle non-linear temporal dynamics and external covariates was key for improving the forecast accuracy.

2.3.3 MLP, RF, and XGB

For analyzing the performance of ML models in daily forecasting of SLM, we focus on three fundamental models that have proven particularly valuable in sea level prediction studies: MLP, RF, and XGB. Each of these approaches brings distinct advantages to the complex task of modeling sea level dynamics. Figure 9 shows the architecture of each of these ML models.

The MLP traces its origins to Rosenblatt's perceptron concept in the late 1950s, with significant advancements coming from Rumelhart, Hinton, & Williams' work on backpropagation in the 1980s. This neural network architecture operates through interconnected layers that transform input data through successive nonlinear operations (Figure 9). The MLP represents the most widely used type of ANN, having been extensively applied in ocean forecasting studies (Wei et al., 2019; Feng et al., 2020). The architecture consists of: (i) an input layer receiving feature vectors, (ii) one or more hidden layers with nonlinear activation functions, and (iii) an output layer producing predictions. The forward propagation for a single hidden layer MLP can be expressed as:

$$\hat{y} = \sigma_2(W_2.\sigma_1(W_1.X + b_1) + b_2) \quad (15)$$

where X is the input vector, W_1,W_2 are weight matrices, b_1, b_2 are bias vectors, and σ_1, σ_2 are activation functions (typically ReLU for hidden layers).

The model learns through backpropagation by minimizing the loss function L (typically MSE (cf. Table 4)) for regression). The model's behavior is shaped by several crucial parameters, including the number and size of hidden layers, choice of activation functions, the learning rate governing parameter updates, batch size for training iterations, total training epochs, optimization method selection such as Adam or SGD, and dropout rates for preventing overfitting. While MLPs excel at identifying complex patterns in sequential data, their effectiveness depends heavily on careful parameter tuning and sufficient training data.

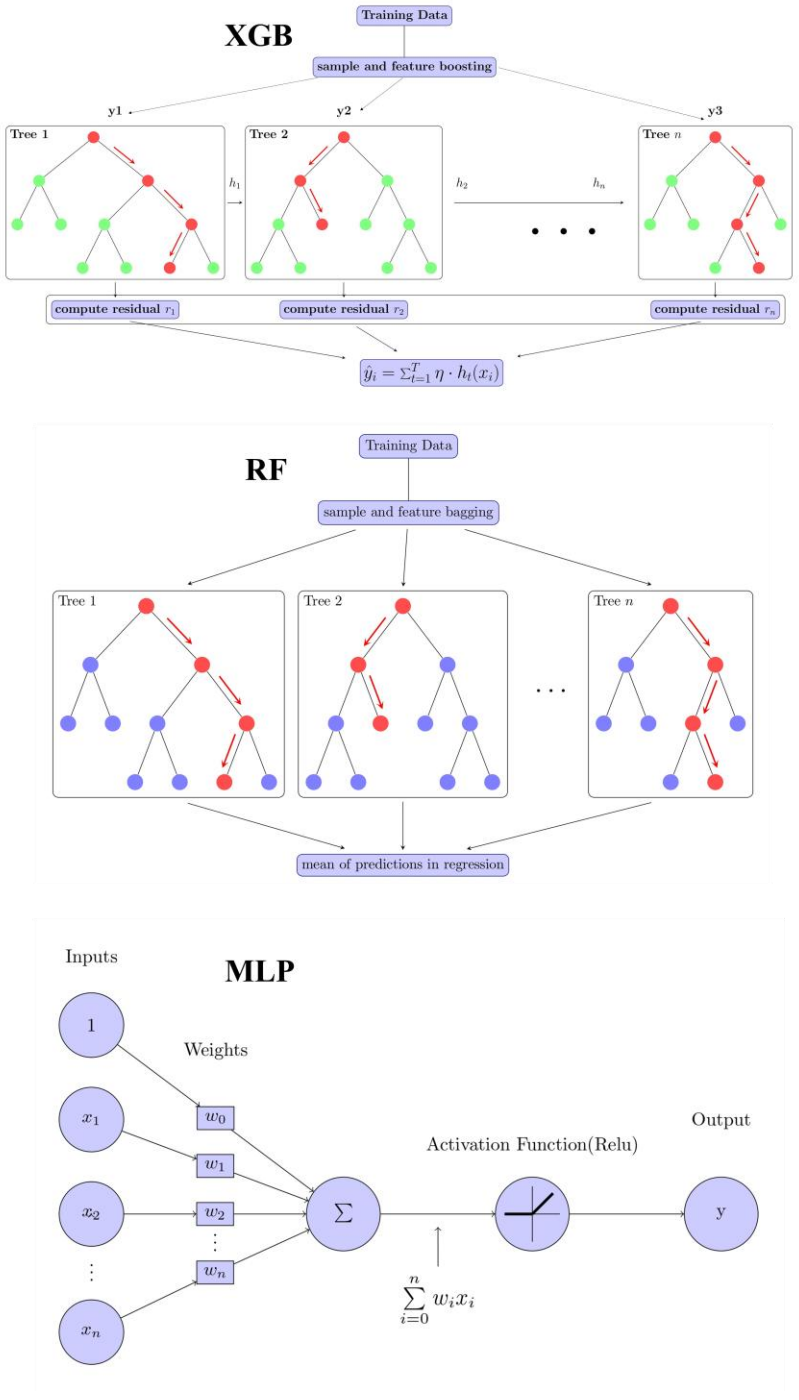


*Figure 9. Architecture of different ML models (XGB, RF, and MLP models) for daily SLM forecasting in the Baltic Sea (Modified from **Publication III**).*

Developed by Breiman (2001), RF is an ensemble method that creates multiple decision trees trained on random subsets of data (bagging) and features. This ensemble method creates numerous decision trees; each trained on randomly sampled subsets of both instances and features from the training data (Figure 9). The RF's predictive capability stems from averaging across all constituent trees, with key parameters including the total number of trees grown, maximum permitted depth for individual trees (also called m_{try}), minimum samples required to form leaf nodes or split internal nodes, the fraction of features considered at each split, and whether bootstrap sampling is employed. The method's inherent feature selection mechanism provides valuable insights into variable importance while maintaining robustness against overfitting, making it particularly suitable for preliminary analysis of sea level data.

Random Forest is a powerful ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of individual trees (for regression tasks). Key characteristics include Bootstrap aggregating (bagging) to reduce variance, Random feature selection at each split to decrease correlation between trees, and Built-in feature importance analysis through permutation metrics. The prediction for regression is the average of individual tree predictions:

$$\hat{y} = {}^{1}/_{B} \sum_{b=1}^{B} T_b(x), \tag{16}$$

where T_b represents the b-th decision tree in the ensemble of size B. Each tree is grown by recursively partitioning the feature space to minimize impurity (typically using MSE for regression).

Extreme gradient boosting (or XGBoost) (Chen, 2014) is an advanced gradient boosting algorithm that builds decision trees sequentially, where each new tree focuses on correcting the residual errors left by earlier trees (Figure 9). XGB builds models sequentially, with each new tree specifically targeting the residual errors of its predecessors. The algorithm's configuration involves setting the number of boosting iterations, learning rate that controls contribution of each tree, maximum depth of individual trees, minimum sum of instance weights needed in child nodes, random subsampling ratios for both features and training instances, regularization terms that penalize model complexity, and minimum loss reduction required to make additional partitions. XGB's ability to handle missing values and provide multiple feature importance metrics, combined with its typically superior predictive performance, makes it particularly valuable for sea level forecasting tasks where data quality and completeness may vary. It minimizes a regularized objective function:

$$L(\emptyset) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + f_t(x_i)) + \Omega(f_t),\ \Omega(f_t) = \gamma T + {}^{1}/_{2}\,\lambda \|w\|^2, \tag{17}$$

where L is the loss function and Ω is the regularization term penalizing model complexity (γ=complexity cost, λ is L2 regularization, T=number of leaves), and f_t represents the t-th tree. The model updates predictions additively:

$$\hat{y}_i(t) = \hat{y}_i(t-1) + \eta f_t(x_i), \tag{18}$$

where η is the learning rate. Key advantages like handling missing values, sophisticated regularization, and multiple feature importance metrics make XGB particularly effective for sea level forecasting.

2.3.4 CNN-LSTM and CNN-GRU Hybrid DL Models

Hybrid models like CNN-LSTM and CNN-GRU represent advanced hybrid deep neural network architecture specifically designed for spatiotemporal forecasting tasks, making them particularly suitable for sea level prediction, where both spatial patterns and temporal dynamics must be captured simultaneously. These models synergistically combine the strengths of CNNs with recurrent architecture, offering superior performance for complex hydrological time series analysis. The foundation of these hybrid approaches traces back to several key developments in DL.

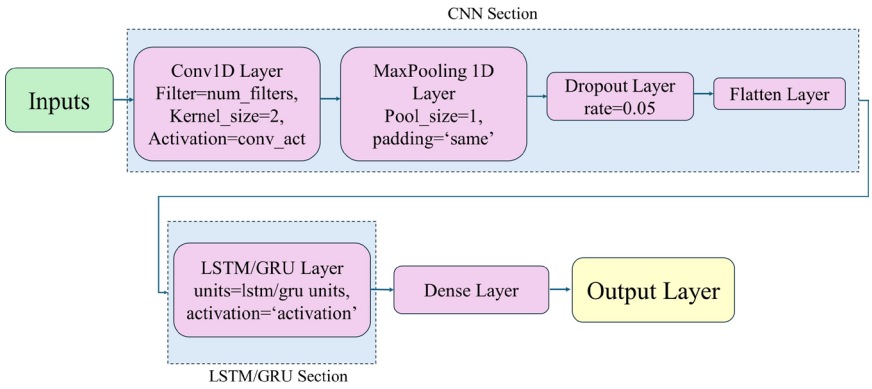


*Figure 10. Proposed CNN-LSTM and CNN-GRU architectures for daily SLM forecasting in the Baltic Sea (from **Publication III**).*

With described earlier the standalone CNN, LSTM, and GRU architectures, in the CNN-LSTM architecture, the spatial processing capabilities of CNNs work in tandem with LSTM's temporal modeling (Figure 10). The CNN component typically consists of multiple convolutional layers that automatically extract relevant spatial features from input data through learned filters, followed by pooling operations that reduce dimensionality while preserving important patterns. These extracted features then feed into LSTM layers that model temporal dependencies through their input, output, and forget gates, which collectively determine what information to retain or discard across time steps. The model's behavior is influenced by several critical parameters, including the number and size of convolutional filters, kernel dimensions, pooling strategies, the number of LSTM units, dropout rates between layers, and the learning rate for optimization. The CNN-LSTM has proven particularly effective in sea level forecasting where coastal topography and atmospheric conditions create complex spatial patterns that evolve over time.

The CNN-GRU variant replaces the LSTM component with GRU cells, offering a more computationally efficient alternative while maintaining strong temporal modeling capabilities. The GRU's simplified architecture combines the forget and input gates into a single update gate and merges the cell state with the hidden state, reducing the parameter count without compromising performance. Key configuration parameters for CNN-GRU include similar CNN hyperparameters as the CNN-LSTM version, along with GRU-specific settings such as the number of GRU units, recurrent dropout probability, and activation functions. The reduced complexity of GRU cells often makes CNN-GRU preferable when working with limited training data or when faster model convergence is desired.

These hybrid architectures (CNN-LSTM and CNN-GRU) offer distinct advantages for sea level forecasting. The CNN components effectively process spatial relationships in data from multiple monitoring stations or gridded climate models, while the recurrent layers (LSTM or GRU) capture the temporal evolution of sea level changes influenced by tides, weather patterns, and climate oscillations. The models automatically learn relevant features at multiple timescales, from short-term storm surges to seasonal variations and long-term sea level rise trends. Both architectures support multivariate input, enabling the incorporation of various environmental factors such as wind fields, atmospheric pressure, and other influential features that influence sea level dynamics.

2.4 Model Explainability

Explainability in ML has become increasingly critical, as many advanced models operate as "black boxes" while they provide predictions, their internal decision-making processes remain opaque. This lack of transparency is addressed through explainability techniques, which help uncover the underlying factors driving model output. For this purpose, we employed SHAP (Lundberg & Lee, 2017), a robust explainability method rooted in cooperative game theory. SHAP quantifies the contribution of each input feature to individual predictions by fairly distributing the "payout" (impact) among all features, ensuring a consistent and interpretable measure of influence. By leveraging SHAP, we not only enhance trust in the model's outputs but also align its decision-making logic with known physical processes in Baltic Sea dynamics. Formula for SHAP analysis expressed as below:

$$\phi_i = \sum\{S \subseteq F\{i\}\}[|S|!\,(|F| - |S| - 1)!]\,/|F|!\;[f(S \cup \{i\})\; -\; f(S)]. \tag{19}$$

where ϕ_i denotes the contribution (SHAP value) of feature $i$to the prediction; the summation is taken over all possible subsets S of the feature set F that exclude feature i; $|S|$ represents the number of features in subset S, and $|F|$ denotes the total number of features; $f(S)$ is the model output when only the features in subset S are used; and $f(S \cup \{i\})$ is the model output when feature i is added to subset S.

We applied the explainability method to the CNN-GRU model for SLM forecasting, for which the results have been shown in Section 4.5 and **Publication III**.

2.5 Extreme Value Theory (EVT)

Extreme value analysis was conducted to systematically evaluate potential underestimation level in the ML models' predictions for long-term sea level forecasting. We implemented a seasonal extreme value analysis framework using the Generalized Extreme Value (GEV) distribution (Coles, 2001). The analysis focused on return periods of 2, 5, 8, 10, 15, 20, 50, and 100 years using the seasonal block maxima method, where we extracted annual maximum sea levels from the observations. The GEV distribution, characterized by its location (μ), scale (σ), and shape (ξ) parameters, was fitted using maximum likelihood estimation:

$$F(x) = \exp\{-[1 + \xi(x - \mu)/\sigma]^{(-1/\xi)}\}. \tag{20}$$

where $F(x)$ denotes the cumulative distribution function (CDF), representing the probability that the variable is less than or equal to x; x is the value at which the distribution is evaluated; $\exp\{\cdot\}$ denotes the exponential function; μ is the location

parameter, which shifts the distribution along the horizontal axis; σ is the scale parameter ($\sigma > 0$), controlling the spread of the distribution; and ξ is the shape parameter, which determines the tail behavior (heavy, light, or bounded).

More details on the approach and results are discussed in Section 4.5 and **Publication III** (Section 2.2.4).

2.6 Evaluation Metrics

This dissertation employs a range of evaluation metrics to assess the performance of ML and DL models applied for sea-level forecasting. The selected metrics are widely used in ML-based forecasting studies and were chosen to enable direct comparison with previous research.

For the regression tasks in **Publications I–III**, the primary evaluation metrics include first appearance much earlier RMSE, MSE, correlation coefficient (R), coefficient of determination (R^2), and the Willmott index (WI). For the detection of extreme sea-level peak events, additional classification-based metrics—precision, recall, and F1-score—were employed.

To quantify forecasting uncertainty in **Publication II**, the prediction interval coverage probability (PICP) index was used. During EDA, the Pearson correlation coefficient and MI index were applied, and the optimal feature lag was determined using the Bayesian Information Criterion (BIC). Table 4 provides a summary of all evaluation metrics used in the dissertation, along with their corresponding equations and descriptions.

In table 4, y_i denote the observed (measured) sea level at time step i, and $\hat{y}_i$ the corresponding predicted sea level obtained from the ML or DL model. The total number of samples is denoted by n. The mean of the observed and predicted sea level series are represented by $\bar{y}$ and $\bar{\hat{y}}$, respectively. In the classification-based evaluation of ESL events, TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. For uncertainty quantification, L_i and U_i represent the lower and upper bounds of the prediction interval for the i-th observation, and $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition inside the brackets is satisfied and 0 otherwise. In the MI formulation, $p(x)$, $p(y)$, and $p(x, y)$ denote the marginal and joint probability distributions of variables X and Y. For the BIC, k represents the number of model parameters and $\hat{L}$ denotes the maximized value of the likelihood function.

Table 4. Different Evaluation metrics used in this dissertation.

Metric	Equation	Description
Mean Squared Error (MSE)	$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	Measures the average squared difference between observed and predicted sea level values
Root Mean Squared Error (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{\text{n}}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$	Square root of MSE, expressed in the same unit as the observations
Pearson Correlation Coefficient (r)	$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$	Measures the linear correlation between observed and predicted values
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Indicates the proportion of variance explained by the model
Willmott Index (WI)	$WI = 1 - \frac{\sum_{t=0}^{m-1}(\hat{y}_i - y_i)^2}{\sum_{t=0}^{m-1}(\lvert\hat{y}_i - \bar{y}\rvert + \lvert y_i - \bar{y}\rvert)^2}$	Measures agreement between observed and predicted values
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	Fraction of correctly detected SLM events
Recall	$\text{Recall} = \frac{TP}{TP + FN}$	Fraction of correctly detected SLM events
F1 Score	$F1 = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall for peak event detection
Prediction Interval Coverage Probability (PICP)	$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(y_i \in [L_i, U_i])$	Measures the proportion of observations lying within the prediction interval
Mutual Information (MI)	$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)log\,(\frac{p(x,y)}{p(x)p(y)})$	Quantifies nonlinear dependency between variables
Bayesian Information Criterion (BIC)	$\text{BIC} = kln\,(n) - 2ln\,(\hat{L})$	Model selection criterion used to determine the optimal feature lag

3 ML for Spatiotemporal Sea Level Forecasting

In this section, we address spatiotemporal sea level forecasting in the Baltic Sea using multivariate DL approaches (Equation 5). The primary target variable is DT, treated as absolute sea level and derived from a corrected hydrodynamic model (Jahanmard, 2024), with the NKG2015 geoid model used as the vertical reference. Alongside historical DT, the multivariate framework incorporated wind speed (zonal and meridional), sea level pressure (SLP), SST, SSS and day of the year (DOY) as supplementary predictors. The dataset covers the period from 1 January 2017 to 31 December 2019, constrained by the availability of the corrected hydrodynamic model outputs.

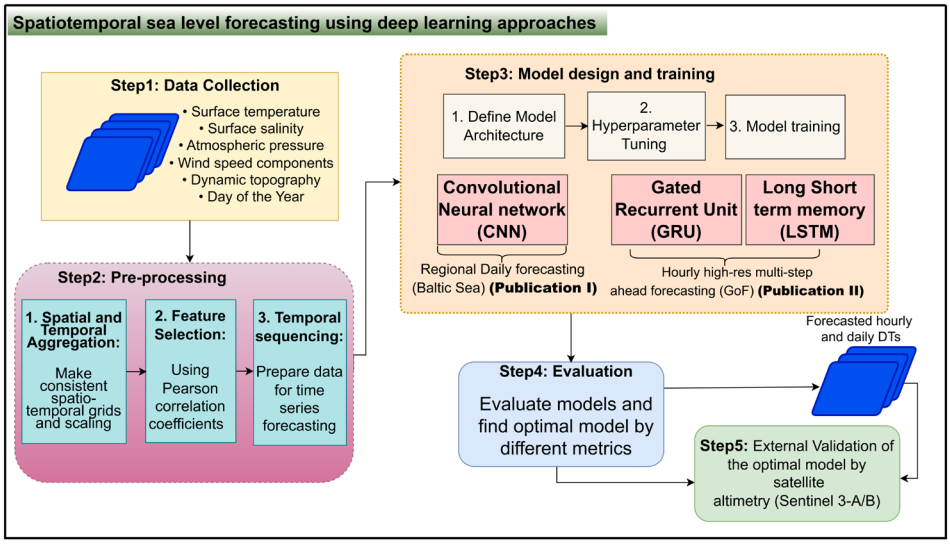


*Figure 11. Scheme of the spatiotemporal sea level forecasting using CNN (**Publication I**) along with LSTM and GRU models (**Publication II**) in the Baltic Sea.*

Publication I developed a CNN model for day-ahead sea level forecasting across the Baltic Sea at 0.25°×0.25° resolution using SLP, wind speed, SST, and day of year (80:20 train-test split). The CNN architecture—selected for grid-structured data—comprised convolutional, pooling, activation, flattening, and fully connected layers with a 7-day lag window. Forecasts were externally validated against Sentinel-3A satellite altimetry.

Publication II extended this framework to higher resolution (one nautical mile, hourly) and multi-step-ahead forecasting in the Gulf of Finland—a dynamically complex region. LSTM and GRU sequence-to-sequence models generated forecasts at 3, 6, 9, 12, and 24-hour lead times using a 12-hour lag window of multivariate inputs (zonal wind, SST, SLP) with an 85:15 data split. Implemented in an HPC environment, models were validated against Sentinel-3A and Sentinel-3B altimetry. Figure 11 summarizes the spatiotemporal methodology.

Hyperparameters for Conv2D (two-dimensional CNN), LSTM, and GRU were selected via trial-and-error (see Table 5). Key findings are summarized below, with full details in **Publications I** and **II**.

*Table 5. Proposed DL Model hyperparameters for spatiotemporal sea level forecasting in the Baltic Sea (**Publications I and II**).*

Models	Model parameters	Chosen hyperparameter
Conv2d **(Publication I)**	Activation functions	ReLU and Linear (last layer)
	Number of Training epochs	10
	Optimizer	'Adam'
	Loss function	'MSE'
LSTM and GRU **(Publication II)**	LSTM/GRU Units	512
	Activation Functions	Sigmoid (gates), Tanh (state)
	Batch Size	128
	Number of Training Epochs	50
	Loss Function	'MSE'
	Optimizer	'Adam'
	Dropout Rate Regularization	0.1
	Kernel Regularization	L2, 0.01

3.1 Feature Selection

Input feature selection combined domain knowledge of the Baltic Sea with EDA. This approach ensured models captured key sea level drivers while maintaining computational efficiency.

Pearson correlation coefficient (PCC) analysis served multiple purposes. It identified the most influential predictors—with pressure and zonal wind consistently dominant—and revealed collinearity among predictors to avoid redundancy.

In **Publication I**, PCC analysis quantified the influence of meteorological variables on sea level variability (Figure 12). Sea level pressure, wind components, and SST showed significant correlations across the Baltic Sea. Wind speed and pressure exhibited particularly strong relationships, with absolute correlations frequently exceeding 0.4 at most locations.

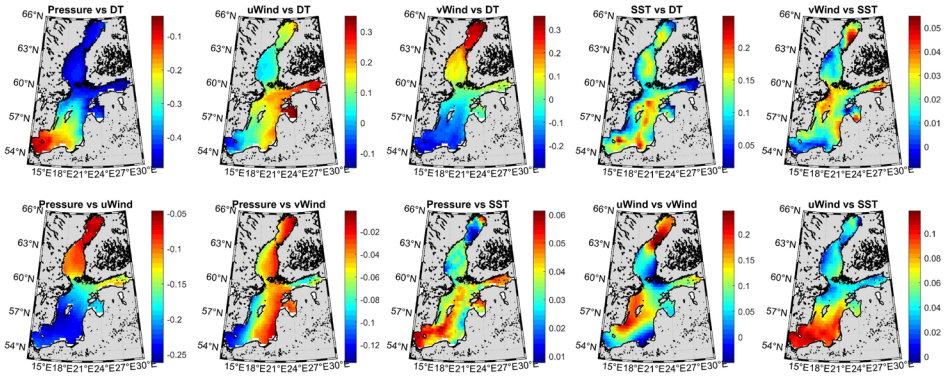


*Figure 12. Spatial correlation between different input features and DT, also the collinearity analysis for the input meteorological features used in Conv2d model (Modified from **Publication I**).*

In **Publication II**, PCC correlation analysis was performed across the Gulf of Finland's high-resolution grid (120 × 259 points) for hourly multistep-ahead forecasting. Initial preprocessing included SSS and SWH alongside previously used predictors. Despite HPC implementation, the high spatial resolution and multistep requirements necessitated balancing complexity and feasibility. Based on PCC analysis, meridional wind, SSS, and SWH—which showed weaker influence—were excluded. While PCC captures only linear relationships, the identified drivers aligned with known Baltic Sea dynamics (e.g., wind-driven surges), providing confidence in feature selection. These additional factors were later addressed **in Publication III** for SLM forecasting.

For four selected grid points (P1–P4) representing eastern, western, southern, and northern Gulf of Finland, PCC analysis served two purposes: assessing regional variations in feature-DT relationships and examining feature collinearity. Results (Figure 13) showed that surface pressure, zonal wind, and SST consistently exhibited the strongest influence across most locations and were retained, while meridional wind and SSS were excluded due to lower impact.

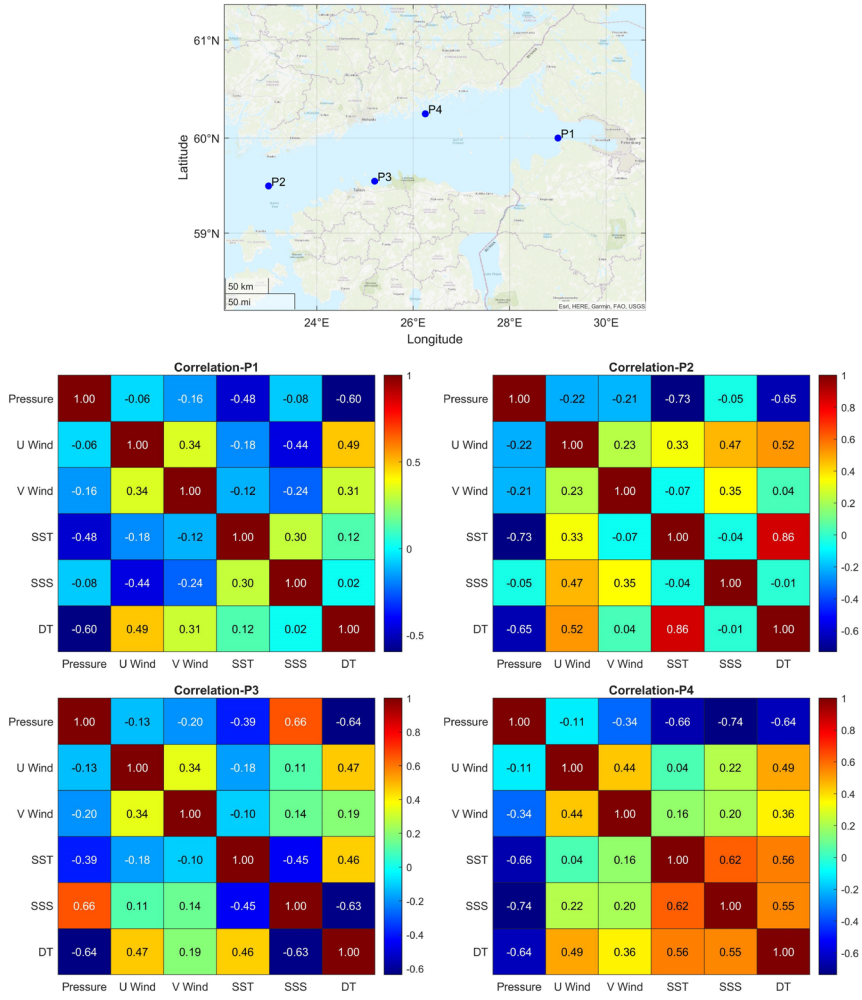


*Figure 13. Pearson correlation analysis for hourly sea level forecasting (Modified from **Publication II**).*

3.2 Overall Performance

For daily sea level forecasting in the Baltic Sea, we implemented a two-dimensional CNN (Conv2D) model, whose architecture and hyperparameters were described previously. As reported in **Publication I** (Table 4), the model achieved strong predictive performance, with $R^2 = 0.95$ and RMSE = 4.7 cm for the training set, and $R^2 = 0.91$ with the same RMSE of 4.7 cm for the test set. The consistently high R^2 values demonstrate substantial explanatory capability, while the identical RMSE across training and test datasets indicates stable generalization and negligible overfitting. These results confirm the robustness and reliability of the proposed sea level forecasting framework.

In **Publication II**, LSTM and GRU models were evaluated for multistep hourly sea level forecasting in the Gulf of Finland (Table 6). During training, GRU achieved R^2 of 0.94 and RMSE of 5.63 cm, slightly outperforming LSTM (R^2 0.93, RMSE 5.81 cm). Both models maintained strong performance across all forecast horizons. On average, GRU achieved RMSE of 4.96 cm and R^2 of 0.93, while LSTM achieved RMSE of 5.3 cm and R^2 of 0.92.

The models demonstrated excellent generalization, with test RMSE for 3 h, 6 h, and 9 h horizons even lower than training errors. Performance decreased slightly at longer horizons (12 h and 24 h) but remained robust—R^2 around 0.92 and RMSE between 5.7 and 6.2 cm—highlighting reliable multistep forecasting capability.

*Table 6. Performance of the LSTM and GRU models for high-resolution sea-level forecasting in the Gulf of Finland at different horizons (Modified from **Publication II**).*

Training and test metrics		Models			
		GRU		LSTM	
		R^2	RMSE (cm)	R^2	RMSE (cm)
Training		0.94	5.63	0.93	5.81
Test (horizons)	3 h	0.96	3.55	0.95	4.13
	6 h	0.95	4.41	0.94	4.85
	9 h	0.92	5.16	0.92	5.47
	12 h	0.91	5.67	0.90	5.87
	24 h	0.89	5.99	0.89	6.17
Test average		0.93	4.96	0.92	5.3

3.3 Time Series Forecasting Performance of Models

To evaluate the time series performance of the proposed Conv2d model for daily sea level forecasting in the Baltic Sea region, we compared the model forecasting results with the observed HDM sea levels at different grid points covering the region during the test period (a1 to a8 in Figure 14). Based on the results in Figure 14, the Conv2D model performs very well in forecasting daily sea levels and successfully captures the trends, dynamics and variations across different regions. However, the model's performance declines when forecasting grid points a1, a3, a5, and a7—most noticeably at a3 and a5—located at Gulf of Finland and Riga, where it tends to underestimate sea levels exceeding 100 cm.

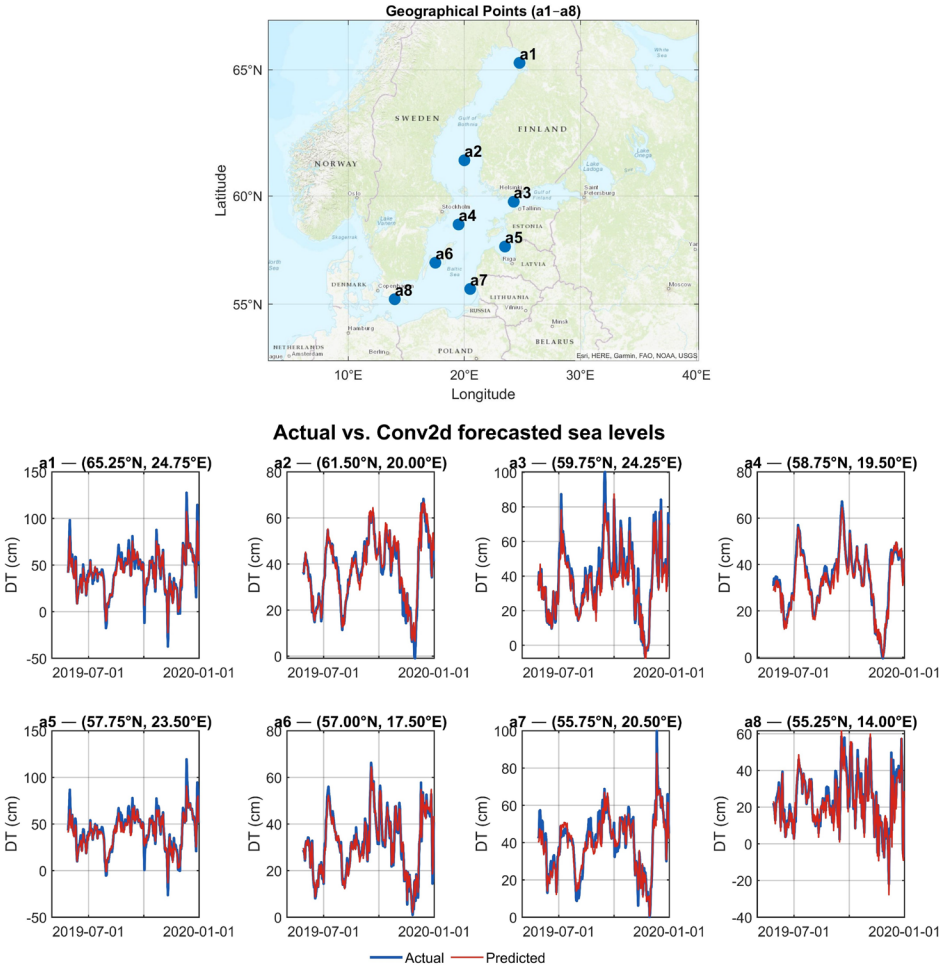


*Figure 14. Conv2d model for multivariate daily sea level forecasting in the Baltic Sea (Modified from **Publication I**).*

In **Publication II**, we extended **Publication I** by incorporating hourly high-resolution spatiotemporal data for the Gulf of Finland—a region characterized by a small Rossby radius and highly dynamic processes requiring finer-scale forecasts. We developed LSTM and GRU sequence-to-sequence models for multi-step-ahead sea-level forecasting. Figure 15 shows model performance at 3-hour and 12-hour horizons across four grid points (P1–P4).

Both models accurately captured general temporal patterns, including short-term fluctuations and longer-term trends. GRU consistently provided a closer fit to observations, particularly during periods of moderate variability. Uncertainty estimates using 90% prediction intervals (PICP metric; Table 4) were generally reliable. However, both models struggled with extreme events above approximately 90–100 cm, where prediction intervals often failed to capture peaks—revealing systematic underestimation.

These findings confirm that high-resolution spatiotemporal data improve forecasting capability but also highlight the need for specialized approaches to capture extreme events critical for coastal risk assessment.

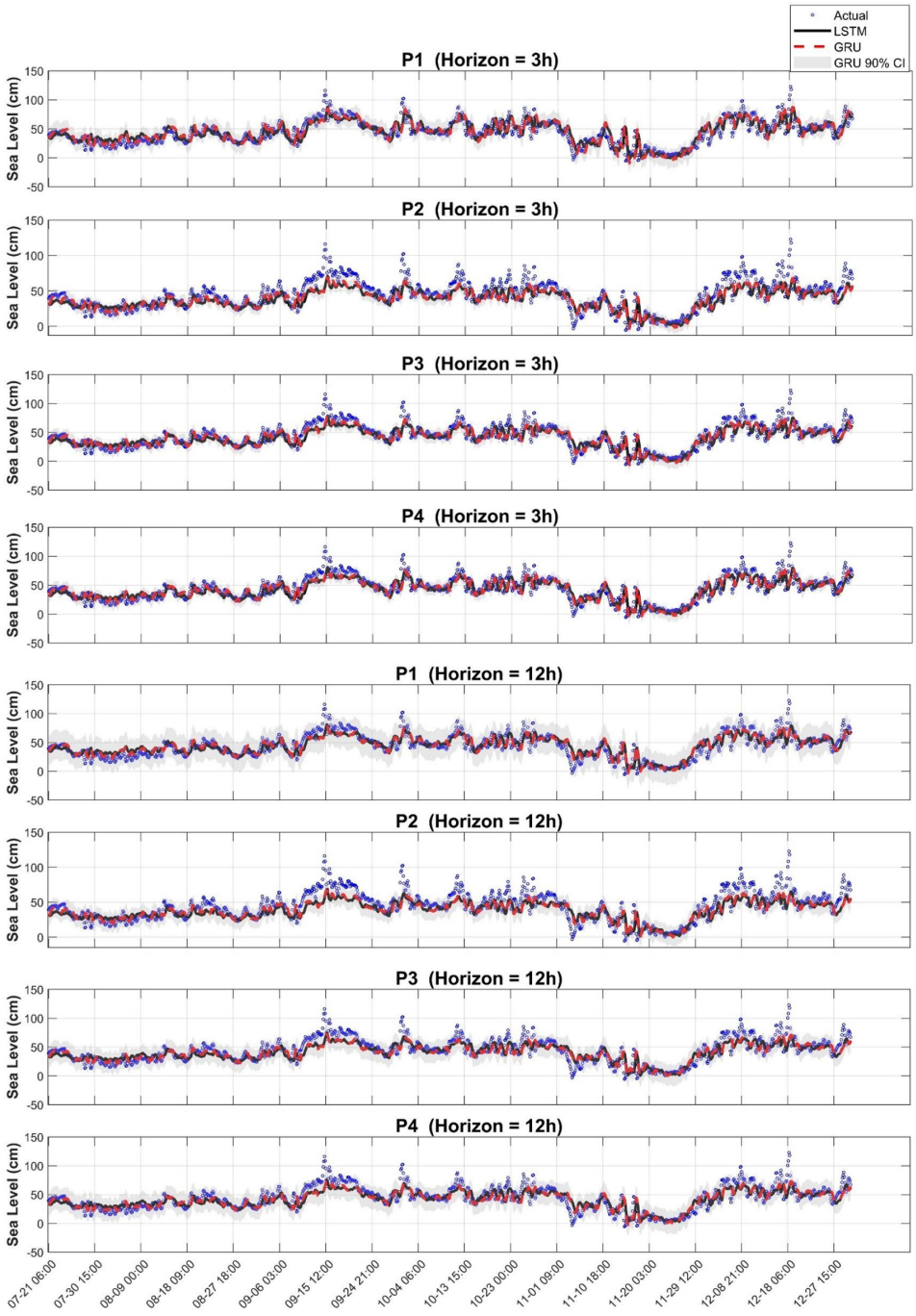


*Figure 15. Time series performance of LSTM and GRU models in different grid points along with the PICP at different horizons (Modified from **Publication II**).*

Seasonal dependence of GRU model prediction errors was examined for a 3-hour forecast horizon at four grid points (P1–P4) using monthly residual boxplots (Figure 16). Median errors remained near zero across all locations, indicating no systematic bias. However, clear seasonal variations in error magnitude and variability were evident.

At P1, December showed the largest spread and most extreme negative errors, suggesting winter performance degradation. P2 exhibited slight underestimation in September–October, with increased December variability. P3 displayed compact error distributions in summer, widening in autumn and winter with larger negative outliers. P4 showed stable summer errors but increased spread in September and December.

Across all points, December consistently exhibited highest variability, indicating that wintertime processes—stronger atmospheric forcing and nonlinear sea level responses—pose greater challenges for the GRU model.

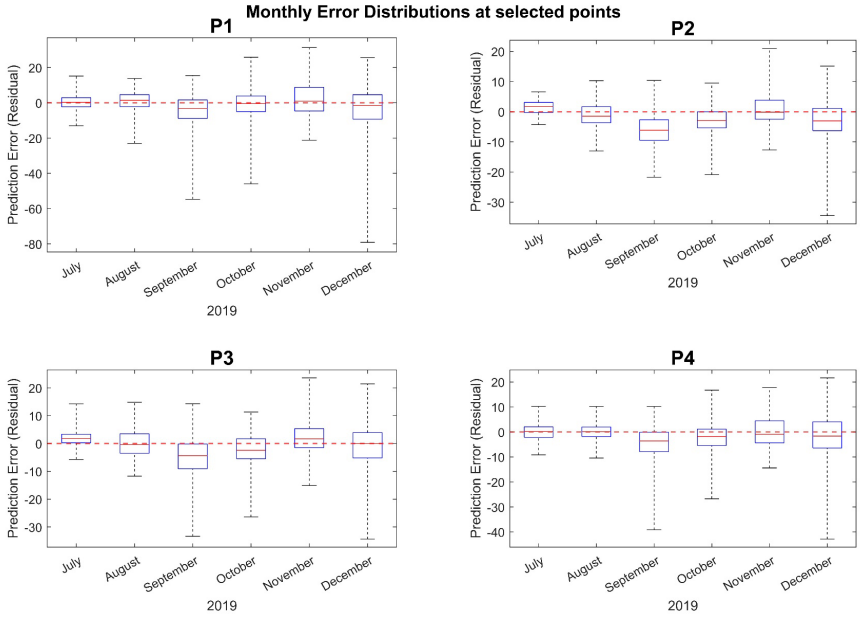


*Figure 16. Boxplot of monthly GRU model errors (predicted minus actual) at selected spatial grid points for horizon 3 h, during different months in the test period (from **Publication II**).*

Residual analysis of GRU model performance across forecast horizons (3–24 hours) was conducted for easternmost (P1) and westernmost (P4) grid points in the Gulf of Finland (Figure 17). At both locations, median residuals remained near zero across all horizons, indicating unbiased predictions. However, interquartile range widened steadily with increasing forecast horizon—more pronounced at P1—reflecting growing uncertainty for longer forecasts.

Histograms of residuals (right panels) showed approximately Gaussian distributions centered near zero for short horizons (3–6 hours), confirming high near-term accuracy. As horizons increased, distributions widened and became slightly skewed, indicating larger errors and increased variance. This effect was stronger at P1, reflecting greater sensitivity of the eastern Gulf of Finland to dynamic variability.

Results confirm that GRU is well-suited for short-term forecasting, but longer horizons require careful consideration of increasing uncertainty.

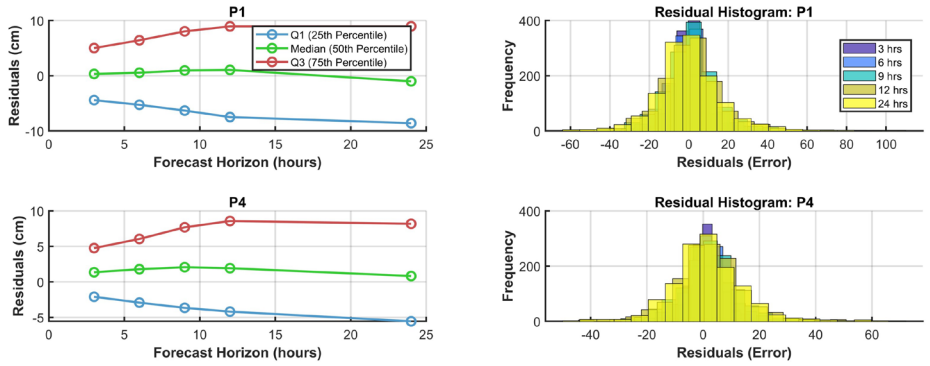


*Figure 17. Residual analysis of GRU model at different horizons for P1 and P4 (Modified from **Publication II**).*

3.4 Spatial Performance of Models

The spatial performance of the Conv2D model during the test period was evaluated using the spatial RMSE computed at each grid point, as shown in Figure 18. This metric provides a detailed view of how the model behaves across different regions of the Baltic Sea by highlighting local variations in forecasting accuracy.

Overall, the Conv2D model demonstrated strong performance throughout most of the domain. For most grid points, particularly in the central and northern parts of the basin, the RMSE remained below 1 cm, indicating that the model was able to consistently reproduce the spatial sea-level patterns with very high accuracy. In the middle Baltic, slightly larger errors were observed, with RMSE values ranging between 3 and 5 cm. These moderate deviations likely reflect the influence of more complex circulation features or variable meteorological forcing that the model captures with somewhat reduced precision.

The largest discrepancies appeared in the southeastern Baltic Sea, especially near the Polish coastline, where the RMSE exceeded 10 cm. These higher errors suggest that the Conv2D model faces challenges in regions characterized by intricate coastal geometry, shallow bathymetry, and stronger local wind-driven dynamics.

The spatial performance of the GRU model in the Gulf of Finland was evaluated for multiple forecast horizons using two approaches: mean performance over the test period (autumn–winter 2019) and an instantaneous example from November 2019.

Mean spatial results (Figure 19) showed that the GRU accurately captured dominant sea-level dynamics. For short horizons (3 h and 6 h), errors remained below 2 cm across most of the domain. Beyond 6 h, spatial error patterns became more heterogeneous, with deviations reaching approximately 5 cm in some areas. Increasing lead time led to progressive underestimation of spatial variability, indicating reduced sensitivity to smaller-scale features.

The instantaneous example from 08 November 2019 (Figure 20) revealed more pronounced degradation. While the broad spatial structure was captured, discrepancies increased with lead time—particularly in the eastern Gulf of Finland, where local errors reached approximately 10 cm at the 24-hour horizon. This highlights that despite robust average performance, accuracy at longer horizons varies significantly depending on specific conditions and location.

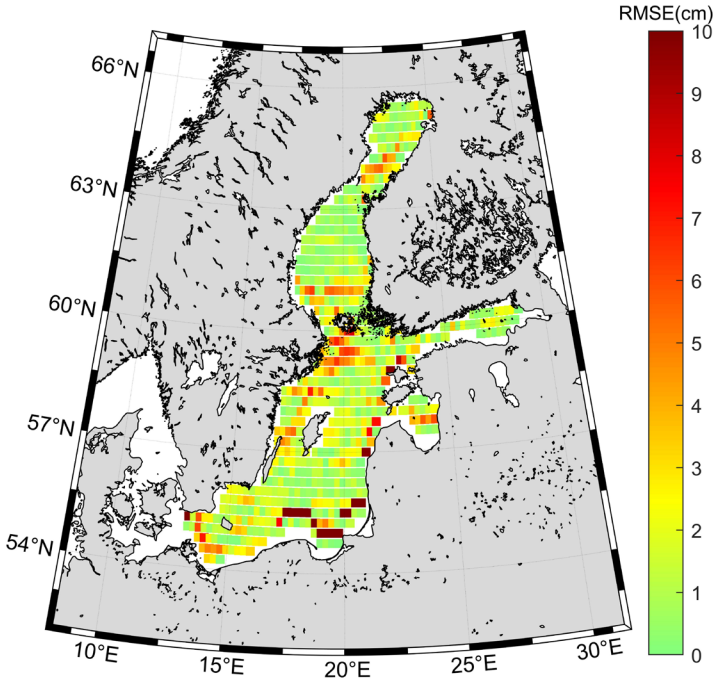


*Figure 18. The spatial performance of the proposed Conv2d model for multivariate daily sea level forecasting in the Baltic Sea (Modified from **Publication I**).*

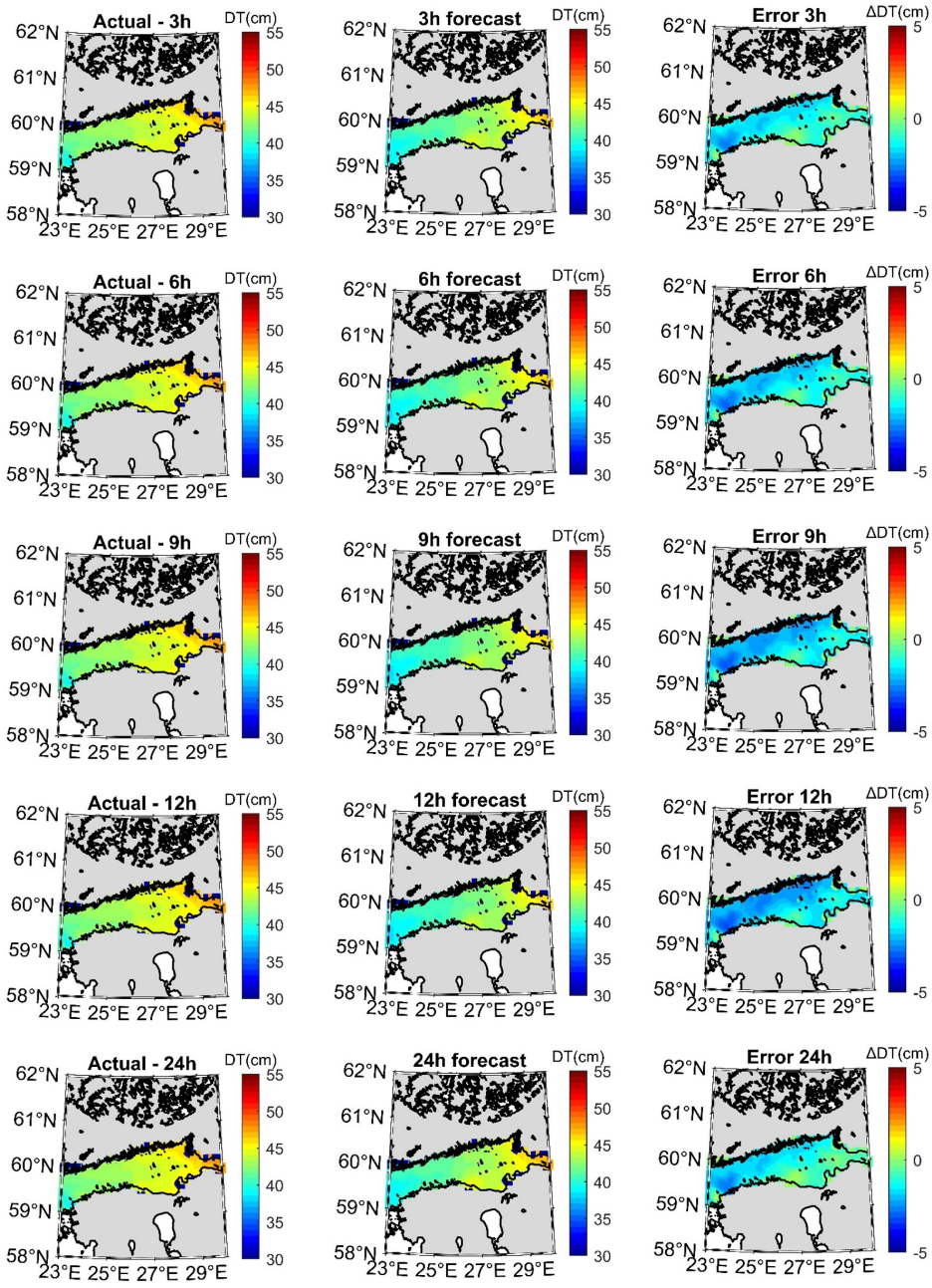


*Figure 19. Spatial performance of GRU model at different horizons (Modified from **Publication II**).*

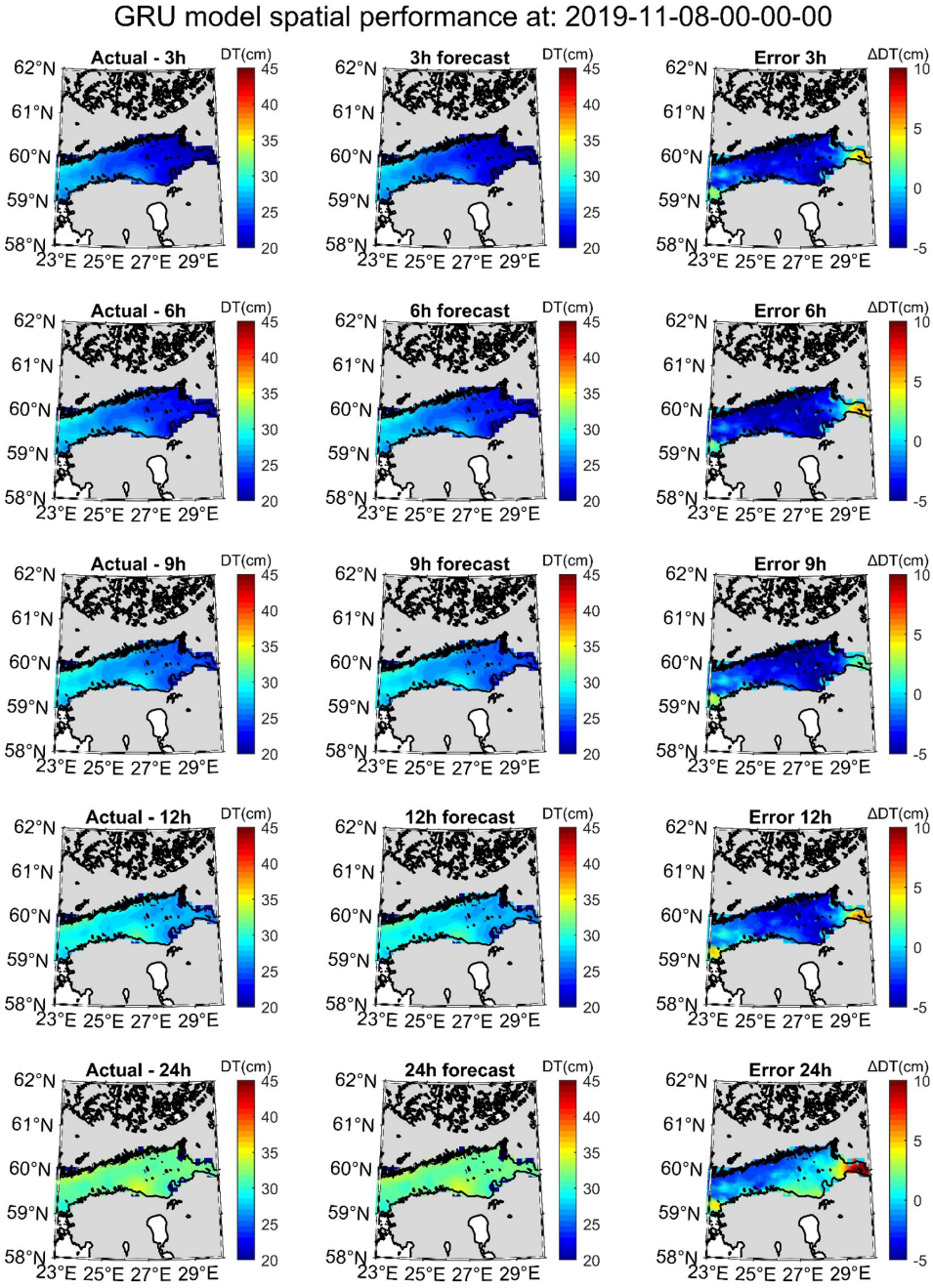


Figure 20. Instantaneous performance of GRU model for '2019-11-08-00-00-00' (Modified from ***Publication II****).*

3.5 External Validation using Satellite Altimetry

In **Publication I**, we utilized SSH data from Sentinel-3A (specifically the Baltic+SEAL product). The Baltic+SEAL dataset provides improved coastal sea level measurements compared to standard SA products (0.5–1 cm), thanks to its specialized retracking algorithm and coastal correction methods (Mostafavi et al., 2021). The algorithm for harmonizing the satellite observations with HDM data was described in **Publication I**, while Figure 21 demonstrated the implementation results and the product's enhanced accuracy in the Baltic Sea.

In **Publication II**, we extended this validation approach by comparing the proposed GRU model's forecasts with independent data from both Sentinel-3A and Sentinel-3B missions (provided by EUMETSAT). Unlike **Publication I**, which used Baltic+SEAL, the new validation method covered a different testing period. To address noise in the SA data—primarily caused by land contamination in coastal zones (where radar echoes mix land and sea signals)—we applied a Median Absolute Deviation (MAD)-based outlier detection method. This filtering step removed spurious observations while preserving valid sea level signals.

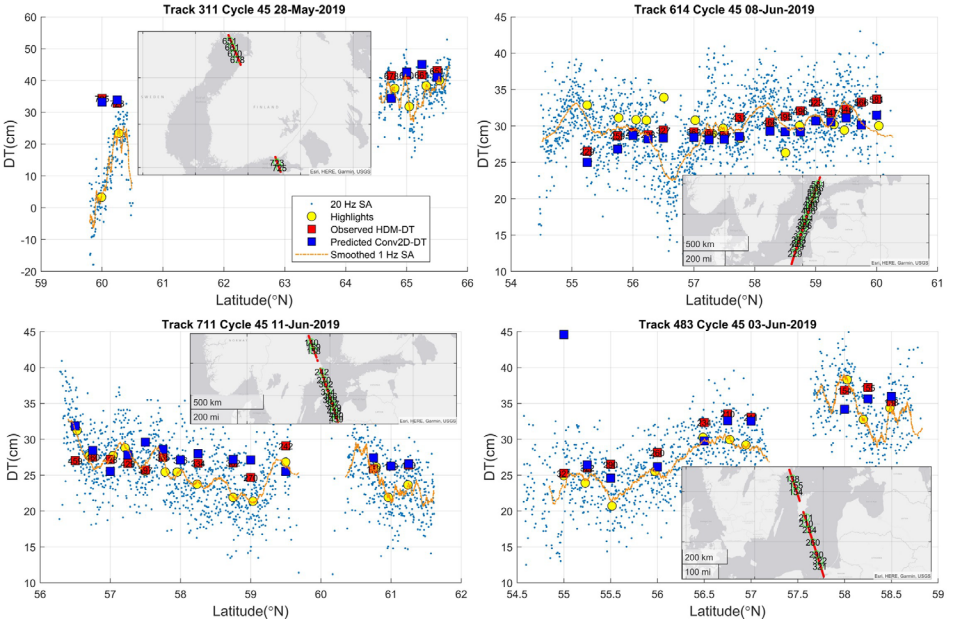


*Figure 21. Conv2d model for multivariate daily sea level forecasting in the Baltic Sea (Modified from **Publication I**).*

The results of this validation are presented in Figure 22 where for the selected Gulf of Finland tracks, SA measurements show stronger high-frequency variability than HDM or GRU DTs, reflecting instantaneous sea-surface conditions, while the HDM represents predominantly lower-frequency sea-level variations. GRU forecasts and HDM DTs agree well with SA DTs, with discrepancies below 5 cm for tracks most of the tracks. Larger errors (10–15 cm) occur on some other tracks, likely due to HDM's limitations in capturing ocean dynamics.

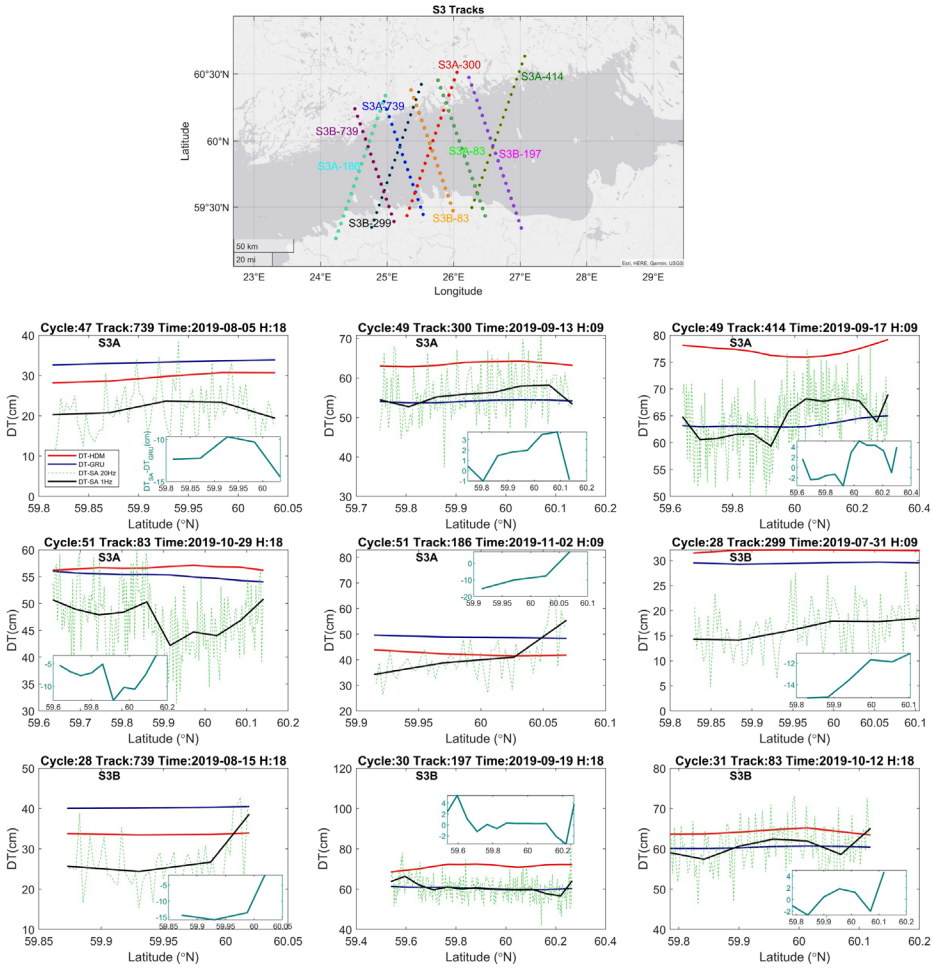


Figure 22. Validation of GRU model against Sentinel 3A and 3B satellite altimetry measurements for 3-hours ahead sea level forecasting.

4 ML for Sea Level Maxima Forecasting

Forecasting daily SLM is a critical task in coastal risk assessment, as ESLs are directly associated with coastal flooding, infrastructure damage, and socio-economic impacts. The fundamental problem addressed in this Section is how to develop a robust and reliable ML framework that can accurately forecast short-term daily SLM particularly during extreme events, while overcoming systematic model bias, data imbalance, and limited explainability.

Previous investigations presented in **Publications I** and **II** revealed that the proposed data-driven models frequently exhibited systematic underestimation during extreme sea-level events. This behaviour was evident in the corresponding performance analyses and graphical results (cf. Figures 14 and 15) and was primarily attributed to the inherent tendency of ML models to learn dominant patterns in the data. Because extreme events are rare compared to normal conditions, the models became biased toward typical sea-level behaviour rather than adequately representing infrequent but high-impact extremes. This limitation was further exacerbated in the high-resolution hourly forecasting framework adopted in **Publication II**, where the dataset was strongly imbalanced, a challenge also discussed in related literature such as Ramos-Valle et al. (2021).

Forecasting SLM is further complicated by the multivariate and physically driven nature of extreme events. Extremes Sea-levels are not governed by a single variable but rather by the combined influence of atmospheric and oceanographic drivers, including storm surges, wave conditions, and large-scale climate indices.

Considering these challenges, the primary objective of this Section is to rigorously evaluate the performance of different ML models under extreme conditions. To address the previously observed limitations, the study focuses specifically on forecasting daily SLM rather than daily means or hourly levels. A multivariate forecasting framework is adopted, incorporating an extended set of physically meaningful predictors, including zonal and meridional wind speed, wind gust, atmospheric pressure, precipitation, evaporation, SWH, and the Baltic Sea Index (BSI). BSI is defined as the difference in normalized pressure anomalies between 53°30′N, 14°30′E (Szczecin, Poland) and 59°30′N, 10°30′E (Oslo, Norway) (Lehmann et al., 2002). These predictors were selected based on domain knowledge and EDA to ensure physical relevance and statistical robustness.

The modelling framework is expanded to include a diverse range of approaches. Baseline ML models such as MLP, XGB, and RF are considered alongside hybrid DL architectures, including CNN–LSTM and CNN–GRU models. To ensure fair comparison and robust evaluation, state-of-the-art optimization techniques, such as BO, is employed to identify optimal hyperparameter configurations for each modelling framework. This ensures that performance differences reflect model capability rather than suboptimal tuning.

The temporal coverage of the training dataset is also extended through the incorporation of long and consistent TG records (1971–2022). Furthermore, the final proposed modelling framework is designed to be explainable, addressing a key limitation of the DL models presented in **Publications I** and **II**. Improving interpretability strengthens confidence in the forecasting system and provides insights into the relative contribution of different physical drivers.

The overall goal is to generate reliable short-term forecasts of daily SLM and subsequently apply extreme value analysis for long-term risk assessment. Model performance is evaluated using multiple statistical metrics, while peak event detection capability is assessed using a separate set of threshold-based performance measures (cf. Table 4 for the metrics equations). The following Sections provide detailed descriptions of extreme patterns in the Baltic Sea, input feature selection, model configuration, performance evaluation, explainability analysis, and extreme value analysis.

The main novelties for which the results are presented in **Publication III**, include:

- The development of a physically informed multivariate forecasting framework that integrates atmospheric and oceanographic drivers, including wind components, wind gust, atmospheric pressure, precipitation, evaporation, SWH, and the BSI.
- A rigorous and fair comparison of heterogeneous modelling approaches, including baseline ML models (MLP, RF, XGB) and hybrid DL architectures (CNN–LSTM and CNN–GRU).
- The application of BO for hyperparameter tuning across all models to ensure unbiased performance comparison and optimal model configurations.
- The extension of the temporal training dataset using long-term TG records to improve representation of rare extreme events and enhance model generalization.
- To account for spatial heterogeneity in sea-level drivers, station-specific models were developed, trained, and evaluated independently for each tide-gauge location.
- The integration of explainability techniques within the final modelling framework, addressing the lack of explainability in previous DL models and improving transparency and physical consistency.
- The implementation of a two-stage modelling strategy, combining short-term ML-based forecasts of daily SLM with subsequent extreme value analysis for long-term coastal risk assessment.

Figure 23 shows the adopted strategy for daily SLM forecasting using ML approaches in the selected Baltic Sea TG stations. A comprehensive description of the methodology and results is available in **Publication III** and the corresponding Appendices. A dedicated focus on forecasting daily sea-level maxima under extreme conditions, directly addressing the systematic underestimation of peak events identified in **Publications I** and **II**.

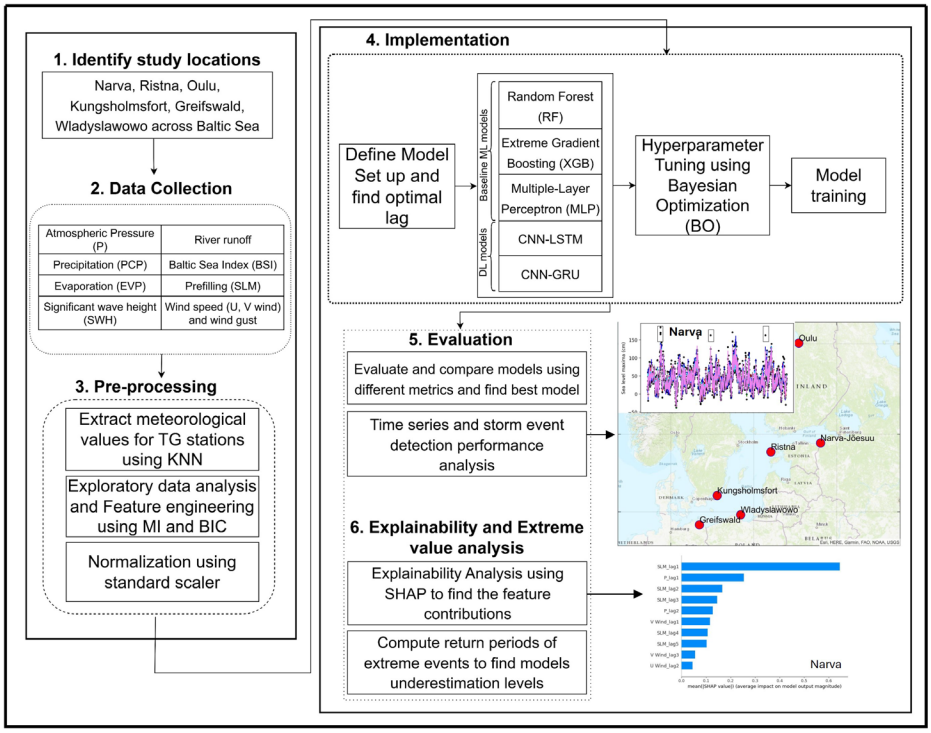


*Figure 23. Diagram for daily SLM forecasting using MLP, XGB, RF, CNN-LSTM and CNN-GRU models (from **Publication III**).*

4.1 Extreme Patterns in the Baltic Sea

For SLM forecasting, sea-level measurements were collected from multiple TG stations (Narva, Ristna, Oulu, Kungsholmsfort, Greifswald and Władysławowo) covering the period from 1 January 1971 to 31 December 2022. The selection of these TG stations was based on data availability, minimal data gaps, spatial coverage across different parts of the Baltic Sea region, and documented evidence of extreme conditions reported in previous studies. VLM corrections were not applied to the measurements, as the primary objective was to analyze the relative sea-level values as observed in order to study extremes. However, to ensure consistency across stations, all sea-level measurements were converted to a common reference using the procedure described in **Publication III**, aligning the data with the BSCD2000 geoid model. The time periods used for the training, validation, and test datasets for each station are illustrated in Figure 24a. It is also worth noting that a separate validation set was used in **Publication III**. Since the performance of different models was being optimized using the BO method, this validation set served as a testbed to evaluate and identify the optimal hyperparameter settings.

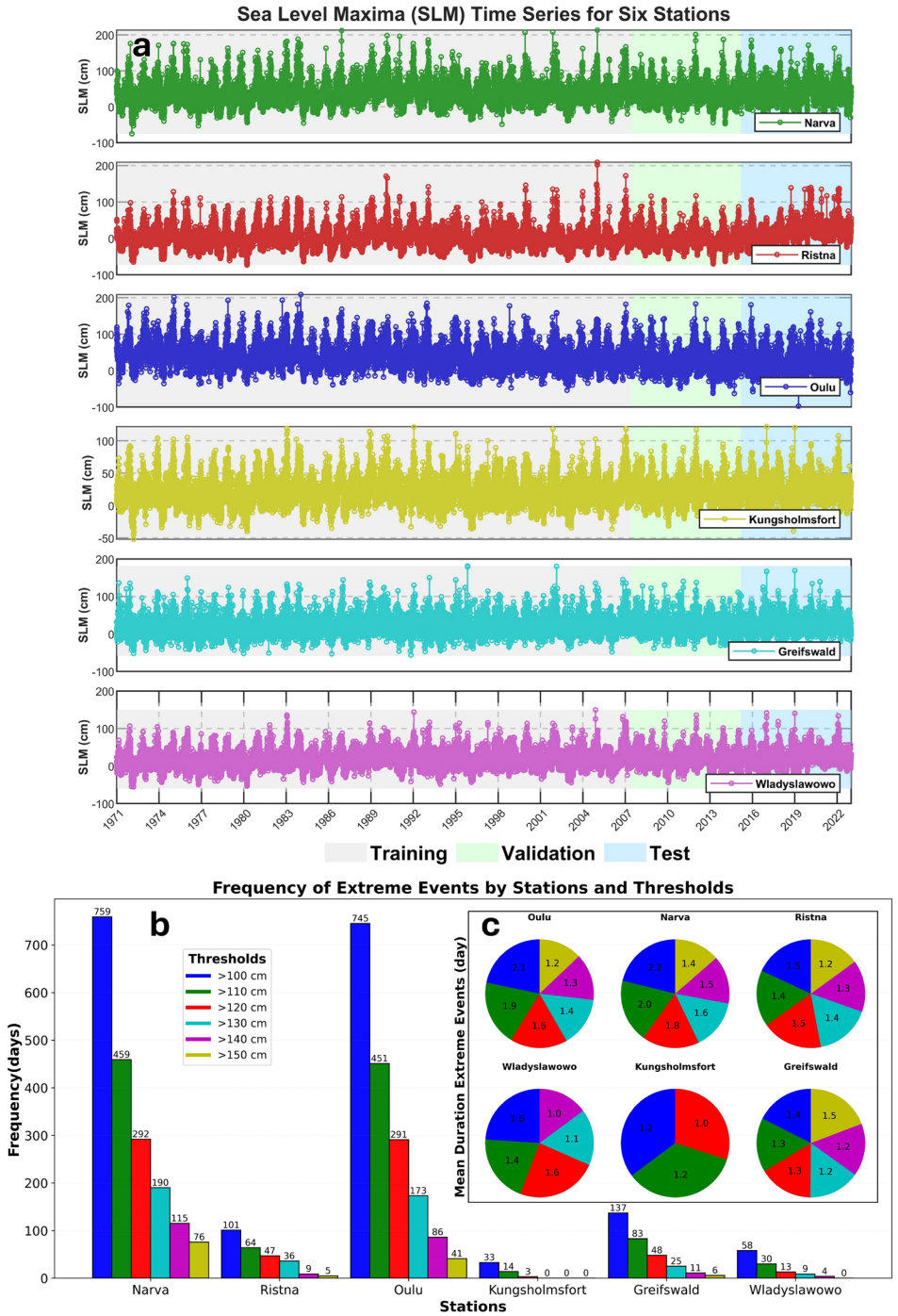


*Figure 24. a) Training, validation and test sample splitting for the studied TG stations, along with b) Pattern of SLM including duration and number of detected extremes based on different thresholds from 100 to 150 cm (Modified from **Publication III**).*

To investigate extreme patterns in SLM, different thresholds ranging from 100 cm to 150 cm were analyzed to examine how the frequency of extreme events varies with threshold levels, as shown in Figure 24b. For instance, the Narva and Oulu stations, located in the eastern and northern Baltic Sea, exhibited the highest number of extreme events, followed by Greifswald, Ristna, Władysławowo, and Kungsholmsfort. These results are consistent with previous studies conducted in the Baltic Sea (Wolski et al., 2014).

As illustrated in Figure 24b, very high peaks exceeding 150 cm were rare, except for the Narva station. In addition, the duration of extreme events was compared across different thresholds and is presented in Figure 24c. The average duration of events above 100 cm ranged from 1.2 to 2.2 days across the various stations.

A decadal analysis of ESL, based on thresholds of 100 cm and 150 cm, was also conducted and is presented in Figure 25a for different stations. For the northeastern stations at the 100 cm threshold, the number of extreme events increased from the 1970s to the 1980s, then declined until a resurgence in the 2010s. At the 150 cm threshold, decadal patterns of extremes were more variable across stations. However, for Narva and Oulu—the most vulnerable stations—the period from 1971 to 1990 experienced more extreme events compared to subsequent decades.

Seasonal analysis of extremes, shown in Figure 25b, indicates that the rarest extremes occur during winter and autumn, with few events detected during spring and summer. The highest sea-level peaks, exceeding 200 cm, were observed at the Narva, Ristna, and Oulu stations.

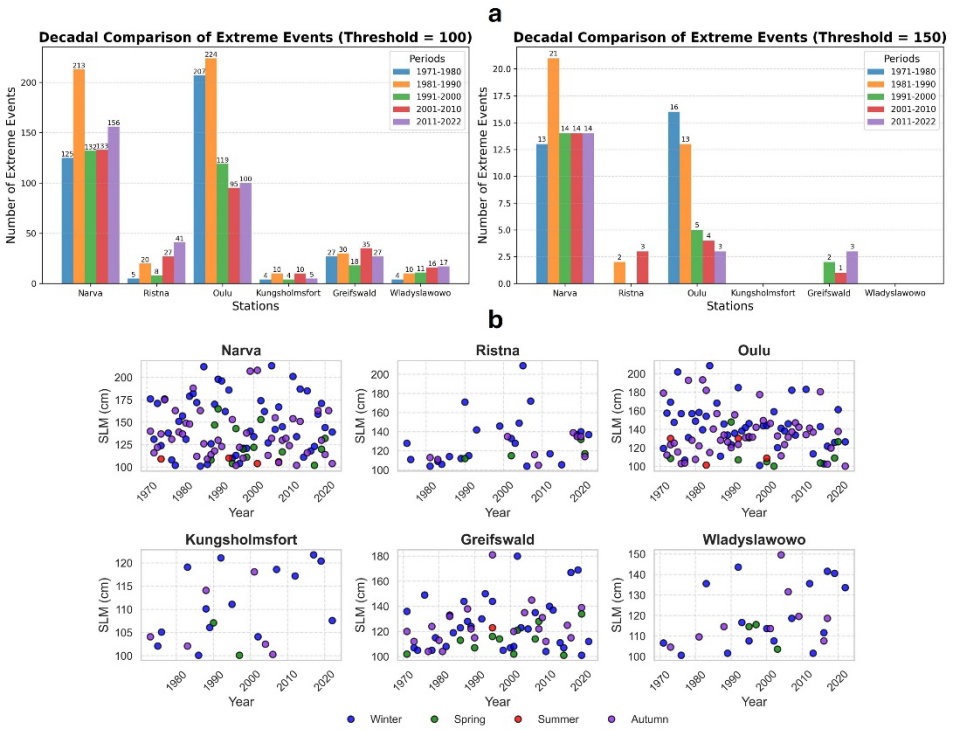


*Figure 25. a) Decadal number of detected extremes along with b) the seasonal patterns of the extremes based at different stations during different periods (Modified from **Publication III**).*

4.2 Feature Selection

For daily SLM forecasting in the Baltic Sea, we extended previously selected features by including indicators such as wind speed components (zonal, meridional and gust), SLP, river runoff, SWH, evaporation, precipitation, BSI and daily SLMs from six TG stations in the Baltic Sea based on both domain knowledge and EDA. The sources of each data source were mentioned in **Publication III**.

The original data were provided at an hourly resolution, but we extracted daily maxima for all variables except for pressure (given the inverse barometric effect) daily minimum was used instead. We also assessed the potential of other engineered features like incorporating wind speed adjusted for coastline orientation suggested in literature (Guillou & Chapalain, 2021); However, this variable did not significantly improve the results, so we excluded it from the final model.

For EDA, we conducted an MI-based analysis. The MI index captures both linear and nonlinear dependencies, offering key advantages over the PCC method used in earlier studies (Pak et al., 2020). Unlike PCC, which only measures linear relationships, MI identifies complex, non-monotonic interactions common in marine processes. It is also robust to non-Gaussian data distributions and outliers, making it better suited for heterogeneous climate variables.

The MI results for each station are presented in Figure 26, illustrating that different features influence SLMs across the stations. Specifically, pressure, zonal wind (uwind), vwind, and SWH were identified as the primary features due to their higher MI indices. Additionally, wind gust was included for the Narva, Oulu, and Greifswald stations, and the BSI was added for Oulu, Kungsholmsfort, Greifswald, and Władysławowo stations.

Collinearity analysis revealed a high correlation between wind gust and SWH at Ristna, Kungsholmsfort, and Władysławowo, so the wind gust variable was excluded for these stations. The effects of river runoff and evaporation minus precipitation were found to be negligible, and these variables were also excluded from the final input sets.

For stations exhibiting more extreme patterns, such as Narva and Oulu, pressure and wind speed components appeared to have a stronger influence. Furthermore, eastern stations like Narva and Ristna were more affected by SWH.

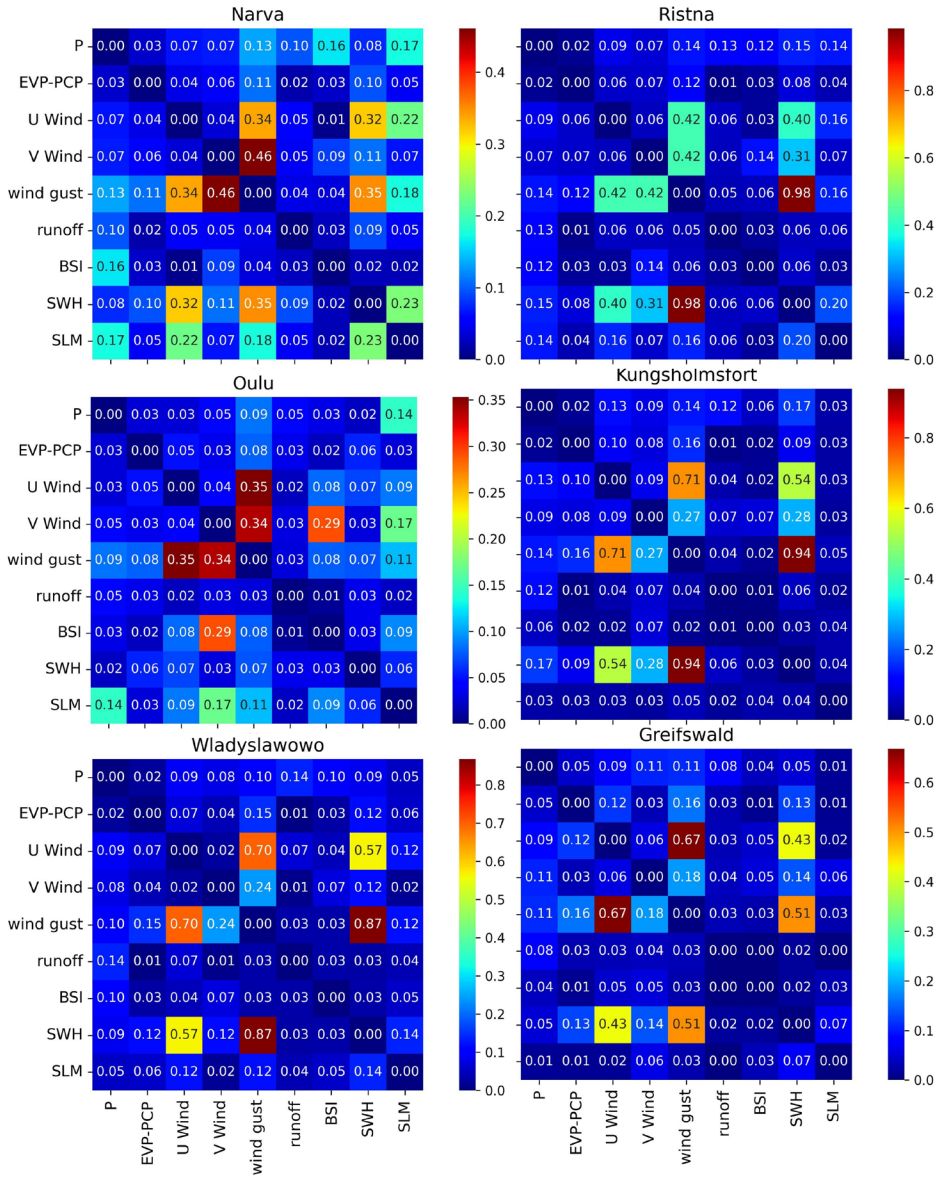


*Figure 26. MI index as the feature selection technique for daily SLM forecasting (Modified from **Publication III**). Please refer to the main text body for the utilized symbols.*

4.3 Results of Bayesian Optimization for Hyperparameter Tuning

To ensure a fair comparison of different ML models, each must be optimized to its best possible performance. Common hyperparameter tuning methods include random search, grid search, genetic algorithms, and sequential feature selection. For the SLM forecasting, we employed BO, which offers distinct advantages over traditional approaches. The selected ranges and final optimal hyperparameter value by the BO method for each model and station have been displayed in **Publication III** (Table S5).

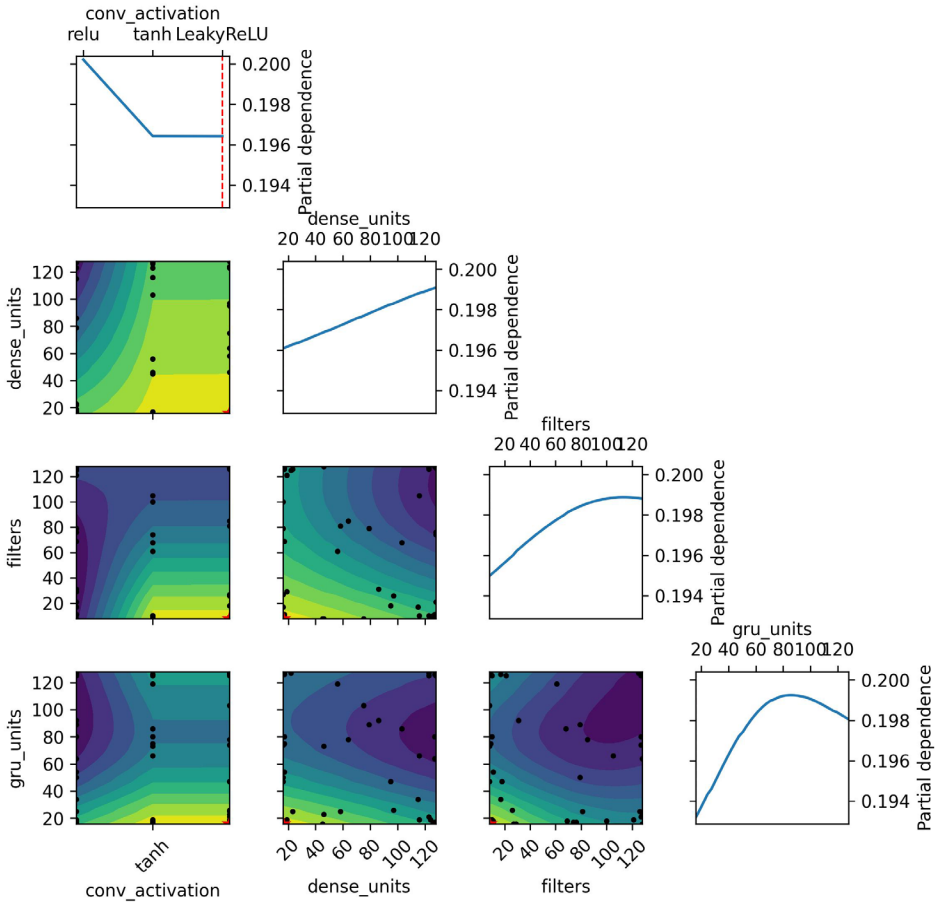


Figure 27. BO optimization results for the CNN-GRU model, computed for Narva station as an example.

Unlike brute-force methods like grid or random search, BO uses probabilistic models to intelligently guide the search toward promising hyperparameters, significantly reducing computational effort. It balances exploration of new parameter combinations with exploitation of known high-performing regions, enabling faster convergence to optimal configurations—particularly valuable in complex, nonlinear problems like sea level modeling. This approach efficiently navigates high-dimensional parameter spaces where simpler methods often struggle, ensuring robust model performance without exhaustive trial-and-error. Moreover, a visual example of BO method implementation for CNN-GRU model at Narva station was depicted in Figure 27.

4.4 Models' Performance

Figure 28 summarizes the training and test performance of the evaluated ML (RF, MLP, XGB) and DL (CNN-LSTM, CNN-GRU) models across six TG stations using RMSE and R^2 metrics. Overall, the results indicate spatial variability in model performance.

During training, RMSE values ranged from about 4 to 13 cm, with the smallest errors (<8 cm) at Kungsholmsfort and Władysławowo, and larger errors at Narva, Oulu, and Greifswald due to more complex local dynamics. Overall model performance was strong,

with training R^2 values mostly above 0.80, and XGB and CNN-based models showing the best fit by effectively capturing nonlinear patterns.

Test-period performance followed similar spatial patterns but with increased RMSE values, as expected. The best generalization performance was achieved at Kungsholsfort, Władysławowo, and Ristna, where test RMSE values typically ranged from 7 to 10 cm and R^2 values exceeded 0.80 for most models. These stations consistently showed smaller performance gaps between training and testing, indicating good model robustness. Conversely, Narva, Oulu, and Greifswald exhibited higher test RMSE values—reaching up to 14–15 cm at Oulu—along with lower R^2 scores, highlighting the greater difficulty in forecasting sea levels at these locations.

Across all stations and metrics, the DL models outperformed the traditional ML approaches. In particular, the CNN-GRU model consistently achieved the lowest or near-lowest RMSE values during the test period and maintained competitive R^2 scores, indicating superior generalization ability. The CNN-LSTM model also performed strongly, though with slightly higher errors than CNN-GRU in several stations. Among the ML models, MLP generally performed better than RF and XGB.

Overall, Figure 28 demonstrates that incorporating convolutional feature extraction with recurrent temporal modelling, especially through the CNN-GRU architecture, provide clear advantages for daily sea level forecasting. The results also emphasize the importance of station-specific characteristics, as model performance is strongly influenced by local sea level variability and dynamics.

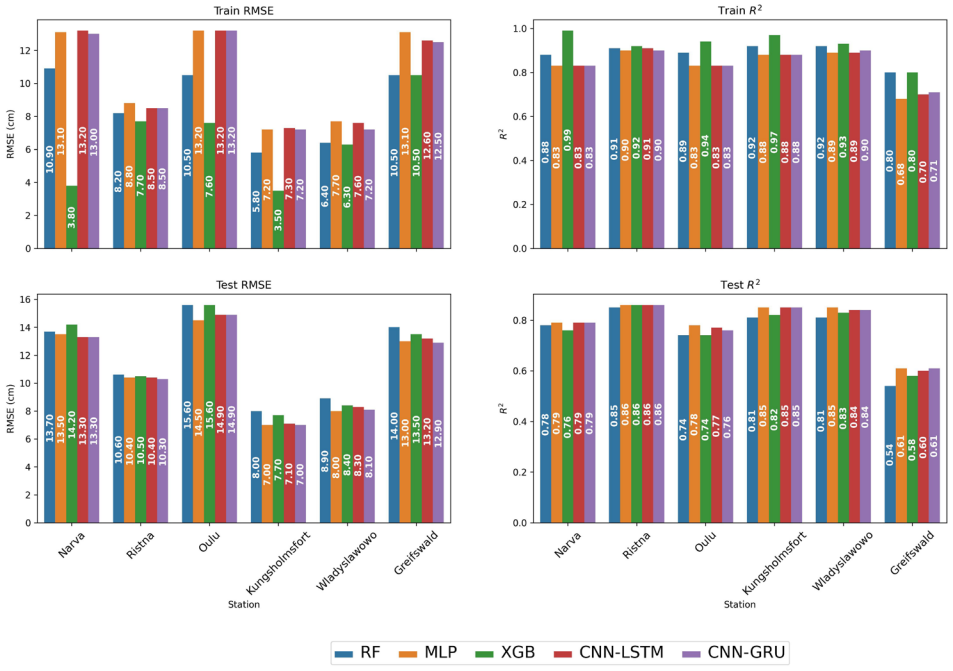


*Figure 28. General performance of employed ML and DL models for daily SLM forecasting (Modified from **Publication III**).*

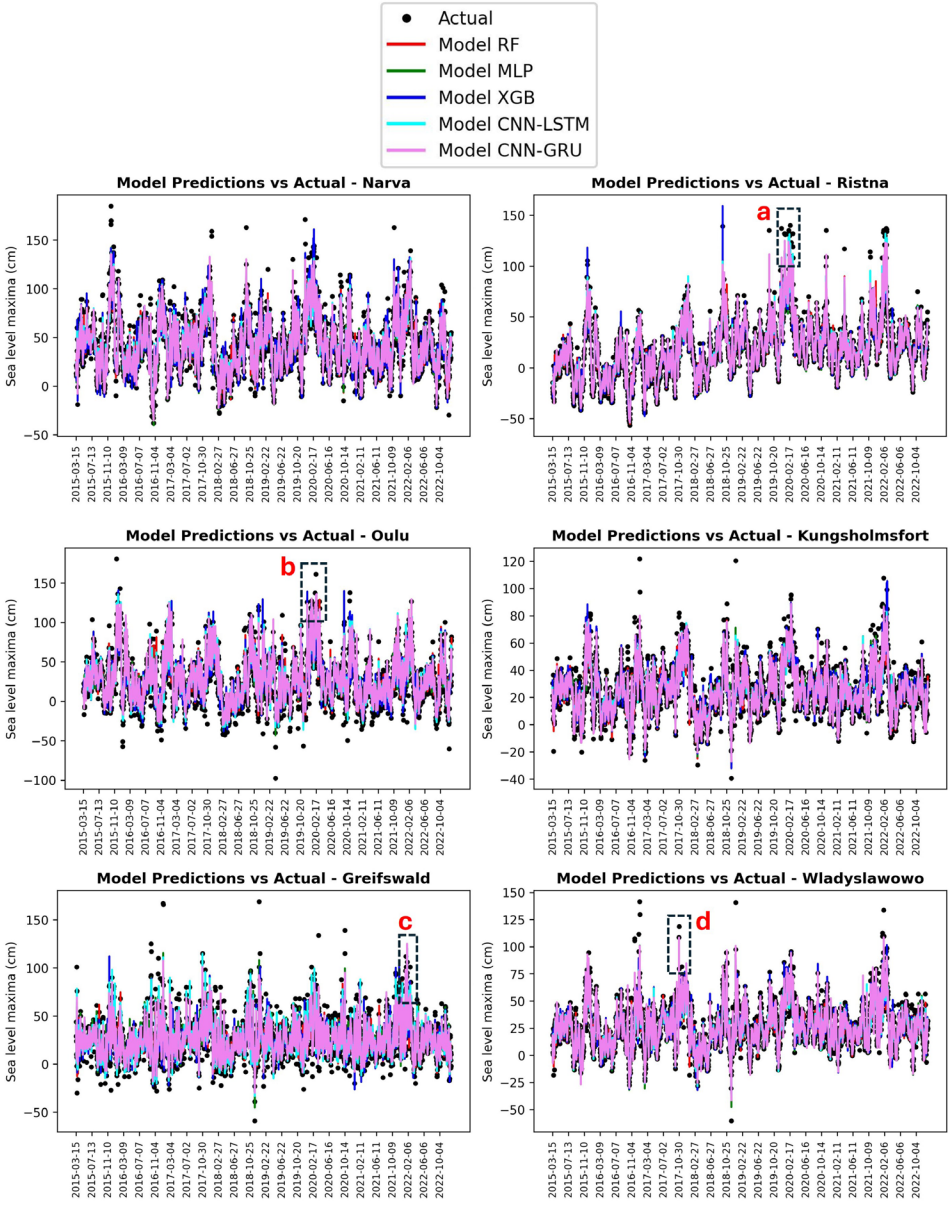


*Figure 29. Time series forecasting of different models vs. the actual daily SLMs at studied TG stations (Modified from **Publication III**).*

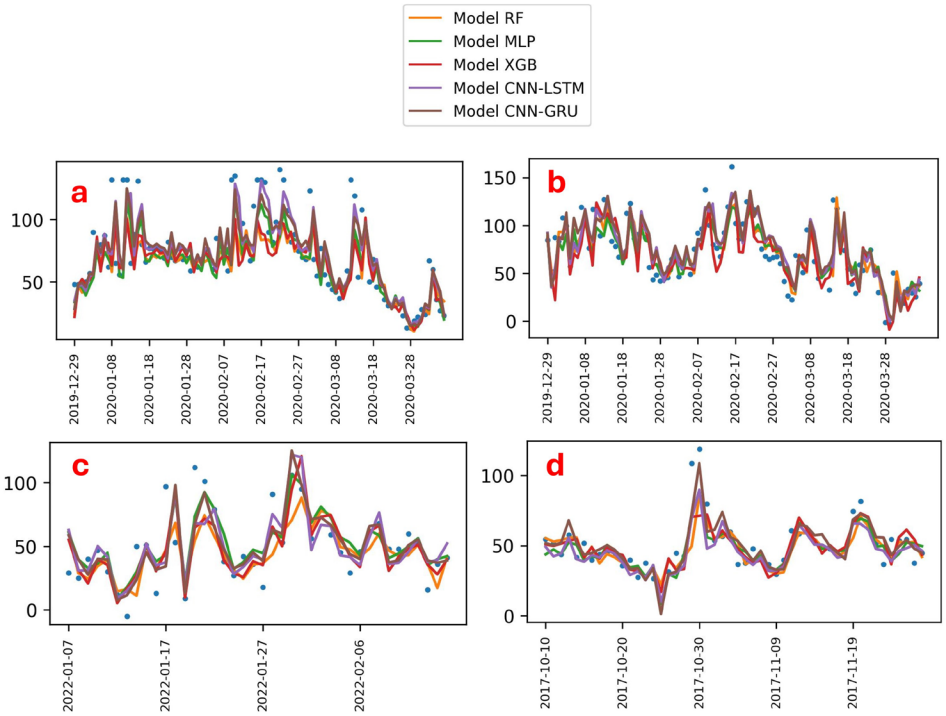


*Figure 30. Same as Figure 29, but for the magnified regions a-d (Modified from **Publication III**).*

The time-series performance of the models during the test period was evaluated across multiple TG stations, as illustrated in Figure 29. Overall, all models were able to reproduce the general temporal evolution of SLM, capturing both short-term fluctuations and longer-term variability. This indicates that the input features and model structures were adequate for learning the dominant sea level dynamics at the daily scale. Among the evaluated approaches, the DL models—particularly the CNN-GRU—demonstrated superior performance in tracking observed SLM patterns across most stations.

A closer inspection of the station-specific plots in Figure 30 reveals that the neural network models (CNN-GRU, CNN-LSTM, and MLP) consistently followed the observed peaks more closely than the traditional ML models (RF and XGB). The convolutional layers enabled effective extraction of local temporal features, while the recurrent components improved the representation of sequential dependencies. The CNN-GRU model, in particular, exhibited a stronger ability to adapt to abrupt rises and falls, producing predictions that aligned more tightly with the observed time series.

Despite this overall strong performance, all models struggled to fully capture rare extreme events. As highlighted in Figure 30, extreme SLM values exceeding 150 cm at the Narva, Greifswald, and Oulu stations, as well as those above 120 cm at the Kungsholmsfort and Władysławowo stations, were systematically underestimated. These discrepancies are especially visible during isolated high-magnitude peaks. This behavior is consistent across both ML and DL models, although the underestimation is less pronounced for the CNN-GRU model.

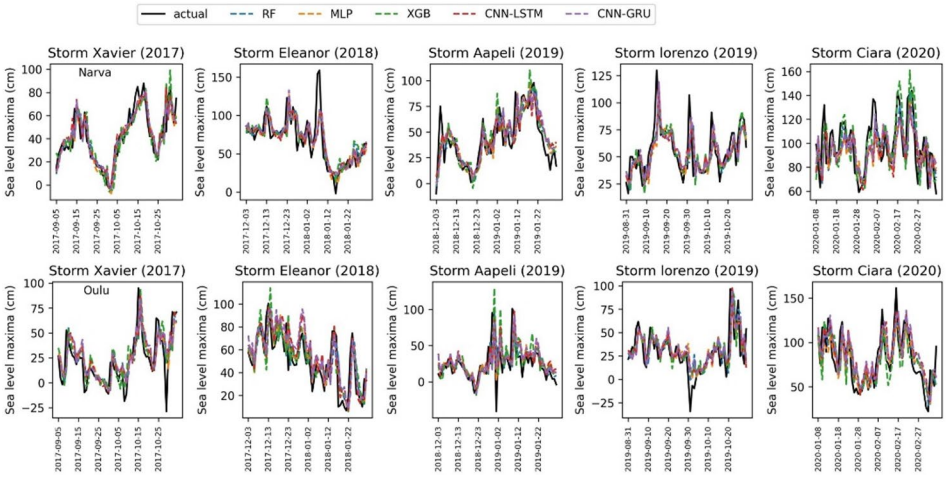


*Figure 31. Comparison of the employed ML and DL models at storm events detection for Narva and Oulu stations (Modified from **Publication III**).*

To further evaluate the ability of the models to detect and reproduce ESL behaviour during storm conditions, their performance was examined at two highly vulnerable TG stations—Narva and Oulu—during major storm events in the test period, namely Xavier (2017), Eleanor (2018), Aapeli (2019), Lorenzo (2019), and Ciara (2020). The time-series comparisons between observed and forecasted SLM for these events are presented in Figure 31.

Overall, the models were able to capture the timing and general evolution of storm-induced sea level fluctuations at both stations. In particular, the onset, duration, and decay phases of the storm surges were reasonably well reproduced, indicating that the models learned the dominant temporal signatures associated with extreme meteorological forcing.

At the Narva station, the CNN-GRU and CNN-LSTM models tracked rapid sea level rises and short-lived peaks more accurately than the traditional ML models. For storms such as Eleanor (2018) and Ciara (2020), these DL models were better able to follow the sharp intensification and subsequent relaxation of sea levels, whereas RF and MLP occasionally exhibited smoother responses and slight temporal lag. XGB also showed some overestimation patterns for these storm events. Nevertheless, even for the CNN-based models, the highest storm peaks (above 150 cm)—particularly during Eleanour (2018) were still partially underestimated.

A similar pattern was observed at the Oulu station, where all models successfully captured the overall storm-driven variability but showed increasing difficulty in reproducing high-magnitude peaks. The underestimation was most evident during Ciara (2020), when observed maxima rose rapidly over short time intervals. Although the CNN-GRU model again provided the closest match to the observations, peak amplitudes remained slightly damped, reflecting the inherent challenge of predicting extreme surges with limited historical examples.

The systematic underestimation of storm peaks across both stations can be attributed to two main factors. First, extreme storm events represent a small proportion of the overall dataset, reducing their influence during model training. Second, storm-driven sea level extremes are strongly influenced by localized atmospheric and hydrodynamic

processes—such as wind direction, pressure gradients, and ice effects at higher latitudes—that may not be fully represented in the predictor set.

In summary, figure 31 demonstrates that while all models are capable of detecting storm events and reproducing their temporal structure, DL models—especially CNN-GRU—offer clear advantages in capturing the intensity and dynamics of storm-induced SLM. However, improving the prediction of peak magnitudes during rare extreme events remains a key challenge and highlights the need for targeted strategies.

4.5 CNN-GRU Model's Explainability

For the best-performing forecasting approach (CNN-GRU model), we applied SHAP-based explainability to reveal how key variables affect SLM predictions. Figure 32 presents the SHAP-based explainability analysis of the CNN-GRU model across six TG stations including Narva, Kungsholmsfort, Ristna, Greifswald, Oulu, and Władysławowo, along with the averaged feature importance and percentage contribution of each predictor to the total model explainability shown in Figure 33.

The results provide valuable insights into the dominant drivers of SLM prediction and the relative influence of prefilling, atmospheric pressure, wind components, and wave-related variables.

Across all stations, lagged sea level maxima (SLM_lag1) emerge as the most influential predictor by a large margin. At individual stations, its mean absolute SHAP value reaches approximately 0.64 at Narva, 0.85 at Ristna, 0.70 at Władysławowo, and remains dominant at Kungsholmsfort, Greifswald, and Oulu. This highlights the strong temporal persistence in SLM behavior and confirms that recent past sea levels or prefilling play a critical role in forecasting current maxima. The aggregated analysis across the stations in Figure 33 further reinforces this finding, with SLM_lag1 alone accounting for about 39.3% of the total explainability, making it the single most important contributor to the CNN-GRU model predictions.

Atmospheric pressure lags (P_lag1 and P_lag2) are consistently ranked among the top predictors at all stations, though their relative importance varies spatially. For example, pressure-related features are recognized as the dominant meteorological feature at Narva, Ristna and Władysławowo, reflecting regional differences in storm surge sensitivity to pressure gradients. When averaged across all stations (Figure 33), pressure predictors together contribute more than 22% of the total explainability (13.2% for P_lag1 and 8.7% for P_lag2), underscoring their critical role in SLM variability.

Higher-order sea level lags (SLM_lag2 and SLM_lag3) also contribute meaningfully to model predictions. These features capture longer memory effects and multi-day persistence in sea level dynamics, particularly at stations with smoother hydrodynamic responses such as Kungsholmsfort and Władysławowo. In the averaged explainability results, SLM_lag2 and SLM_lag3 account for approximately 10.7% and 6.0% of the total contribution, respectively, confirming that multi-lag temporal dependencies are effectively leveraged by the CNN-GRU architecture.

Wind-related variables, including both zonal (U) and meridional (V) components, exhibit non-negligible importance across all stations. Their influence is particularly noticeable at Narva, Ristna and Greifswald. Collectively, wind predictors contribute around 16% of the total explainability, with V_Wind_lag1 and U_Wind_lag1 being the most influential among them. This suggests that wind direction and persistence are key secondary drivers of SLM variability, especially during extreme weather conditions.

Wave-related features (SWH lags), while generally less dominant, still appear among the top ten predictors at several stations, particularly Narva and Ristna. Their relatively smaller contribution (approximately 5.9% in the averaged analysis) indicates that wave effects play a supporting role, potentially enhancing sea level extremes during storm events but exerting less influence on daily maxima compared to prefilling and atmospheric forcing.

In addition to local atmospheric and oceanic predictors, the BSI also contributes meaningfully to the CNN-GRU model explainability. BSI-related lags, particularly BSI_lag1 and BSI_lag2, consistently appear among the top-ranked features at several stations, with more pronounced influence at Greifswald, Oulu, and Władysławowo. This highlights the importance of basin-scale atmospheric circulation patterns in modulating local SLM, beyond the effects of site-specific wind and pressure forcing. BSI persistent contribution across stations indicates that incorporating regional climate indices enhances the physical consistency of the CNN-GRU model. Overall, the inclusion of BSI improves the model's representation of synoptic scale forcing and supports more robust SLM forecasting.

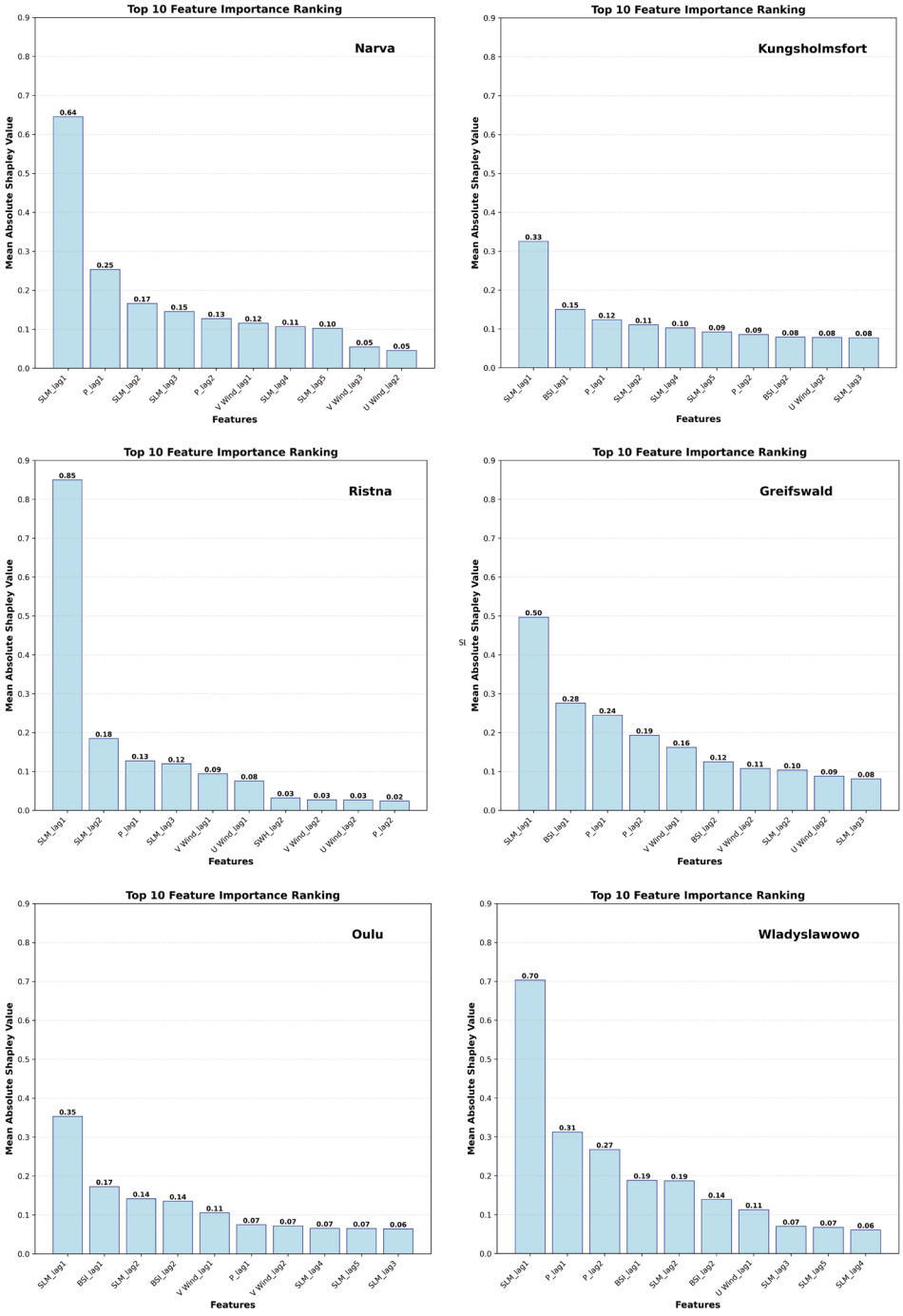


*Figure 32. SHAP analysis results of explainability for CNN-GRU model at different stations, the bars show the top ten features at each TG station (Modified from **Publication III**).*

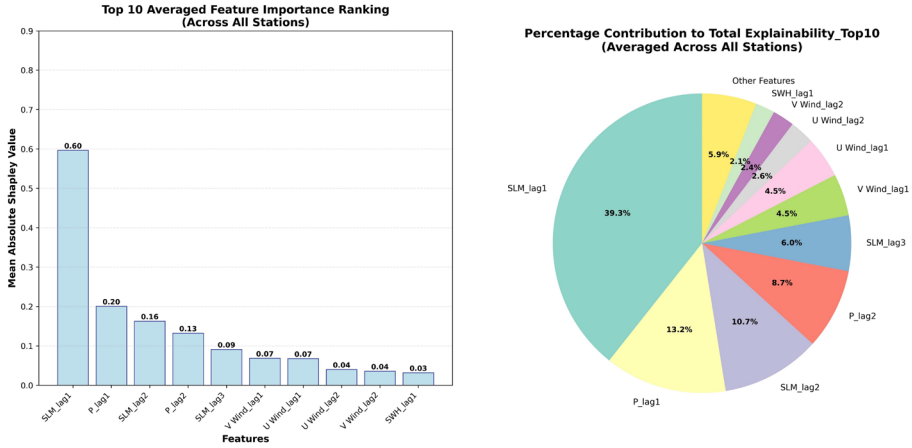


*Figure 33. Percentage of each feature's contributions on the total explainability (Modified from **Publication III**).*

4.6 Extreme Value Analysis Results

To investigate the behavior of ESL and to better understand the underestimation observed in the daily ML-based forecasts, a threshold- and return-level analysis was conducted across multiple stations and seasons, as illustrated in Figure 34. The seasonal return level curves reveal distinct spatial and temporal patterns in extremes behavior, with substantially higher return levels during winter and autumn compared to spring and summer at all stations. This seasonal contrast highlights the dominant role of storm activity and large-scale atmospheric forcing during colder months, which is consistent with the periods where the ML models showed the largest underestimation of peak events.

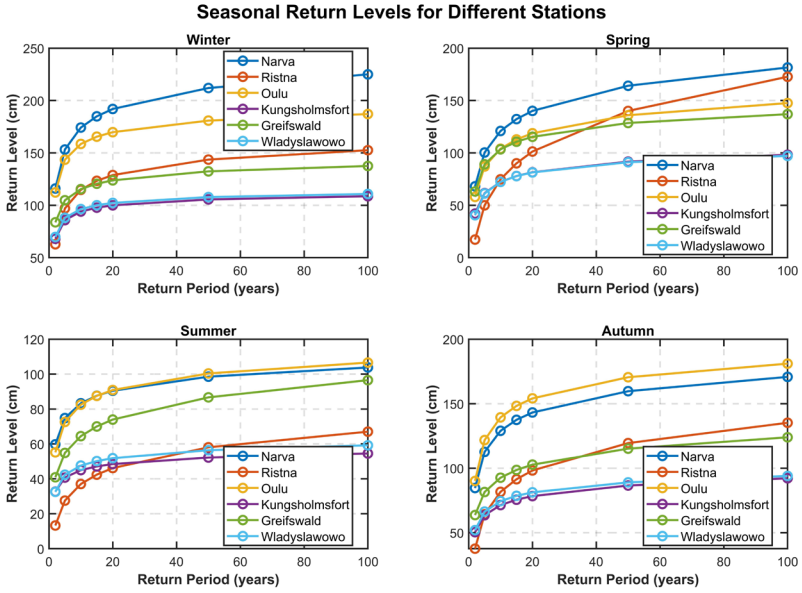


*Figure 34. Estimation of return periods of the ESL (Modified from **Publication III**).*

Stations such as Narva, Oulu, and Greifswald exhibit particularly high winter and autumn return levels, with 5–7-year return levels exceeding 150 cm in winter at Narva and approaching similar magnitudes at Oulu. These high thresholds correspond closely to the extreme peaks that were underestimated in the daily forecasting experiments, indicating that the ML models tend to dampen events that lie in the upper tail of the distribution. In contrast, stations such as Kungsholmsfort and Władysławowo show lower seasonal return levels and smoother growth with increasing return period, which aligns with the comparatively better performance and lower underestimation observed in the daily predictions at these locations.

5 Discussions and Conclusions

5.1 Discussion and Conclusions

Data-driven approaches, including ML, offer potential improvements in sea level forecasting accuracy while enhancing computational efficiency and cost-effectiveness. However, several challenges limit current ML-based forecasting: integrating heterogeneous observational data sources, achieving the spatiotemporal resolution and accuracy needed to capture localized dynamics, accurately representing extreme events, and overcoming the "black-box" nature of conventional models that hinders explainability. This dissertation addresses these challenges by exploring multiple ML-based methods for sea level forecasting in the semi-enclosed Baltic Sea region, characterized by strong stratification and complex oceanographic dynamics (Wulff et al., 2001).

The first method focused on spatiotemporal sea level forecasting using deep neural network models, including CNN (or conv2d), LSTM, and GRU. A CNN model was developed to provide the first regional forecasts of geoid-referenced sea levels (DT) across the Baltic Sea. CNN models perform well for grid-like data and were prioritized for their ability to capture spatial patterns and dependencies across the study domain. The model forecasts daily sea levels at 0.25° × 0.25° resolution using predictors—SST, zonal and meridional wind components, SLP, and DOY—selected through prior knowledge and statistical analysis.

The CNN model demonstrated strong spatial and temporal performance. Spatially, it captured sea levels accurately across most locations, with RMSE mostly below 5 cm. Temporally, it modeled sea levels up to 100 cm and tracked trends and variations effectively. Overall, the model achieved an RMSE of 4.1 cm and an R^2 of 0.91. However, performance was lower in the eastern and southeastern Baltic Sea (~RMSE of 10 cm), likely due to the coarse input resolution limiting representation of small-scale processes (Ishida et al., 2020). For SLM (>70 cm) events in the Gulf of Riga, accuracy dropped to approximately 73%. Analysis showed the CNN captured storm surges driven by rapid pressure drops and wind intensification, while missed events likely resulted from unrepresented mechanisms such as river discharge or wave setup.

To achieve the spatiotemporal resolution needed for capturing small-scale coastal dynamics, LSTM and GRU models with sequence-to-sequence architectures were developed for multistep-ahead forecasting. These models were selected for their ability to process sequential data, capture both short- and long-term sea level trends, and mitigate vanishing or exploding gradients through their gated structures. The models forecasted sea levels 3, 6, 9, 12, and 24 hours ahead at hourly intervals with a high spatial resolution of one nautical mile in the Gulf of Finland. This region exhibits sea level variability over very short spatial scales due to a small internal Rossby radius (approximately 2–4 km) (Soomere et al., 2008) and rapid fluctuations in local storm forcing, making it prone to extremes (Weisse et al., 2021). Resolutions of 0.5–1 nautical mile (≈1/3–1/2 of the Rossby radius) are therefore necessary to capture such dynamics and improve forecast accuracy (Kowalewski & Kowalewska-Kalkowska, 2017). Applying ML/DL models at this resolution was unprecedented in sea level forecasting.

Both LSTM and GRU models performed well for normal sea level conditions, generally up to 80–90 cm, with GRU models showing slightly superior performance. The GRU achieved superior results with an average RMSE of 4.96 cm and R^2 of 0.93 to forecast

between 3 and 24 hours ahead. The GRU's superior performance may be attributed to its simpler architecture with fewer parameters, which could facilitate optimization and improve generalization. Despite this, both models underestimated extreme events, particularly in the eastern Gulf of Finland, highlighting the challenges of forecasting in regions with highly nonlinear and ESL dynamics.

The spatiotemporal models (CNN, LSTM and GRU) consistently underestimated extremes because data-driven approaches are inherently biased toward normal conditions, with extreme events being rare in training data (Ramos-Valle et al., 2021). This issue was amplified in LSTM and GRU models for high-resolution forecasting due to highly imbalanced datasets. Even physics-based models face similar challenges; for instance, artificial wind-forcing adjustments can reduce bias but introduce new errors (Lorenz & Gräwe, 2023). The eastern Gulf of Finland exhibited higher forecasting errors, particularly during the winter 2019 testing period—a time prone to ESLs. Beyond seasonal factors, this region is inherently difficult to model due to its complex hydrodynamics (Alenius et al., 1998), driven by: (i) sharp salinity gradients from river inflows; (ii) large winter sea level fluctuations from waves and seiches; (iii) strong local currents from high river discharge and a small Rossby radius; and (iv) seasonal ice cover. These interacting factors collectively challenge accurate sea level and circulation modeling in the region.

The DL models (CNN and GRU) were validated using external SA observations from Sentinel-3A and Sentinel-3B. To ensure consistency, the geoid-referenced, bias-corrected HDM was aligned with satellite data through ellipsoid-to-geoid conversion (Jahanmard et al., 2022; Varbla et al., 2022). Although satellite data offer higher resolution than the nautical-mile-scale forecasts, careful corrections—including atmospheric corrections for tropospheric delays, temporal matching, outlier filtering, and excluding points within 5 km of the coast—ensured comparability and validity. Previous studies confirming strong correlations between SA and corrected HDM data further support this validation approach (Mostafavi et al., 2023, 2024). Overall, both the bias-corrected HDM and the DL models showed good agreement with satellite measurements, with RMSE generally within 5 cm. Larger discrepancies occurred in areas affected by coastal land contamination, where satellite measurements are known to be less reliable.

The second method addressed key limitations of the first approach. The sea level data timeframe was extended to 1971–2022, overcoming the previous 2017–2019 constraint imposed by the availability of corrected NEMO Nordic model data (Jahanmard et al., 2022). Meteorological inputs were also expanded beyond wind speed and pressure to include SWH, rainfall, evaporation, precipitation, wind gust, and the BSI—an indicator of large-scale North Sea influences. A variety of ML and DL models were developed, including RF, MLP, XGB, and hybrid CNN-LSTM and CNN-GRU models. These models were selected based on their demonstrated capabilities in ESL prediction studies, and to represent a diverse range of approaches—from bagging and boosting to simple neural networks and more advanced hybrid models. BO applied for hyperparameter tuning, and the models were trained to forecast SLM. Optimal station-specific feature sets were identified using MI, and the look-back parameter (δ; Equation 5) was selected via BIC.

Neural network-based models (MLP, CNN-LSTM, and CNN-GRU) demonstrated strong skill in forecasting daily SLM across the six examined TG stations: eastern (Narva, Ristna), northern (Oulu), central (Kungsholmsfort), southern (Władysławowo), and southwestern (Greifswald). Model performance was evaluated for both time series forecasting and storm event detection, with the CNN-GRU model achieving the best results, followed by the MLP. CNN-GRU and MLP models achieved an RMSE of 7 to 15 cm and an R^2 of 0.61

to 0.86 for the test period, indicating strong predictive capability. A notable feature of this approach was its ability to accurately capture rapid peaks and intense storms in the range of 100–150 cm—representing a significant improvement over the first method, which tended to underestimate extreme events (above 100 cm). The models were separately evaluated using classification metrics based on different alarming thresholds (60–100 cm) to assess their performance under various extreme conditions. The CNN-GRU model performed best in recall and F1-score, while the MLP ranked highest in accuracy. However, all models showed a tendency to underestimate extreme events above 150 cm, particularly at Narva and Oulu.

The BO-optimized CNN-GRU model was made explainable using SHAP-based XAI—addressing a key limitation of the first method. This revealed that the previous sea level state (prefilling) was the dominant predictor, contributing approximately 40% of predictive power across all stations. Meteorological influences varied spatially. At Narva and Ristna, pressure, wind, and significant wave height were most influential. At southern Baltic stations, the Baltic Sea Index (BSI), pressure, and wind played larger roles—highlighting a spatial gradient from local-scale drivers in the northeast to large-scale drivers in the southwest. To evaluate the model's underestimation of 150 cm events observed at northeastern stations (Narva and Oulu), extreme value analysis using seasonal block maxima was performed. Results showed these extremes have a return period of 5–7 years, confirming the model's sensitivity in capturing such rare events.

The key findings are summarized below according to the two proposed strategies:

- DL models demonstrated high predictive skill for regional sea-level variability, achieving typical forecast errors of about 5 cm for 24-hour predictions.
- Among tested architectures, GRU provided better balance between accuracy and computational efficiency than LSTM, making it the most suitable model for high-resolution operational forecasting. However, CNN acquired RMSE of 4.7 cm used a lag window of previous 7 days compared to 12 hours lag by GRU model.
- A major limitation of spatiotemporal models was systematic underestimation of rare extreme sea levels due to imbalanced training data.
- The second framework significantly improved sea-level maxima forecasting, extending reliable prediction capability from approximately 100 cm to about 150 cm events, which correspond to typical winter return periods of about 5–7 years in the northeastern Baltic Sea.
- Hybrid DL models, particularly BO-CNN-GRU (Bayesian optimized CNN-GRU model), achieved the best overall performance for extreme sea-level prediction across Baltic stations.
- Explainability analysis showed strong physical consistency, with previous sea levels (prefilling), atmospheric pressure, and wind forcing as the main drivers of extreme sea-level variability. Large-scale atmospheric indices, such as the BSI, had stronger influence at western and southern Baltic stations, while eastern stations were more affected by local meteorological conditions, reflecting regional differences in drivers.
- Persistent underestimation of the most extreme events indicates that future improvements require longer datasets, better representation of rare events, and integration of additional physical predictors.
- Overall, the proposed ML–DL frameworks provide accurate, interpretable, and operationally applicable tools that complement traditional hydrodynamic models for sea-level forecasting.

5.2 Future Research Suggestions

Future studies should prioritize

- Addressing the underestimation of the rarest ESL events
- Extending the temporal coverage of training data and incorporating long-term datasets could improve the representation of infrequent extremes.
- The integration of physics-informed techniques may help enforce physically consistent predictions, although preliminary experiments with extreme-penalizing loss functions did not show significant improvements.
- Exploring a broader set of meteorological and oceanographic predictors to better capture both local and large-scale drivers of sea level variability.
- Developing probabilistic forecasting frameworks with uncertainty quantification could support more robust risk-based decision-making, particularly for extreme events.
- Advancing to multi-step-ahead forecasting would enable sea-level forecasts several days in advance, particularly when developing methods tailored to specific grid points.
- Using ensemble DL approaches could improve accuracy and robustness by leveraging complementary strengths of multiple model architectures.
- Applying and validating the methodology in other geographical regions with different hydrodynamic regimes (e.g., using transfer learning) would help assess model generalizability and scalability.

List of Figures

List of Tables

References

Abdalla, S., Kolahchi, A. A., Ablain, M., Adusumilli, S., Bhowmick, S. A., Alou-Font, E., Amarouche, L., Andersen, O. B., Antich, H., Aouf, L., Arbic, B., Armitage, T., Arnault, S., Artana, C., Aulicino, G., Ayoub, N., Badulin, S., Baker, S., Banks, C., … & Zlotnicki, V. (2021). Altimetry for the future: Building on 25 years of progress. Advances in Space Research, 68(2), 319–363.

Ablain, M., Lalau, N., Meyssignac, B., Fraudeau, R., Barnoud, A., Dibarboure, G., ... & Donlon, C. (2025). Benefits of a second tandem flight phase between two successive satellite altimetry missions for assessing instrumental stability. Ocean Science, 21(1), 343–358.

Accarino, G., Chiarelli, M., Fiore, S., Federico, I., Causio, S., Coppini, G., & 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, 124, 1–9.

Ågren, J., Strykowski, G., Bilker-Koivula, M., Omang, O., Märdla, S., Forsberg, R., Ellmann, A., Oja, T., Liepiņš, I., Paršeliūnas, E., Kaminskis, J., Sjöberg, L. E., & Valsson, G. (2016). The NKG2015 gravimetric geoid model for the Nordic-Baltic region. Proceedings of the 1st Joint Commission, 2.

Al Kajbaf, A., & Bensi, M. (2020). Application of surrogate models in estimation of storm surge: A comparative assessment. Applied Soft Computing, 91, 106184.

Alemseged, T. H., & Rientjes, T. H. M. (2007). Uncertainty issues in hydrodynamic flood modeling. In Proceedings of the 5th international symposium on spatial data quality SDQ (Vol. 36, No. 2, p. c43).

Alenius, P., Myrberg, K., & Nekrasov, A. (1998). The physical oceanography of the Gulf of Finland: a review. Boreal Environ. Res, 3(2), 97–125.

Altunkaynak, A. (2007). Forecasting surface water level fluctuations of Lake Van by artificial neural networks. Water Resources Management, 21(2), 399–408.

Ayinde, A. S., Huaming, Y. U., & Kejian, W. U. (2024). Review of machine learning methods for sea level change modeling and prediction. Science of The Total Environment, 954, 176410.

Ayinde, A. S., Yu, H., & Wu, K. (2023). Sea level variability and modeling in the Gulf of Guinea using supervised machine learning. Scientific Reports, 13(1), 21318. https://doi.org/10.1038/s41598-023-21318-x

Balogun, A. L., & Adebisi, N. (2021). Sea level prediction using ARIMA, SVR and LSTM neural network: Assessing the impact of ensemble ocean–atmospheric processes on models' accuracy. Geomatics, Natural Hazards and Risk, 12(1), 653–674.

Barzandeh, A., Ličer, M., Rus, M., Kristan, M., Maljutenko, I., Elken, J., Lagemaa, P. & Uiboupin, R. (2025). Application of the HIDRA2 deep-learning model for sea level forecasting along the Estonian coast of the Baltic Sea. Ocean Science, 21(4), 1315-1327.

Bellinghausen, K., Hünicke, B., & Zorita, E. (2025). Using random forests to forecast daily extreme sea level occurrences at the Baltic Coast. Natural Hazards and Earth System Sciences, 25(3), 1139–1162.

Bengio, Y., Frasconi, P., & Simard, P. (1993, March). The problem of learning long-term dependencies in recurrent networks. In IEEE international conference on neural networks (pp. 1183–1188). IEEE.

Braakmann-Folgmann, A., Roscher, R., Wenzel, S., Uebbing, B., & Kusche, J. (2017). Sea level anomaly prediction using recurrent neural networks. arXiv Preprint arXiv:1710.07099.

Bruneau, N., Polton, J., Williams, J., & Holt, J. (2020). Estimation of global coastal sea level extremes using neural networks. Environmental Research Letters, 15(7), 074030.

Buch, E. (2002). Forecasting the Baltic. In Elsevier Oceanography Series (Vol. 66, pp. 179–188). Elsevier.

Burchard, H., & Bolding, K. K. (2002). GETM, a general estuarine transport model.

Calkoen, F., Luijendijk, A., Rivero, C. R., Kras, E., & Baart, F. (2021). Traditional vs. machine-learning methods for forecasting sandy shoreline evolution using historic satellite-derived shorelines. Remote Sensing, 13(5), 934.

Cazenave, A., Dieng, H. B., Meyssignac, B., Von Schuckmann, K., Decharme, B., & Berthier, E. (2014). The rate of sea-level rise. Nature Climate Change, 4(5), 358–361.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Church, J. A., Clark, P. U., Cazenave, A., Gregory, J. M., Jevrejeva, S., Levermann, A., Merrifield, M. A., Milne, G. A., Nerem, R. S., Nunn, P. D., Payne, A. J., Pfeffer, W. T., Stammer, D., & Unnikrishnan, A. S. (2013). Sea level change. PM Cambridge University Press.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). An introduction to statistical modeling of extreme values (Vol. 208, p. 208). London: Springer.

Dangendorf, S., Calafat, F. M., Arns, A., Wahl, T., Haigh, I. D., & Jensen, J. (2014). Mean sea level variability in the North Sea: Processes and implications. Journal of Geophysical Research: Oceans, 119(10).

Delpeche-Ellmann, N., Giudici, A., Rätsep, M., & Soomere, T. (2021). Observations of surface drift and effects induced by wind and surface waves in the Baltic Sea for the period 2011–2018. Estuarine, Coastal and Shelf Science, 249, 107071.

Delpeche-Ellmann, N., Mingelaitė, T., & Soomere, T. (2017). Examining Lagrangian surface transport during a coastal upwelling in the Gulf of Finland, Baltic Sea. Journal of Marine Systems, 171, 21–30.

Ekman, M. (1989). Impacts of geodynamic phenomena on systems for height and gravity. Bulletin Géodésique, 63(3), 281-296.

El Aouni, A., Gaudel, Q., Regnier, C., Van Gennip, S., Le Galloudec, O., Drevillon, M., ... & Lellouche, J. M. (2025). GLONET: Mercator's end-to-end neural Global Ocean forecasting system. Journal of Geophysical Research: Machine Learning and Computation, 2(3), e2025JH000686.

Fang, Z., & Hong, Z. (2024). Research on deep learning models combining Transformer and BERT: Application in network intrusion detection. In Proceedings of the 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM) (pp. 218–223). IEEE.

Feng, X., Ma, G., Su, S. F., Huang, C., Boswell, M. K., & Xue, P. (2020). A multi-layer perceptron approach for accelerated wave forecasting in Lake Michigan. Ocean Engineering, 211, 107526.

Filippo, A., Torres, A. R., Jr., Kjerfve, B., & Monat, A. (2012). Application of artificial neural network (ANN) to improve forecasting of sea level. Ocean & Coastal Management, 55, 101–110.

Frederikse, T., Landerer, F., Caron, L., Adhikari, S., Parkes, D., Humphrey, V. W., Dangendorf, S., Hogarth, P., Zanna, L., Cheng, L., & Wu, Y. H. (2020). The causes of sea-level rise since 1900. Nature, 584(7821), 393–397.

Ganju, N. K., Brush, M. J., Rashleigh, B., Aretxabaleta, A. L., Del Barrio, P., Grear, J. S., Harris, A. L., Lake S. J., McCardell G., O'Donnell J., Ralston, D. K., Signell, R. P., Testa, J. M., & Vaudrey, J. M. P. (2016). Progress and challenges in coupled hydrodynamic-ecological estuarine modeling, Estuar. Coast., 39, 311–332.

Ghorbani, M. A., Deo, R. C., Karimi, V., Yaseen, Z. M., & Terzi, O. (2018). Implementation of a hybrid MLP–FFA model for water level prediction of Lake Eğirdir, Turkey. Stochastic Environmental Research and Risk Assessment, 32(6), 1683–1697.

Ghorbani, M. A., Khatibi, R., Aytek, A., Makarynskyy, O., & Shiri, J. (2010). Sea water level forecasting using genetic programming and comparing the performance with artificial neural networks. Computers & Geosciences, 36(5), 620–627. https://doi.org/10.1016/j.cageo.2009.09.007

Ghosh, S., Sharma, A., Gupta, J., Subramanian, A., & Shekhar, S. (2024). Towards kriging-informed conditional diffusion for regional sea-level data downscaling. In Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems (pp. 372–383).

Guillou, N., & Chapalain, G. (2021). Machine learning methods applied to sea level predictions in the upper part of a tidal estuary. Oceanologia, 63(4), 531–544.

Han, Q., Jiang, X., Zhao, Y., Wang, X., Li, Z., & Zhang, R. (2024). Generative diffusion model-based downscaling of observed sea surface height over Kuroshio Extension since 2000. arXiv Preprint arXiv:2408.12632.

Harter, L., Pineau-Guillou, L., & Chapron, B. (2024). Underestimation of extremes in sea level surge reconstruction. Scientific Reports, 14(1), 14875.

Hazrin, N. A., Chong, K. L., Huang, Y. F., Ahmed, A. N., Ng, J. L., Koo, C. H., Tan, K. W., Sherif, M., & El-Shafie, A. (2023). Predicting sea levels using ML algorithms in selected locations along coastal Malaysia. Heliyon, 9(9), e19327.

Hieronymus, M., Hieronymus, J., & Arneborg, L. (2017). Sea level modelling in the Baltic and the North Sea: The respective role of different parts of the forcing. Ocean Modelling, 118, 59–72.

Hofstede, J., & Hamann, M. (2022). The 1872 catastrophic storm surge at the Baltic Sea coast of Schleswig-Holstein; lessons learned. Die Küste, Karlsruhe, Bundesanstalt für Wasserbau, https://doi. org/10.18171/1.092101.

Hordoir, R., Axell, L., Höglund, A., Dieterich, C., Fransner, F., Gröger, M., Liu, Y., Pemberton, P., Schimanke, S., Andersson, H., Ljungemyr, P., Nygren, P., Falahat, S., Nord, A., Jönsson, A., Lake, I., Döös, K., Hieronymus, M., Dietze, H., Löptien, U., Kuznetsov, I., Westerlund, A., Tuomi, L., & Haapala, J. (2019). Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas–research and operational applications. Geoscientific Model Development, 12(1), 363–386.

Hordoir, R., Jahanmard, V., Isachsen, P. E., Loeptien, U., Dietze, H., Sandø, A. B., & Lien, V. S. (2026). Barents Sea atlantification driven by a shift in atmospheric synoptic timescale. Nature Climate Change, 1–8.

Horton, B. P., Kopp, R. E., Garner, A. J., Hay, C. C., Khan, N. S., Roy, K., & Shaw, T. A. (2018). Mapping sea-level change in time, space, and probability. Annual Review of Environment and Resources, 43(1), 481–521.

Hünicke, B., & Zorita, E. (2008). Trends in the amplitude of Baltic Sea level annual cycle. Tellus A: Dynamic Meteorology and Oceanography, 60(1), 154–164.

IHO. (1987). IHO Standards for Hydrographic Surveys and Classification Criteria for Deep Sea Soundings and Procedures for Elimination of Doubtful Data (No. 44). International Hydrographic Bureau.

Imani, M., Kao, H. C., Lan, W. H., & Kuo, C. Y. (2018). Daily sea level prediction at Chiayi coast, Taiwan using extreme learning machine and relevance vector machine. Global and Planetary Change, 161, 211–221.

IPCC, 2021: Climate change 2021-the physical science basis. Interaction, 49(4), 44–45.

Ishida, K., Tsujimoto, G., Ercan, A., Tu, T., Kiyama, M., & Amagasaki, M. (2020). Hourly-scale coastal sea level modeling in a changing climate using long short-term memory neural network. Science of the Total Environment, 720, 137613.

Jahanmard, V. (2024). Developments Towards Deriving Realistic Dynamic Topography by Synergizing High-Resolution Geoid with Sea Level Data [TalTech Press]. https://doi.org/10.23658/taltech.3/2024.

Jahanmard, V., Delpeche-Ellmann, N., & Ellmann, A. (2021). Realistic dynamic topography through coupling geoid and hydrodynamic models of the Baltic Sea. Continental Shelf Research, 222, 104421.

Jahanmard, V., Delpeche-Ellmann, N., & Ellmann, A. (2022). Towards realistic dynamic topography from coast to offshore by incorporating hydrodynamic and geoid models. Ocean Modelling, 180, 102124.

Jahanmard, V., Hordoir, R., Delpeche-Ellmann, N., & Ellmann, A. (2023). Quantification of hydrodynamic model sea level bias utilizing deep learning and synergistic integration of data sources. Ocean Modelling, 186, 102286.

Jevrejeva, S., Calafat, F. M., De Dominicis, M., Hirschi, J. J. M., Mecking, J. V., Polton, J. A., Sinha, B.; Wise, A., & Holt, J. (2024). Challenges, advances and opportunities in regional sea level projections: The role of ocean-shelf dynamics. Earth's Future, 12(8), e2024EF004886.

Jevrejeva, S., Moore, J. C., Grinsted, A., & Woodworth, P. L. (2008). Recent global sea level acceleration started over 200 years ago?. Geophysical Research Letters, 35(8).

Jevrejeva, S., Williams, J., Vousdoukas, M. I., & Jackson, L. P. (2023). Future sea level rise dominates changes in worst case extreme sea levels along the global coastline by 2100. Environmental Research Letters, 18(2), 024037.

Jia, Y., Xiao, K., Lin, M., & Zhang, X. (2022). Analysis of global sea level change based on multi-source data. Remote Sensing, 14(19), 4854.

Johansson, M., Boman, H., Kahma, K. K., & Launiainen, J. (2001). Trends in sea level variability in the Baltic Sea. Boreal Environment Research, 6(3), 159–180.

Karimi, S., Kisi, O., Shiri, J., & Makarynskyy, O. (2013). Neuro-fuzzy and neural network techniques for forecasting sea level in Darwin Harbor, Australia. Computers & Geosciences, 52, 50–59.

Khaledian, M. R., Isazadeh, M., Biazar, S. M., & Pham, Q. B. (2020). Simulating Caspian Sea surface water level by artificial neural network and support vector machine models. Acta Geophysica, 68(2), 553–563.

Kowalewski, M., & Kowalewska-Kalkowska, H. (2017). Sensitivity of the Baltic Sea level prediction to spatial model resolution. Journal of Marine Systems, 173, 101–113.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436–444.

Krasnopolsky, V. M., & Fox-Rabinovitz, M. S. (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Networks, 19(2), 122–134.

Kulikov, E. A., & Medvedev, I. P. (2013). Variability of the Baltic Sea level and floods in the Gulf of Finland. Oceanology, 53(2), 145–151.

Kulikov, E. A., Medvedev, I. P., & Koltermann, K. P. (2015). Baltic sea level low-frequency variability. Tellus A: Dynamic Meteorology and Oceanography, 67(1), 25642.

Lagemaa, P., Elken, J., & Kõuts, T. (2011). Operational sea level forecasting in Estonia. Estonian Journal of Engineering, 17(4), 301.

Lehmann, A., & Myrberg, K. (2008). Upwelling in the Baltic Sea—A review. Journal of Marine Systems, 74, S3–S12.

Lehmann, A., Krauß, W., & Hinrichsen, H. H. (2002). Effects of remote and local atmospheric forcing on circulation and upwelling in the Baltic Sea. Tellus A: Dynamic meteorology and oceanography, 54(3), 299–316.

Lehmann, A., Myrberg, K., Post, P., Chubarenko, I., Dailidiene, I., Hinrichsen, H. H., Hüssy, K., Liblik, T., Meier, H. E. M., Lips, U. & Bukanova, T. (2022). Salinity dynamics of the Baltic Sea. Earth System Dynamics, 13(1), 373–392.

Leppäranta, M., & Myrberg, K. (2009). Physical oceanography of the Baltic Sea. Berlin, Heidelberg: Springer Berlin Heidelberg.

Liebsch, G., Schwabe, J., Varbla, S., Ågren, J., Teitsson, H., Ellmann, A., Forsberg, R., Strykowski, G., Bilker-Koivula, M., Liepiņš, I., Paršeliūnas, E., Keller, K., Vestøl, O., Omang, O., Kaminskis, J., Wilde-Piórko, M., Pyrchla, K., Olsson, P.-A., Förste, C., Ince, E. S., Somla, J., Westfeld, P., & Hammarklint, T. (2023). Release note for the BSCD2000 height transformation grid. The International Hydrographic Review, 29(2).

Liu, J., Jin, B., Wang, L., & Xu, L. (2020). Sea surface height prediction with deep learning based on attention mechanism. IEEE Geoscience and Remote Sensing Letters, 19, 1–5.

Lorenz, M., & Gräwe, U. (2023). Uncertainties and discrepancies in the representation of recent storm surges in a non-tidal semi-enclosed basin: a hindcast ensemble for the Baltic Sea. Ocean Science, 19(6), 1753–1771.

Lorenz, M., Viigand, K., Gräwe, U., 2025. Untangling the waves: decomposing extreme sea levels in a non-tidal basin, the Baltic Sea. Natural Hazards and Earth System Sciences, 25(4), 1439–1458. https://doi.org/10.5194/nhess-25-1439-2025

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Makarynskyy, O., Makarynska, D., Kuhn, M., & Featherstone, W. E. (2004). Predicting sea level variations with artificial neural networks at Hillarys Boat Harbour, Western Australia. Estuarine, Coastal and Shelf Science, 61(2), 351–360.

Mäkinen, J. (2021). The permanent tide and the international height reference frame IHRF. Journal of Geodesy, 95(9), 106.

Mardani, N., Suara, K., Fairweather, H., Brown, R., McCallum, A., & Sidle, R. C. (2020). Improving the accuracy of hydrodynamic model predictions using Lagrangian calibration. Water, 12(2), 575.

Medvedev, I. P., & Kulikov, E. A. (2021). Extreme Storm Surges in the Gulf of Finland: Frequency-Spectral Properties and the Influence of Low-Frequency Sea Level Oscillations. Oceanology, 61(4), 459–468.

Meier, H. M., Broman, B., & Kjellström, E. (2004). Simulated sea level in past and future climates of the Baltic Sea. Climate research, 27, 59–75.

Meier, M., Andréasson, J., Broman, B., Graham, P., Kjellström, E., & Persson, G. (2006). Climate change scenario simulations of wind, sea level, and river discharge in the Baltic Sea and Lake Mälaren region–a dynamical downscaling approach from global to local scales. SMHI.

Milne, G. A., Gehrels, W. R., Hughes, C. W., & Tamisiea, M. E. (2009). Identifying the causes of sea-level change. Nature Geoscience, 2(7), 471–478.

Mostafavi, M., Delpeche-Ellmann, N., & Ellmann, A. (2021). Accurate sea surface heights from Sentinel-3A and Jason-3 retrackers by incorporating high-resolution marine geoid and hydrodynamic models. Journal of Geodetic Science, 11(1), 58–74.

Mostafavi, M., Delpeche-Ellmann, N., Ellmann, A., & Jahanmard, V. (2023). Determination of accurate dynamic topography for the Baltic Sea using satellite altimetry and a marine geoid model. Remote Sensing, 15(8), 2189.

Mostafavi, M., Ellmann, A., & Delpeche-Ellmann, N. (2024). Long-Term and Decadal Sea-Level Trends of the Baltic Sea Using Along-Track Satellite Altimetry. Remote Sensing, 16(5), 760.

Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C., & Ward, P. J. (2016). A global reanalysis of storm surges and extreme sea levels. Nature communications, 7(1), 11969.

Nguyen, D. H., Le, X. H., Heo, J. Y., & Bae, D. H. (2021). Development of an extreme gradient boosting model integrated with evolutionary algorithms for hourly water level prediction. IEEE Access, 9, 125853–125867.

Nieves, V., Radin, C., & Camps-Valls, G. (2021). Predicting regional coastal sea level Orseau, S., Huybrechts, N., Tassi, P., Kaidi, S., & Klein, F. (2021). NavTEL: Open-source decision support tool for ship routing and underkeel clearance management in Estuarine Channels. Journal of Waterway, Port, Coastal, and Ocean Engineering, 147(2), 04020053. changes with machine learning. Scientific reports, 11(1), 7650.

Orseau, S., Huybrechts, N., Tassi, P., Kaidi, S., & Klein, F. (2021). NavTEL: Open-source decision support tool for ship routing and underkeel clearance management in Estuarine Channels. Journal of Waterway, Port, Coastal, and Ocean Engineering, 147(2), 04020053.

Pachev, B., Arora, P., del-Castillo-Negrete, C., Valseth, E., & Dawson, C. (2023). A framework for flexible peak storm surge prediction. Coastal Engineering, 186, 104406.

Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based PM2. 5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. Science of the Total Environment, 699, 133561.

Pashova, L., & Popova, S. (2011). Daily sea level forecast at tide gauge Burgas, Bulgaria using artificial neural networks. Journal of Sea Research, 66(2), 154–161.

Passaro, M., & Juhl, M. C. (2023). On the potential of mapping sea level anomalies from satellite altimetry with Random Forest Regression. Ocean Dynamics, 73(2), 107–116.

Passaro, M., Cipollini, P., & Benveniste, J. (2015). Annual sea level variability of the coastal ocean: The Baltic Sea-North Sea transition zone. Journal of Geophysical Research: Oceans, 120(4), 3061–3078.

Passaro, M., Müller, F., Dettmering, D., Abulaitjiang, A., Rautiainen, L., Scarrott, R.G., Chalençon, E., & Sweeney, M. (2021). BALTIC+ theme 3 Baltic+ SEAL (sea level) product handbook.

Ponte, R. M., Carson, M., Cirano, M., Domingues, C. M., Jevrejeva, S., Marcos, M., Mitchum, G., Van De Wal, R., Woodworth, P. L., Ablain, M., Ardhuin, F., Ballu, V., Becker, M., Benveniste, J., Birol, F., Bradshaw, E., Cazenave, A., De Mey-Frémaux, P., Durand, F., Ezer, T., Fu, L.-L., Fukumori, I., Gordon, K., Gravelle, M., Griffies, S. M., Han, W., Hibbert, A., Hughes, C. W., Idier, D., Kourafalou, V. H., Little, C. M., Matthews, A., Melet, A., Merrifield, M., Meyssignac, B., Minobe, S., Penduff, T., Picot, N., Piecuch, C., Ray, R. D., Rickards, L., Santamaría-Gómez, A., Stammer, D., Staneva, J., Testut, L., Thompson, K., Thompson, P., Vignudelli, S., Williams, J., Williams, S. D. P., Wöppelmann, G., Zanna, L., & Zhang, X. (2019). Towards comprehensive observing and modeling systems for monitoring and predicting regional to coastal sea level. Frontiers in Marine Science, 6, 437.

Qin, Y., Su, C., Chu, D., Zhang, J., & Song, J. (2023). A review of application of machine learning in storm surge problems. Journal of Marine Science and Engineering, 11(9), 1729. https://doi.org/10.3390/jmse11091729

Qin, Z., Duan, Y., Wang, Y., Shen, Z., & Xu, K. (1994). Numerical simulation and prediction of storm surges and water levels in Shanghai harbour and its vicinity. Natural hazards, 9(1), 167–188.

Raj, N., & Brown, J. (2023). Prediction of mean sea level with GNSS-VLM correction using a hybrid deep learning model in Australia. Remote Sensing, 15(11), 2881.

Raj, N., Gharineiat, Z., Ahmed, A. A. M., & Stepanyants, Y. (2022). Assessment and prediction of sea level trend in the South Pacific region. Remote Sensing, 14(4), 986.

Rajabi-Kiasari, S., Delpeche-Ellmann, N., & Ellmann, A. (2023). Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea. Computers & Geosciences, 178, 105406, doi: doi.org/10.1016/j.cageo.2023.105406.

Rajabi-Kiasari, S., Delpeche-Ellmann, N., Ellmann, A., & Soomere, T. (2026). Forecasting sea level maxima using Machine learning with explainability and extreme value analysis. International Journal of Applied Earth Observation and Geoinformation, 146, 105064.

Rajabi-Kiasari, S., Ellmann, A., & Delpeche-Ellmann, N. (2025). Sea level forecasting using deep recurrent neural networks with high-resolution hydrodynamic model. Applied Ocean Research, 157, 104496, doi: doi.org/10.1016/j.apor.2025.104496.

Ramos-Valle, A. N., Curchitser, E. N., Bruyère, C. L., & McOwen, S. (2021). Implementation of an artificial neural network for storm surge forecasting. Journal of Geophysical Research: Atmospheres, 126(13), e2020JD033266.

Rus, M., Fettich, A., Kristan, M., & Ličer, M. (2023). HIDRA2: deep-learning ensemble sea level and storm tide forecasting in the presence of seiches–the case of the northern Adriatic. Geoscientific Model Development, 16(1), 271–288.

Samuelsson, M., & Stigebrandt, A. (1996). Main characteristics of the long-term sea level variability in the Baltic sea. Tellus a, 48(5), 672–683.

Särkkä, J., Räihä, J., Rantanen, M., & Kämäräinen, M. (2023). Simulating sea level extremes from synthetic low-pressure systems. Natural Hazards and Earth System Sciences Discussions, 2023, 1–13.

Sertel, E., Cigizoglu, H. K., & Sanli, D. U. (2008). Estimating daily mean sea level heights using artificial neural networks. Journal of Coastal Research, 24(3), 727–734.

Shahabi, A., & Tahvildari, N. (2024). A deep-learning model for rapid spatiotemporal prediction of coastal water levels. Coastal Engineering, 190, 104504.

Shchepetkin, A. F., & McWilliams, J. C. (2005). The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. Ocean modelling, 9(4), 347–404.

She, J., Berg, P., & Berg, J. (2007). Bathymetry impacts on water exchange modelling through the Danish Straits. Journal of Marine Systems, 65(1–4), 450–459.

Sithara, S., Pramada, S. K., & Thampi, S. G. (2020). Sea level prediction using climatic variables: A comparative study of SVM and hybrid wavelet SVM approaches. Acta Geophysica, 68(6), 1779–1790.

Slobbe, D. C., Klees, R., & Gunter, B. C. (2014). Realization of a consistent set of vertical reference surfaces in coastal areas. Journal of Geodesy, 88(6), 601–615.

Slobbe, D. C., Verlaan, M., Klees, R., & Gerritsen, H. (2013). Obtaining instantaneous water levels relative to a geoid with a 2D storm surge model. Continental Shelf Research, 52, 172–189.

Soomere, T. (2003). Anisotropy of wind and wave regimes in the Baltic Proper. Journal of Sea Research, 49(4), 305–316.

Soomere, T., Myrberg, K., Lepparanta, M., & Nekrasov, A. (2008). The progress in knowledge of physical oceanography of the Gulf of Finland: e review for 1997-2007. Oceanologia, 50(3), 287–362.

Stammer, D., Cazenave, A., Ponte, R. M., & Tamisiea, M. E. (2013). Causes for contemporary regional sea level changes. Annual review of marine science, 5(1), 21–46.

Sun, K., & Pan, J. (2023). Model of storm surge maximum water level increase in a coastal area using ensemble machine learning and explicable algorithm. Earth and Space Science, 10(12), e2023EA003243.

Suursaar, Ü., & Sooäär, J. (2007). Decadal variations in mean and extreme sea level values along the Estonian coast of the Baltic Sea. Tellus A: Dynamic Meteorology and Oceanography, 59(2), 249–260.

Sztobryn, M. (2003). Forecast of storm surge by means of artificial neural network. Journal of Sea Research, 49(4), 317–322.

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., & Ward, P. J. (2021). Exploring deep learning capabilities for surge predictions in coastal areas. Scientific Reports, 11(1), 17224.

Tilburg, C. E., & Garvine, R. W. (2004). A simple model for coastal sea level prediction. Weather and forecasting, 19(3), 511–519.

Tur, R., Tas, E., Haghighi, A. T., & Mehr, A. D. (2021). Sea level prediction using machine learning. Water, 13(24), 3566.

Varbla, S. (2023). Iterative Refinement and Accuracy Validation of Marine Geoid Models. TalTech Press]. https://doi. org/10.23658/taltech, 35, 2023.

Varbla, S., Ågren, J., Ellmann, A., & Poutanen, M. (2022). Treatment of tide gauge time series and marine GNSS measurements for vertical land motion with relevance to the implementation of the Baltic Sea Chart Datum 2000. Remote Sensing, 14(4), 920.

Varbla, S., Ellmann, A., & Delpeche-Ellmann, N. (2020). Validation of marine geoid models by utilizing hydrodynamic model and shipborne GNSS profiles. Marine Geodesy, 43(2), 134–162.

Varbla, S., Ellmann, A., & Delpeche-Ellmann, N. (2021). Applications of airborne laser scanning for determining marine geoid and surface waves properties. European Journal of Remote Sensing, 54(1), 558–568.

Wang, B., Wang, B., Wu, W., Xi, C., & Wang, J. (2020). Sea-water-level prediction via combined wavelet decomposition, neuro-fuzzy and neural networks using SLA and wind information. Acta Oceanologica Sinica, 39(5), 157–167.

Wang, G., Wang, X., Wu, X., Liu, K., Qi, Y., Sun, C., & Fu, H. (2022). A hybrid multivariate deep learning network for multistep ahead sea level anomaly forecasting. Journal of Atmospheric and Oceanic Technology, 39(3), 285–301.

Wedam, G. B., McMurdie, L. A., & Mass, C. F. (2009). Comparison of model forecast skill of sea level pressure along the east and west coasts of the United States. Weather and forecasting, 24(3), 843–854.

Wei, L., Guan, L., & Qu, L. (2019). Prediction of sea surface temperature in the South China Sea by artificial neural networks. IEEE geoscience and remote sensing letters, 17(4), 558–562.

Weisse, R., & Hünicke, B. (2019). Baltic sea level: past, present, and future. In Oxford Research Encyclopedia of Climate Science.

Weisse, R., Dailidiene, I., Hünicke, B., Kahma, K., Madsen, K., Omstedt, A., Parnell, K., Schöne, T., Soomere, T., Zhang, W., & Zorita, E. (2021). Sea level dynamics and coastal erosion in the Baltic Sea region. Earth System Dynamics Discussions, 1–40.

Wolski, T., & Wiśniewski, B. (2021). Characteristics and long-term variability of occurrences of Storm surges in the Baltic Sea. Atmosphere, 12(12), 1679.

Wolski, T., Wiśniewski, B., Giza, A., Kowalewska-Kalkowska, H., Boman, H., Grabbi-Kaiv, S., Hammarklint T., Holfort J., & Lydeikaitė, Ž. (2014). Extreme sea levels at selected stations on the Baltic Sea coast. Oceanologia, 56(2), 259-290.

Wulff, F. V., Rahm, L., & Larsson, P. (Eds.). (2001). A systems analysis of the Baltic Sea (Vol. 148). Springer Science & Business Media.

Wunsch, C., & Stammer, D. (1997). Atmospheric loading and the oceanic "inverted barometer" effect. Reviews of Geophysics, 35(1), 79–107.

Zhang, Z., Yin, J., & Wang, L. (2024). Multi-head attention ResUnet with sequential sliding windows for sea surface height anomaly field forecast: A regional study in North Atlantic Ocean. Applied Soft Computing, 157, 111551.

Žust, L., Fettich, A., Kristan, M., & Ličer, M. (2021). HIDRA 1.0: deep-learning-based ensemble sea level forecasting in the northern Adriatic. Geoscientific Model Development, 14(4), 2057–2074.

Acknowledgements

I would like to express my sincere gratitude to all those who contributed to the successful completion of this work.

First and foremost, I would like to thank my supervisor Artu Ellmann and co-supervisor Nicole Depeche-Ellmann for their guidance, encouragement, and valuable support throughout this research journey. Their insights, patience, and constructive comments greatly contributed to the development and improvement of this dissertation, and their encouragement was instrumental in overcoming the challenges encountered during this work.

I am especially grateful to Eduardo Zorita for his guidance during my research visit. His expertise, thoughtful discussions, and generous support greatly enriched my research experience.

I would also like to thank Tarmo Soomere, my co-author, for his collaboration, insightful contributions, and productive cooperation throughout the research process. I also thank colleagues and co-authors in Geodesy research group (Vahidreza Jahanmard, Majid Mostafavi, Sander Varbla, and Aleksei Kupavõh) for their cooperation and assistance. My sincere appreciation goes also to my friends for their support.

I am profoundly thankful to my family, especially my father and mother, for their love, understanding, and continuous support, which have been a constant source of motivation.

Finally, on a personal note, I would like to express my deepest gratitude to my wife, Faeze, whose unwavering patience, constant support, and compassionate encouragement played an essential role throughout all stages of this research. Her understanding, sacrifices, and continuous moral support during the challenging moments of this academic journey were a significant source of motivation and resilience for me. Without her calm, encouragement, and stability, the completion of this dissertation would have been considerably more difficult.

This research was supported by the Estonian Research Council through the grants "Development of an iterative approach for near-coast marine geoid modelling using re-tracked satellite altimetry, in-situ, and modelled data" (PRG330) and "Development of a continuous dynamic vertical reference for maritime and offshore engineering using machine learning strategies (DYNAREF)" (PRG1785).

Data Statement

The authors gratefully acknowledge the Swedish Meteorological and Hydrological Institute (SMHI) for providing the original Nemo Nordic dataset, including sea levels, wind speed, sea surface temperature (SST), and sea surface salinity (SSS). They also thank the Baltic+SEAL project team (Passaro et al., 2021) for the access to a preliminary version of this dataset, which was used for CNN model validation. Moreover, high-resolution sea-level forecasts were validated using Sentinel-3A and Sentinel-3B altimetry from EUMETSAT (https://www.eumetsat.int/).

Geoid data were obtained from NKG2015 (Nordic Geodetic Commission, NKG) and BSCD2000 (The Baltic Sea Hydrographic Commission, BSHC). Additional datasets included SST from NASA Earthdata (https://cmr.earthdata.nasa.gov/search) and wind speed from the CCMP v2.0 product (https://data.remss.com/ccmp/v02.0).

Tide-gauge observations were collected from national agencies in Estonia (envir.ee), Finland (ilmatieteenlaitos.fi), Sweden (smhi.se), Germany (wsv.bund.de), and Poland (imgw.pl). Other meteorological forcing—comprising precipitation, evaporation, runoff, atmospheric pressure, and Baltic Sea Index (BSI)—was derived from ECMWF Reanalysis V5 (ERA5) reanalysis (marine.copernicus.eu). Significant wave height (SWH) data were obtained from the WAM and SWAN models.

Abstract

Development of Machine Learning Approaches for Geoid-referred Sea Level Forecasting

Reliable sea level forecasting is increasingly critical for coastal protection, maritime safety, and climate mitigation and adaptation solutions amid rising seas and intensifying hazards. While recent advances in artificial intelligence (AI) and machine learning (ML) enable accurate and computationally efficient short-term forecasts that are competitive with physics-based ocean models, major challenges remain. These include integration of heterogeneous data sources, achieving high spatiotemporal resolution in multivariate multi-step forecasting, accurately representing extreme events, ensuring vertical consistency across datasets, and enhancing model interpretability.

This thesis aims to develop a vertically consistent, explainable, and hybrid machine learning framework for sea-level forecasting in the Baltic Sea that integrates satellite, in-situ data-sets, and stochastic methodologies. A central innovation of this work is the implementation of a fully geoid-referenced sea-level forecasting framework, in which all datasets are harmonized to the geoid surface (NKG2015 and BSCD2000), ensuring vertical consistency across hydrodynamic models, satellite altimetry and tide-gauge observations. To the best of our knowledge, this is the first application of a geoid-referenced ML framework for sea-level forecasting in the Baltic Sea.

The first methodological component addresses spatiotemporal forecasting using convolutional neural networks (CNN), long short-term memory (LSTM), and gated recurrent unit (GRU) architectures. Two complementary forecasting strategies were developed. A basin-scale CNN model (0.25° × 0.25°) generated daily sea-level forecasts across the Baltic Sea, achieving R^2 = 0.91 and RMSE = 4.7 cm during test period. A high-resolution sub-basin configuration, applied to dynamically complex regions such as the Gulf of Finland, produced hourly forecasts at one-nautical-mile resolution (unprecedented in ML-based sea level forecasting) for lead times of 3–24 hours. In this configuration, the GRU model outperformed the LSTM, achieving an average RMSE of 4.96 cm; R^2 = 0.93. Further validation against independent satellite altimetry for both the CNN and GRU models showed agreement generally within 5 cm, except in nearshore areas affected by land contamination.

Despite strong overall performance, deep learning models showed reduced skill for extreme sea levels exceeding 100 cm. To address this limitation, the second methodological component shifted from continuous forecasting to daily maximum sea-level forecasting using long-term tide-gauge records (1971–2022) from six representative Baltic Sea TG stations. Multiple ML models were evaluated, including Random Forest, Extreme Gradient Boosting, multilayer perceptron, and hybrid CNN–GRU and CNN–LSTM architectures, with model hyperparameters tuned through Bayesian optimization. Neural-network-based models, particularly CNN–GRU, achieved the best performance (RMSE 7–15 cm; R^2 0.61–0.86) and successfully captured storm surges of 120–140 cm. However, rare extremes exceeding 150 cm were generally underestimated. Extreme value analysis indicated that such events correspond to winter return periods of approximately 5–7 years at Narva and Oulu stations.

Explainable AI analysis revealed that pre-existing sea-level conditions are the dominant predictor across all stations (~40% of stations), while meteorological drivers exhibit strong spatial variability. The hybrid CNN–GRU architecture demonstrated superior

capability in capturing spatiotemporal dependencies and improving extreme-event prediction compared to standalone models.

Overall, this thesis demonstrates that integrating machine learning with stochastic extreme value analysis provides a robust and operationally viable framework for geoid-referred sea-level forecasting and coastal risk assessment. While rare extremes remain challenging, the proposed approach advances data-consistent, interpretable, and hybrid forecasting methodologies for regional marine applications.

Lühikokkuvõte

Masinõppe meetodite väljatöötamine geoidi suhtes määratletava merepinna taseme prognoosimiseks

Merepinna taseme prognoosimine on üha olulisem rannikualade kaitse, meresõiduohutuse ning kliimamuutuste leevendamise ja nendega kohanemise lahenduste jaoks merepinna tõusu ja seonduvate ohtude kontekstis. Kuigi tehisintellekti (AI) ja masinõppe (ML) hiljutised edusammud võimaldavad täpseid ja arvutuslikult tõhusaid lühiajalisi prognoose, mis on konkurentsivõimelised olemasolevate ookeanitsirkulatsiooni mudelitega, on suured väljakutsed endiselt lahendamata. Nende hulka kuuluvad heterogeensete andmeallikate integreerimine, kõrge aeg-ruumilise lahutusvõime saavutamine veetaseme mitmemõõtmelises prognoosimises, ekstreemsete sündmuste täpne esitamine, järjepidevuse tagamine andmekogumite vahel ja mudeli tõlgendatavuse parendamine.

Käesoleva väitekirja eesmärk on töötada välja järjepidev, selgitatav ja hübriidne masinõppe raamistik Läänemere merepinna taseme prognoosimiseks, mis integreerib satelliit- ja kohapealseid andmestikke ning stohhastilisi metoodikaid. Selle töö keskne uuendus on geoidi suhtes määratletava merepinna taseme prognoosimine, milles kõik seonduvad hüdrodünaamiliste mudelite, satelliit-altimeetria ja veemõõdujaama vaatluste andmekogumid on geoidi pinnaga samuti ühtlustatud (NKG2015 ja BSCD2000). Meie teada on see masinõppe esimene rakendamine geoidi suhtes määratletava merepinna taseme prognoosimiseks Läänemerel.

Esimene metodoloogiline komponent käsitleb aegruumilist prognoosimist, kasutades konvolutsioonilisi närvivõrke (CNN), pika lühiajalise mälu (LSTM) ja väravatega rekurrentsete üksuste (GRU) arhitektuure. Töötati välja kaks teineteist täiendavat prognoosimisstrateegiat. Vesikonna mastaabis CNN-mudel (0,25° × 0,25°) genereeris igapäevaseid merepinna prognoose kogu Läänemeres, saavutades testperioodil $R^2 = 0,91$ ja RMSE = 4,7 cm. Kõrge eraldusvõimega alamvesikonna konfiguratsioon, mida rakendati dünaamiliselt keerulistele piirkondadele, näiteks Soome lahele, andis tunniprognoose ühe meremiili resolutsiooniga (masinaõppel põhinevas merepinna prognoosimises enneolematu) 3–24-tunnise etteteatamisajaga. Selles konfiguratsioonis edestas GRU mudel LSTM-i, saavutades keskmise RMSE 4,96 cm; $R^2 = 0,93$. Edasine sõltumatu valideerimine satelliit-altimeetria abil nii CNN kui ka GRU mudelite puhul näitas üldiselt 5 cm piires kokkulangevust, välja arvatud maismaa häiringutest mõjutatud rannikulähedastel aladel.

Vaatamata tugevale üldisele tulemuslikkusele näitasid süvaõppe mudelid kesisemaid oskusi ekstreemsete merevee tasemete puhul, mis ületasid 100 cm. Selle puudujäägi lahendamiseks nihkus teise metodoloogiline komponendi fookus pidevprognoosist igapäevase maksimaalse merevee taseme prognoosimisele, kasutades pikaajalisi (1971–2022) veemõõdujaamade andmeid kuuest representatiivsest Läänemere veemõõdujaamast. Hinnati mitmeid masinõppe mudeleid, sealhulgas Random Forest, Extreme Gradient Boosting ning hübriidsed CNN-GRU ja CNN-LSTM arhitektuurid, kusjuures modelleerimise hüperparameetrid olid häälestatud Bayesi optimeerimise abil. Neuraalvõrgupõhised mudelid, eriti CNN-GRU, saavutasid parima tulemuse (RMSE 7–15 cm; R^2 0,61–0,86) ja tuvastasid edukalt 120–140 cm kõrguseid tormilaineid. 150 cm ületavaid ekstreemsündmused jäid aga üldiselt alahinnatuks. Ekstreemumite väärtuste analüüs näitas, et sellised sündmused vastavad Narva ja Oulu jaamades umbes

5–7-aastastele talvistele kordumisperioodidele. Selgitava tehisintellekti analüüs näitas, et olemasolevad merepinna tingimused on domineerivaks faktoriks kõigis jaamades (~40% jaamadest), samas kui meteoroloogilistel teguritel on tugev ruumiline varieeruvus. Hübriidne CNN-GRU arhitektuur näitas paremat võimekust ruumiliste ja ajaliste sõltuvuste tuvastamisel ning ekstreemsündmuste ennustamise parendamisel võrreldes eraldiseisvate mudelitega.

Üldiselt näitab see väitekiri, et masinõppe meetodite integreerimine stohhastilise äärmusväärtuste analüüsiga pakub tugeva ja operatiivse raamistiku geoidi suhtes määratletava merepinna taseme prognoosimiseks ja rannikualade riskide hindamiseks. Kuigi harvajuhtuvate suurte ekstreemumite tuvastamine on endiselt keeruline, edendab pakutud lähenemisviis andmepõhiseid, tõlgendatavaid ja hübriidseid prognoosimismeetodeid piirkondlike mereliste tegevuste tarbeks.

Appendix 1

Publication I

Rajabi-Kiasari, S., Delpeche-Ellmann, N., & Ellmann, A. (2023). Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea. Computers & Geosciences, 178, 105406, doi: doi.org/10.1016/j.cageo.2023.105406.

Computers & Geosciences 178 (2023) 105406

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea

Saeed Rajabi-Kiasari [a,*], Nicole Delpeche-Ellmann [b], Artu Ellmann [a]

[a] Department of Civil Engineering and Architecture, Tallinn University of Technology, Ehitajate road 5, 19086, Tallinn, Estonia
[b] Department of Cybernetics, School of Science, Tallinn University of Technology, Ehitajate road 5, 19086, Tallinn, Estonia

ARTICLE INFO

Keywords:
Dynamic topography
Deep learning
Baltic sea
Sea-level prediction
Hydro-geodesy
Geoid

ABSTRACT

Accurate sea-level forecasting is crucial for navigation, engineering and coastal conservation. One of the major obstacles in obtaining accurate sea-level data, both at coastal and offshore areas, has been in defining a realistic vertical datum. The Baltic Sea countries, however, have collaborated in calculating a high-resolution regional geoid model (NKG2015). This now paves the way for determining accurate sea-levels over the entire sea. Accordingly, this study explores the application of a deep learning two-dimensional Convolutional Neural Network (Conv2D) technique along with using essential inputs (e.g. accurate dynamic topography (DT), wind speed and direction, surface pressure and temperature). The method was tested for a three-year 2017–2019 period in the Baltic Sea. The evaluation was based on two statistical criteria: spatial root mean squared error (*RMSE*) and R-squared (R^2). Results revealed that the proposed Conv2D model allows predicting the DT of the Baltic Sea, with an R^2 of 0.91. The spatial *RMSE* plot also confirms accurate DT predictions for most of the Baltic Sea points, with the discrepancies distributed within ±4 cm. Spatially, some larger *RMSE* values (~10 cm) were obtained at particular locations in south-eastern Baltic Sea, which may be due to the input sources utilized. Examination of sea level maxima also showed that the Conv2D model reproduced the maxima events in most scenarios (9 of 10 and 8 of 11 in Gulf of Finland and Riga, respectively) with residuals that varied up to 7 cm–18 cm. For the higher residuals, the Conv2D tended to underestimate the sea level. This suggests the importance of considering the necessary inputs (e.g waves) for forecasting storm surges and that the Conv2D model can still be improved. External validation was also performed using along-track sea level satellite altimetry data, with differences between forecasted model and satellite being within 5 cm. This confirms the validity of the forecasted model and the occurrences of model biases to be minimum. The method utilized shows the potential to contribute toward operational sea level forecasting in the Baltic Sea.

1. Introduction

Accurate sea level forecasting has become more imperative than before, particularly for coastal protection, marine engineering, determining safe shipping routes, fisheries, and geo-hazards management (Liu et al., 2022; Song et al., 2021). It also assists in a better understanding of regional variations and in the development of models for global sea level change (Braakmann-Folgmann et al., 2017; Shao et al., 2022; Wang et al., 2022). Due to advancements in computing technology (language processing, computational power, etc.), machine learning (ML)/Deep learning (DL) algorithms have been widely acknowledged as robust tools in finding patterns and forecasting in various fields (Braakmann-Folgmann et al., 2017; Zhou et al., 2023). As a result, this study explores using ML/DL for accurate sea-level forecasting. It is however, first important to have an understanding of strengths and limitations of various sea-level data sources available. Thereafter, it is essential to have background knowledge and weaknesses of the ML/DL techniques.

For instance, in terms of the limitations of the various sources available, sea level can be determined from several sources, such as tide gauges, satellite altimeters, hydrodynamic models (HDMs) and Global Navigation Satellite System. These different sources, whilst being limited spatially (e.g. tide gauge) and also temporally (e.g. satellite altimetry (SA)), most importantly suffer from inconsistent vertical reference datum (Jahanmard et al., 2021). This results in different sea-level values, thus making inter-comparisons more challenging. In

* Corresponding author.
E-mail address: saeed.rajabi@taltech.ee (S. Rajabi-Kiasari).

https://doi.org/10.1016/j.cageo.2023.105406
Received 3 February 2023; Received in revised form 8 May 2023; Accepted 20 June 2023
Available online 21 June 2023

fact, HDM is one of the most reasonable sources in terms of spatial (simulates sea level for coastal and offshore areas) and temporal coverage. One of its main limitations is that it tends to have an undisclosed vertical reference datum (Slobbe et al., 2014).

The key component that links such data sources is that of a high-resolution geoid (equipotential surface of the Earth). Most countries are limited in developing high-resolution geoid models due to a lack of vital technical expertise and expensive data equipment (ship- or airborne gravity measurements). An exception (amongst a few) is illustrated concerning the Baltic Sea countries; pursuing a collaboration among nations has resulted in a regionally calculated accurate geoid-model (NKG2015) (Ågren et al., 2016). This effort now paves the way for further developments in marine studies that were not possible before. Recent studies have also demonstrated that it is possible to correct HDM models utilizing geoid-referred tide gauge-stations along with accurate and high-resolution geoid model and interpolation methods (Jahanmard et al., 2021, 2022). This allows the derivation of dynamic topography (DT), which now represents a realistic sea-level, and can be used for forecasting and deriving ocean currents (Karimi et al., 2021; Pegliasco et al., 2021). This advancement in deriving accurate sea-level data (i.e. DT), combined with improvements in ML/DL now enables us to explore accurate SL forecasting for both coastal and offshore.

Several methods of ML/DL exist for sea-level predictions, which may be classified into two types. The first category consists of autoregressive models based solely on historically lagged measurements of sea-levels. Some of these methods have achieved a correlation coefficient of 0.9, 0.85 for 1-day and 2-day ahead predictions (Ali Ghorbani et al., 2010; Imani et al., 2018; Karimi et al., 2013; Kurniawan et al., 2014; Makarynskyy et al., 2004; Pashova and Popova, 2011). The second group relates the sea-level variations to physics-based (hydro-meteorological) characteristics. This approach performs well for longer-term predictions. The currently best physics-based model proposed to predict the relative sea-level has a correlation coefficient of 0.9 and forecasts the sea level one week ahead (Bellinghausen et al., 2023; Qiao et al., 2019).

Thus, based on the result of above-mentioned studies, both the autoregressive and physics-based ML model approaches appears to perform reasonably well. It is intuitive that sea level would be influenced by hydro-meteorological characteristics (e.g., atmospheric pressure, winds, tidal components, etc.). As indicated by a number of studies (Filippo et al., 2012; Liang et al., 2008; Makarynska and Makarynskyy, 2008; Tur et al., 2021), the outcomes of sea-level predictions are more interpretable when hydro-meteorological characteristics (e.g., atmospheric pressure, winds, tidal components, etc.) are also included as inputs. Such multivariate analysis can also assist to raise the model robustness, and in addition allows a better understanding on the extent of influence of the hydro-meteorological characteristics on sea level. Employing only historical values of target sea level as the training data (i.e auto-regressive approach) may also cause the model to be biased (Pour et al., 2020). Besides, the majority of such studies have employed ML/DL models to map relative sea-levels (like sea-level anomaly or extremes) with encouraging results (Balogun and Adebisi, 2021; Guillou and Chapalain, 2021; Tur et al., 2021). The setbacks being that in these studies, the analysis was limited to the location at tide gauge stations and the sea level was not referred to a vetical datum that could link with other sources of sea level in the offshore domain. This is unfortunate for the offshore areas also requires sea level forecasting for various application (e.g. engineering, navigation etc.).

In fact, the literature is still scarce for absolute sea-level predictions (i.e. sea surface height or DT). Recently, Song et al. (2021) set up a DL approach to forecast 15-day ahead sea surface height within the Southern China Sea. Within the same area, a DL method was adopted by Liu et al. (2022) to predict sea surface height, built on the attention mechanism by assiging more weights to important features. Such studies have paved the way for time-series forecasting of absolute sea-levels, yet solely based on autoregressive models. Another aspect as hinted above is that insufficient consideration has been given to the offshore domain, particularly in semi-enclosed seas, where its dynamics may significantly affect coastal areas. Consequently, forecasting sea level in open waters, particularly in areas with inadequate data, can help modelers better understand regional variations and as a consequence assists in developing more accurate models of global sea level change. As a result, numerous studies have recently conducted sea-level prediction for both coastal areas and the open sea (e.g., (Braakmann-Folgmann et al., 2017; Shao et al., 2022; Wang et al., 2022).

One of the challenging tasks in ML/DL implementation is identifying the most appropriate inputs (Primo de Siqueira and Paiva, 2021). This process should consider both the model's lower complexity and the dependent variable's explanation. For a semi-enclosed basin like Baltic Sea in North Europe, that is influenced by rapidly changing dynamic process (usually caused by intrinsic and extrinsic factors (Agha Karimi et al., 2021), it could be daunting to inspect the DT variability thoroughly. Whereas the interaction between DT and large-scale contributions like North Atlantic Ocean (NAO) remains complicated, previous studies have reported the most relevant features that serve as a valuable tracer for DT prediction, comprising wind speed, Sea Level Pressure and Sea Surface Temperature (SST) (Balogun and Adebisi, 2021; Bruneau et al., 2020; Guillou and Chapalain, 2021; Nieves et al., 2021). Furthermore, temporal information such as day of the year (DOY) has frequently been regarded as predictive indicators in data-driven approaches to explain seasonal variability (Janik et al., 2018; Kang et al., 2022; Rajabi-Kiasari and Hasanlou, 2020). These parameters in the model can be regarded as a surrogate to explain other remaining factors (Chen et al., 2020), such as precipitation, ice melting, NAO and ocean currents.

The validation of the final model is also an essential step in the ML/DL implementation. It is often accomplished using a different subset of training data, primarily through hold-out or cross-validation techniques (Jang et al., 2022). However, recent studies have suggested exploiting external data sources for spatiotemporal modeling marine parameters (Meyer et al., 2019; Stock, 2022). Utilization of external sea level data sources (e.g SA, marine buoys etc.) has the advantage of providing an independent verification and it can also quantify any biases that may exist in the final model. This also demonstrates the advantages and importance of having a common and accurate vertical reference for all utilized sea level sources.

Hence in this study, the primary goal is to spatio-temporally predict the one-step ahead (i.e. one-day ahead) DT of the Baltic Sea, utilizing both physically and historically-lagged inputs. These inputs include SST, winds, pressure, DOY, and DT. The forecasted results are then externally validated using along-track satellite altimetry data. To authors' knowledge, this is first examination on the spatial prediction of absolute sea-levels (i.e. DT) for the entire Baltic Sea.

The manuscript is presented as follows: Sections 2 and 3 describe the adopted methodology to derive DT and implementation strategy. Section 4 describes the physical characteristics of the Baltic Sea as the study area and datasets utilized. Section 5 presents the results from predictive performance and discussions on the implications are given in Section 6. Conclusions and future research suggestions are addressed in Section 7.

2. Background

2.1. Derivation of dynamic topography

As mentioned earlier, various sea-level sources may refer to a different vertical reference datum. Compared to other sources, HDMs provide data with the best spatiotemporal coverage; hence HDM appears to be appropriate for the forecasting objective of this study. However, it is often typical for the vertical reference datum of HDM sea-levels to be undisclosed. This indicates that a bias may exists in HDMs compared to other sources associated with plausible vertical reference datum. In Jahanmard et al. (2022), a novel method was developed that used geoid-referred tide gauges and interpolation methods to correct the

HDM sea-level data. This produced more realistic quantification of DT from HDMs.

The *DT* can also be estimated from any Sea Surface height (*SSH*) dataset, e.g., derived from SA. However, the SA technique measures the *SSH*, hence to calculate *DT* at a location of interest (gridpoint *s*) with coordinates (φ_s,λ_s) and the time instant *t*, a geoid (*N*) correction is given as follows:

$$DT_{(\varphi_s,\lambda_s,t)} = SSH_{(\varphi_s,\lambda_s,t)} - N_{(\varphi,\lambda)} \quad (1)$$

Note in section 5.3.3, we use SA data to validate the forecasted *DT*.

Whilst the HDM simulates mostly the long wavelength features of the sea level, possible integration with remote sensing (e.g. SA) and in-situ (tide gauges, Global Navigation Satellite System buoys etc.) sources may prove to be useful in capturing the small-scale dynamics overlooked by the hydrodynamic models (Jahanmard et al., 2022; Mostafavi et al., 2023). In this study, however, we first examine the applicability of using only the HDM with ML methods. This is followed by a validation with SA. For future studies, it may be possible to integrate the solution together with hydrodynamic models, remote sensing and/or in-situ sources.

2.2. Deep learning: Convolutional Neural Network (CNN) algorithm

DL is a subtype of ML that is a branch of Artificial Intelligence, in which, a computer learns to perform specific tasks based on the experiences it gains during training. The basic components of ML is based on data (as the input), a model (i.e a hypothesis) and a loss function. Whereby ML uses hypothesis maps to predict quantities of interest. The discrepancy (difference between prediction and observed) is quantified through a loss function. An iterative approach is used until the loss function is at a minimum.

DL is generated in the same way as ML, but it has many more levels, so that it attempts to function similar to brain, in that it can take an input, processes it and then make its own intuitive decisions/predictions. This, makes it ideal for large and nonlinear data processing. DL-based algorithms have recently absorbed attention as a powerful prediction tool and have been successfully applied in sea-level modelling (Balogun and Adebisi, 2021; Žust et al., 2021). Convolutional Neural Networks (CNNs) and Recurrent-based Neural Networks such as Long Short-Term Memory Networks have been well-known DL methods. CNN has been successfully applied in various regression prediction tasks, including water-depth prediction (Kabir et al., 2020), wave fields (Bai et al., 2022), and soil properties (Ng et al., 2019). CNN was first introduced by LeCun et al. (2015), and its network structure inspires brain's visual cortex. In this study, the CNN algorithm was utilized due to its superiority in using many nonlinear layers and preserving the spatial correlation between variables (Kamangir et al., 2021), making it appropriate for grid-like data. The general explanation of CNN is given in the next section.

2.2.1. 2D-Convolutional Neural Network (Conv2d)

The process of the Conv2d is basically that it takes input data and performs a convolution process (in a hidden layer that is described below), followed by classification/prediction outputs (that involves flattening and a fully connected layer), to produce a model output. In this study, a 2D CNN model for multivariate time-series forecasting was utilized by treating the input data similar as to that performed with images. A description is now summarized below:

There exist three primary layers (see Fig. 1) in the CNN:

(i) convolutional layer (filters + activation function): The convolution process is the most critical step and is basically a linear process that involves multiplication of input data with a set of weights (often referred to as filter or a kernel). Note, the filter contains both weights and biases that are generated during the training process of the data (see Eq. (3)). This filter is usually smaller than the input data and this is intentional as it allows the filter to be multiplied by the input array multiple times at different points on the input. Thus, the filtered output can capture low to high-level frequencies from the input data. Convolutions can have 1, 2, or 3 dimensions, and these convolutions are named based on the number of filter directions. We used a two-dimensional CNN (Conv2d) algorithm (see Fig. 1). Filters are sometimes followed by padding and stride layers for preserving the spatial resolution of the inputs and reducing the number of parameters in the network, respectively.

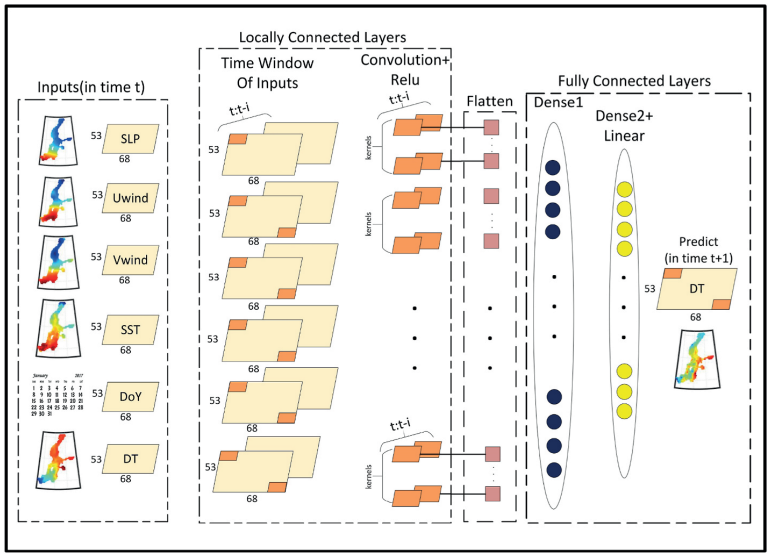


Fig. 1. The architecture of an original CNN model, adapted to the Baltic Sea DT determination. Time windowed SLP, u-wind, v-wind, SST, Day of Year (DOY), and DT from timestep *t-i* to *t* comprise the input data. Final forecasted DT field at timestep *t* +1 will be attained after performing convolutional, flattening and dense layers.

So far, the layers applied have taken on the approach of extracting linear properties. In reality, the ocean dynamic processes are nonlinear; thus, a nonlinear activation function (here, Rectified Linear Unit (ReLU)) is commonly applied that creates nonlinear patterns in neural networks (Nair and Hinton, 2010). The ReLU function is not a separate component of the Conv2d but a supplementary step (Fig. 1) and is usually applied after the convolution process. The training process in the ReLU function is simple and faster than other activation functions (Le et al., 2019; Taherisadr et al., 2018). If function receives any non-positive input, it returns 0; otherwise, it returns that value (x). Hence the ReLU function is formulated as follows:

$$g(x) = \begin{cases} x & if\ x > 0 \\ 0 & if\ x \le 0 \end{cases} \tag{2}$$

The output of a convolutional layer is as follows:

$$H_{i,j}^{k} = g\left(\sum_{l=1}^{L} \sum_{m=1}^{M} W_{l,m}^{(k)} X_{i+l-1,j+m-1}^{(l,m)} + b_k \right) \tag{3}$$

$$\widehat{DT}_{(\varphi_s,\lambda_s,t+1)} = F\left(DOY, pressure_{(\varphi_s,\lambda_s)}, uwind_{(\varphi_s,\lambda_s)}, vwind_{(\varphi_s,\lambda_s)}, SST_{(\varphi_s,\lambda_s)}, DT_{(\varphi_s,\lambda_s)} \right)_{(t-i:t)} \tag{5}$$

Where $X_{i+l-1,j+m-1}^{(l,m)}$ represents the l-th row and m-th column of the input data, $W_{l,m}^{(k)}$ represents the *k-th* filter, $H_{i,j}^{k}$ represents the feature map output by the *k-th* filter, b_k is a bias term, g is the activation function, and L and M are the dimensions of the filter.

(ii) Flattening layer: the extracted feature arrays from the previous layers, or so-called feature maps, are stacked in columns by flattening layer and given as a vector input to the fully connected layer. The output of the flattening layer is a one-dimensional vector with length equal to the product of input samples, and number of filters in the convolutional layer.

(iii) Fully Connected Layer: The flattened data is passed onto a neural network using a fully connected layer. Consequently, the model can locate spatio-temporal connections and discern between dominating and distinct low-level characteristics (Kow et al., 2020). Finally, the fully connected layer performs the neural network processing by minimizing the error term (loss function) between observed and forecasted DT.

The output of a dense layer is as follows:

$$H^{k} = f\left(\sum_{l=1}^{L} W_{l,k} X^{l} + b_k \right) \tag{4}$$

where X^l represents the lth input neuron, $W_{l,k}$ represents the weight connecting the l-th input neuron to the k-th output neuron, b_k is a bias term, and f is the activation function.

A more detailed description of each layer is provided in section 4.4 and Table 2.

3. Methodology

3.1. Computing environment

The CNN model was implemented with a Conv2D structure in open-source Python 3.9 (https://www.python.org) utilizing the Keras framework (2.9) (Chollet, 2015) and the Tensorflow backend (2.9.1) (Abadi et al., 2016).

3.2. Application of CNN to sea level data

The inputs include lagged measurements of the most influential physical features and DT data. The input features can be selected based on the previous studies and their potential impacts on DT. The daily scale was also chosen to infer all possible meteorological inputs retrievable in near-real-time so that it can allow the model to perform in operational-forecasting applications. To perform the modeling process, spatially and temporally consistent data are needed. The datasets used in this research were collected with different spatial resolutions. To create a regular grid of points, the gridpoints were generated at $0.25^{\circ} \times 0.25^{\circ}$ intervals using an interpolation strategy, which will be explained in Section 4.3.

Making consistent various data types, finally, the one-step ahead spatio-temporal prediction formula for a location φ_s, λ_s (e.g., each gridpoint) at time instant t (e.g., daily) can be written as

The duration term $(t-i:t)$ in the above equation refers to the historical measurements for each variable at point (φ_s, λ_s) with a time window of up to a predefined duration of i units (e.g., days in the case of daily estimates) till the time instant (e.g., current date), results in the spatial DT prediction ($\widehat{DT}$) in the next time step ($t+1$).

We aim to find the best function F by relating input parameters to the desired DT quantity that forecasts a future time instant (e.g., one day ahead). The flowchart in Fig. 2 illustrates the methodology employed. In the pre-processing stage, we first conducted data cleaning for corrected DT and ancillary data and then resampled input data to the output resolution both spatially and temporally. Secondly, we applied the necessary atmospheric and geophysical corrections and subtracted the geoid term from *SSH* (based on Eq. (1)) to derive SA-DT. In the third stage, we divided our final dataset into a train set for building a model and a test set to see model performance on a new independent dataset. Then, we applied the CNN algorithm to forecast DT and finally validated it spatially for various SA tracks (see Fig. 2).

3.3. Statistical criteria for model assessment

Three statistical indices were used for the performance assessment of CNN on DT prediction for the test set. The relations are provided in Eqs. (6)–(8).

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{s=1}^{m}\sum_{t=1}^{n}\left(\widehat{DT}_{(\varphi_s,\lambda_s,t)} - DT_{(\varphi_s,\lambda_s,t)}\right)^2}{\sum_{s=1}^{m}\sum_{t=1}^{n}\left(\widehat{DT}_{(\varphi_s,\lambda_s,t)} - \overline{DT}_{(\varphi_s,\lambda_s)}\right)^2}, where\ \overline{DT}_{(\varphi_s,\lambda_s)} = \frac{1}{n}\sum_{t=1}^{n} DT_{(\varphi_s,\lambda_s,t)} \tag{6}$$

where t denotes the time series duration from 1 to n. s is the gridpoint identifier (from 1 to m, where m is the amount of gridnodes); n refers to the quantity of the training and test days at each location (φ_s, λ_s). $DT_{(\varphi_s,\lambda_s,t)}$ denotes the observed DT at each location and time t, $\overline{DT}$ is the arithmetical mean of the observed $DT_{(\varphi_s,\lambda_s,t)}$ and $\widehat{DT}_{(\varphi_s,\lambda_s,t)}$ refers to the predicted DT over the period t by the DL model.

Mean Squared Error (*MSE*) of discrepancies at each gridpoint (φ_s,λ_s):

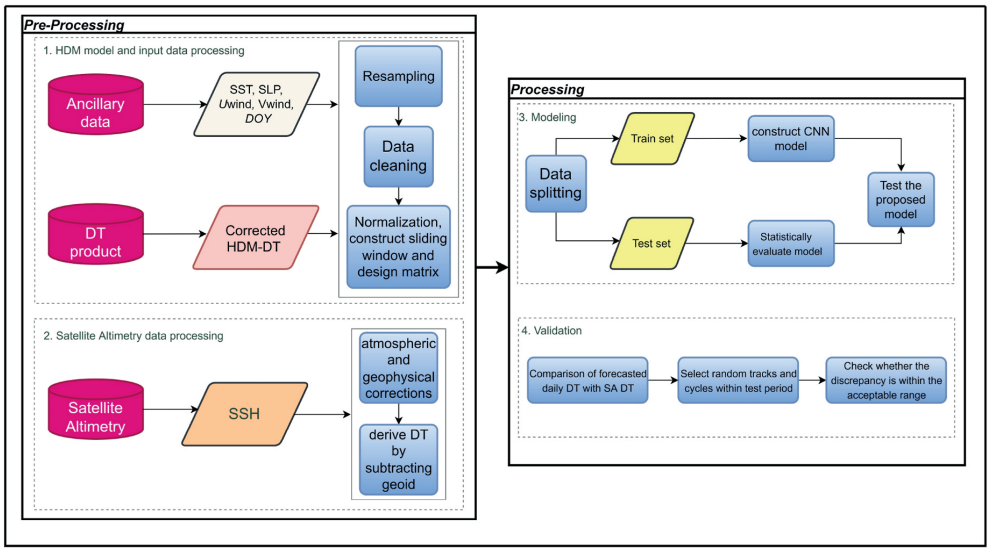


Fig. 2. The proposed framework to forecast DT shows the pre-processing, processing and validation phases.

$$MSE_{(\varphi_s,\lambda_s)} = \frac{\sum_{t=1}^{n}\left(\widehat{DT}_{(\varphi_s,\lambda_s,t)} - DT_{(\varphi_s,\lambda_s,t)}\right)^2}{n} \tag{7}$$

Spatial Root Mean Squared Error (*RMSE*) of discrepancies at each gridpoint (φ_s, λ_s):

$$RMSE_{(\varphi_s,\lambda_s)} = \sqrt{\frac{\sum_{t=1}^{n}\left(\widehat{DT}_{(\varphi_s,\lambda_s,t)} - DT_{(\varphi_s,\lambda_s,t)}\right)^2}{n}} \tag{8}$$

The goodness of a fit is defined as a larger R^2 and lower *RMSE*. Further theoretical aspects of the developed methodology will be explained in the context of the data used (Section 4.4).

4. Case study

4.1. Study site

The Baltic Sea is a microtidal and semi-enclosed sea in northern Europe with a mean depth of roughly 54 m and a total surface area of

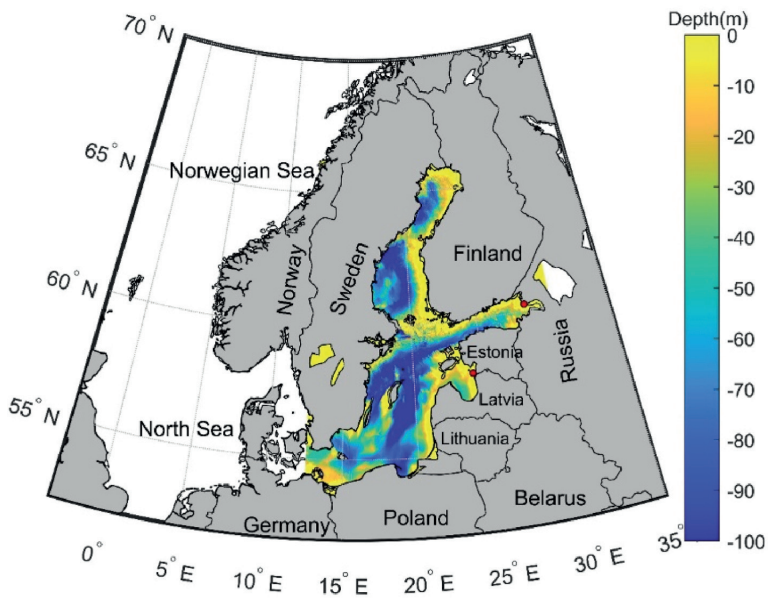


Fig. 3. The study area of the Baltic Sea showing the bathymetry. Red circles are the stations in the Gulf of Finland and Gulf of Riga that were used for the sea level maxima computation.

around 393,000 km^2 (Fig. 3). The geographical limits of Baltic Sea range from 53 °N to 66 °N and 10 °E to 30 °E. It is divided into several sub-basins based on bathymetry and geomorphology. The Danish Straits link Baltic Sea to the Atlantic Ocean through a narrow and shallow channel from which salt water penetrates into, whilst the Baltic Sea freshwater is sourced from numerous rivers from surrounding lands. Due to the estuarine environment, a stratified multi-layer structure often exists in most basins (Delpeche-Ellmann et al., 2021).

Several components based on different time frames affect the sea-level dynamics of the Baltic Sea (Suursaar and Sooäär, 2007). For example, long-term sea-level fluctuations are controlled by the thermal expansion of seawater and glacial melting, but temperature, precipitation, and evaporation have more significant impacts on a decadal timeframe. Short-term effects (yearly, seasonal, daily, etc.) are induced by variations in the hydrological cycle (such as transitory saltwater inflows from the Atlantic, river runoff, and sea ice). Relevant to this study are other localized occurrences during shorter periods (e.g., weekly, daily, and hourly) that are mainly impacted by meteorological conditions, particularly winds. Extreme sea levels as a result of storm surges frequently occur in Baltic by the interaction of stong winds, waves and pressure along with pre-filling of the basins (Leppäranta and Myrberg, 2009; Soomere et al., 2015; Wolski and Wiśniewski, 2014). These events also tend to be more prevalent during the autumn and winter seasons (Wolski and Wiśniewski, 2023). The inner regions of the northern and eastern Baltic Sea i.e Gulf of Bothnia, Gulf of Finland, and Gulf of Riga are particularly prone to experiencing longer periods of high sea levels (above 50 h and ≥70 cm relative to zero of tide gauges). In fact, during the period of 1960–2020, Pärnu Bay of the Gulf of Riga and the northern Gulf of Finland recorded 544 and 519 occasions of annual storm surges, respectively. This can be attributed to bathymetry and geometrical configuration of these coastal areas, and their exposure to prevailing low-pressure weather systems (Wolski and Wiśniewski, 2023).

The high-frequency changes caused by impact of winds and pressure gradients are mainly determined by internal meteorological variables. The wind is usually cited as the primary cause for sea-level fluctuations in the Baltic Sea (Kulikov et al., 2015; Passaro et al., 2015). Coastal upwellings are also common in the Baltic Sea, having seasonal patterns (Delpeche-Ellmann et al., 2017). In Hieronymus et al. (2017) and HÜnicke and Zorita (2008), it is revealed that pressure, which is regulated by NAO Index (Agha Karimi et al., 2021) is also significant in the Baltic Sea.

4.2. Data

The predictors include the lagged measurements of meteorological and temporal factors such as SST, *u*-wind, *v*-wind, pressure, DOY and DT. By a trial and error approach, we found that a window size of up to $i = 7$ days lag (in Eq. (5)) can produce more robust predictions. Table 1 describes the data used in this study and their statistical details.

4.2.1. Hydrodynamic model: dynamic topography

The original Nemo-Nordic (hereinafter Nemo-HDM) (NS01) is a 3D integrated ocean-ice model. It was developed by the Swedish Meteorological and Hydrological Institute (Hordoir et al., 2019) for the Baltic and North Seas. Nemo-HDM delivers hourly sea level measurements with a one nautical mile spatial resolution. As previously stated, Nemo-HDM is an appropriate source for our target DT variable due to its spatiotemporal coverage. However, the limitation with the Nemo-HDM data is that its vertical datum is not precisely defined. Jahanmard et al. (2022) established a methodology for correcting the Nemo-HDM in the Baltic Sea using a vast network of 73 geoid-referred tide gauge stations (typically assumed to be the most valid sea-level source) and interpolation methods. The corrected-HDM data showed an improvement by a scale factor of 1.5 compared to original Nemo-HDM and this was also validated by SA. This study uses the corrected-HDM DT.

4.2.2. Sea surface temperature (SST)

SST has been measured by satellite sensors since 1981 and when compared to other climatic variables, SST has the longest accessible satellite records. Different sensors' measurements were calibrated together in order to produce a Multi-scale Ultra-high Resolution-based SST. As a consequence, a global, gap-free, gridded, and consistent daily 0.01° SST dataset is derived, which was used in this study.

4.2.3. Wind speed and direction

The Cross-Calibrated Multi-Platform delivers a gridded-Level-4 (L4) wind-speed product across the world's seas. It is a composite of ocean surface (10 m) wind retrievals from several satellite microwave sensors, including the scatterometers QuikScat and ASCAT-A, as well as radiometers SSM/I, SSMIS, TMI, GMI, ASMR-E, AMSR2, and WindSat, and also a background field from reanalysis. The final product is spatially gap-filled and updated every 6 h. In this study, we used version 2, level-3 data with 0.25° spatial resolution. More information about data can be found at https://www.remss.com/measurements/ccmp/.

4.2.4. Sea level pressure

The atmospheric pressure at the sea surface is known as surface pressure. It is the weight of all the air in a vertical column above a point on Earth's surface. Pressure is frequently used in conjunction with temperature as an indicator of air density. We used ERA5 surface pressure data with a regular hourly grid of 0.25°. More information can be accessed at https://cds.climate.copernicus.eu/.

4.2.5. Sea surface height from satellite altimetry

SA products were used for the final validation of our DT forecast. The Baltic + SEAL (Passaro et al., 2021) Sentinel-3A high frequency (20 Hz

Table 1
Dataset used in the research from 2017 to 2019.

Group	Variable (unit)	Temporal resolution	Spatial resolution	Statistical summary					Data source
				Min.	Med.	Mean.	Max.	Std.	
Environmental features	Sea surface temperature (°C)	Daily	0.01° grid	−1.65	7.65	8.9	26	6.1	Multi-scale Ultra-high Resolution (MUR) https://cmr.earthdata.nasa.gov/search
	Uwind (zonal component of wind speed-m/s)	6-hourly	0.25°grid	−16.1	1.7	1.6	16.4	7.9	Cross-Calibrated Multi-Platform (CCMP) https://data.remss.com/ccmp/v02.0/
	Vwind (meridional component of wind speed-m/s)	6-hourly		−21.9	0.7	0.6	17.1	4.9	
	Surface air pressure (atm)	Hourly	0.25°grid	1.28	1.3	1.3	1.31	0.004	Copernicus https://marine.copernicus.eu/
DT	Corrected HDM-DT (m)	Hourly	1 nautical mile (1 min of latitude) grid	−0.8	0.34	0.34	1.45	0.2	Jahanmard et al. (2021)
Temporal feature	Day of the year (DOY)								feature engineering

and 300 m resolution) along-track sea surface height data, had been particularly developed for the Baltic Sea coastlines and sea-ice conditions were utilized. Such a product performed better (0.5–1 cm) when compared with standard products (Mostafavi et al., 2021). To obtain an accurate instantaneous sea surface height, along-track data was corrected for the atmospheric and geophysical corrections. We utilized the dynamic atmospheric corrections-decorrected SA data as discussed in (Jahanmard et al., 2022), since we are comparing SA with HDM data, which was not reduced for dynamic atmospheric correction.

4.3. Gridding and resampling

Given the availability of predictors (input data) at various temporal resolutions (see Table 1) and the focus of this study on daily DT forecasting, data were resampled to their arithmetical daily averages. In terms of location, there also exists inconsistency between input data and location of our forecasted DT. As a result, a gridding strategy was implemented in various steps, summarized by a flowchart in Fig. 4. To illustrate this, we selected a small fraction in Baltic Sea identified as a red box (Fig. 5a). A schematic of the interpolation technique within that area was shown in Fig. 5b. Finally removing NaN values available on land, the strategy resulted in 731 gridpoints for each day in the Baltic Sea. Each gridpoint encompasses a time series of daily DT.

4.4. Hyperparameter selection and training process

Hyperparameter selection is the process of optimizing the parameters inside a model (e.g., number of epochs, activation algorithms, etc. (Table 2)). If this selection is performed incorrectly, it may lead to a time-consuming training process. The parameters can be tuned using various methods, e.g., trial and error, grid search, or random search strategies. In this study for the CNN method employed, we did not employ a separate validation set inside the training set; instead, we validated the model for various dates and tracks utilizing SA measurements to check its generalization. Table 2 presents the hyperparameter and the recommended settings. The model was also initiated with a fixed seed number to guarantee the reproducibility of the results.

Table 3 contains model structure used to forecast DT. There are six features in our prediction problem (*DT, SST, pressure, DOY, u*-wind and *v*-wind). It must have four dimensions in order to fit a Conv2D model. The padding layers are set as the "same" since we wanted to keep the information. Moreover, based on different experimental tests, 24 filters were attained. As a result, considering $i = 7$ in Eq. (5), the flattened layer returns a 7*6*24 = 1008 length array. After being flattened, the proposed Conv2D model is then connected with two dense layers with size of 1000 and 100 neurons activated with a ReLU function and optimized by adaptive moment estimation Finally, the output prediction layer is obtained by an activated linear function.

During the model training, the weights of the network are updated using backpropagation to minimize the loss function (here MSE) between the predicted and actual values. Regarding the loss function, there are other functions than MSE, for instance mean absolute error, huber, quantile and some other. However, it is customary to use MSE as the loss function of regression problems for it uses only the prediction and actual values and if outliers exists it has a larger value on theses errors due to the squaring part of the function; It is also differentiable, convex and statistically interpretable function. Thus, we minimized this metric while evaluating the model training performance (cf. Eqs. (7) and (9)).

At each epoch of training, the weights of the network are updated using an optimization algorithm based on adaptive moment estimation. There are also different types of optimizers that can be applied such as Stochastic Gradient Descent (with/without momentum), Adaptive Gradient Algorithm and Root Mean Square Propagation. However, adaptive moment estimation optimizer can quickly learn, be robust and overcome the local minimum issues by stochastic gradient descent while updating the network weights. In addition, it has the ability to adaptively adjust the learning rate for each parameter (utilizing the estimated first and second moments of the gradients), making it the most popular optimizer for complex structures (Kumari and Toshniwal, 2021).

The optimization algorithm computes the gradient of the loss function with respect to the weights of the network and updates the weights in the direction that minimizes the loss through following equations:

$$\frac{\partial MSE}{\partial W_{l,k}} = \frac{2}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)\frac{\partial \hat{y}_i}{\partial W_{l,k}}, \tag{9}$$

$$W_{l,k} \leftarrow W_{l,k} - (learning\ rate)\frac{\partial MSE}{\partial W_{l,k}}$$

where *learning rate* is a hyperparameter that controls the size of the weight updates.

Then, trained Conv2D model is used to predict future values of the time series. To do this, we feed the past values of the time series into the network and obtain the predicted values for the future time step.

The loss curves of the Conv2D training process are shown in Fig. 6. The model has experienced a quick drop in loss value and convergence at the first steps, which can be due to employing the Adam optimizer. Then the model tries to learn till the last epoch gradually. After ten iterations, we observed the model's stability and a relatively constant *MSE* (0.0022 m).

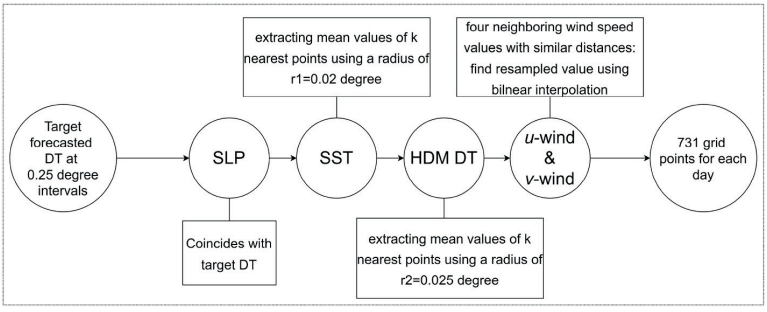


Fig. 4. Illustration of the gridding steps of various data sources, including SLP (sea level pressure), *u*&*v*-wind, SST (sea surface temperature), and HDM DT (corrected dynamic topography from HDM model) with different spatial resolutions with respect to our target forecasted DT. *r*1 and *r*2 are two radii considered for gridding SST and HDM DT, respectively.

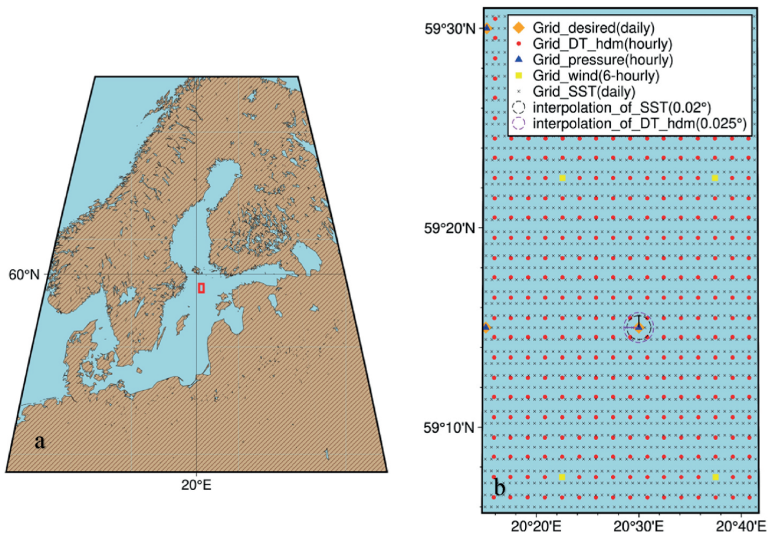


Fig. 5. Gridding strategy to interpolate input samples with respect to the desired location of interest DT for a small fraction in the Baltic Sea, a) selected rectangle (59.095–59.517°N, 20.245–20.695°E) b) magnified gridpoints, and interpolation strategy for the red box.

Table 2
Hyperparameters chosen for the training of the proposed Conv2D model. Other parameters were set to default values.

Model	Parameters	Value
Conv2D	Activation	{'ReLU'} and {'Linear'}
	Number of epochs	10
	Optimizer	{'Adam'}
	Loss function	Mean Squared Error (MSE)

Table 3
Model summary of the adopted Conv2d model. Three main layers in the CNN model include the convolution, flatten, and FCL.

Layer (Type)	Output shape	Parameters #
conv2d_1 (Conv2D)	(None, 7, 6, 24)	72
flatten_1 (Flatten)	(None, 1008)	0
dense_3 (Dense)	(None, 1000)	1009000 [(1008+bias)*1000]
dense_4 (Dense)	(None, 100)	100100 [(1000+bias)*100]
dense_5 (Dense)	(None, 1)	101 [(100+bias)*1]
Total parameters: 1,109,273		
Trainable parameters: 1,109,273		
Non-trainable parameters: 0		

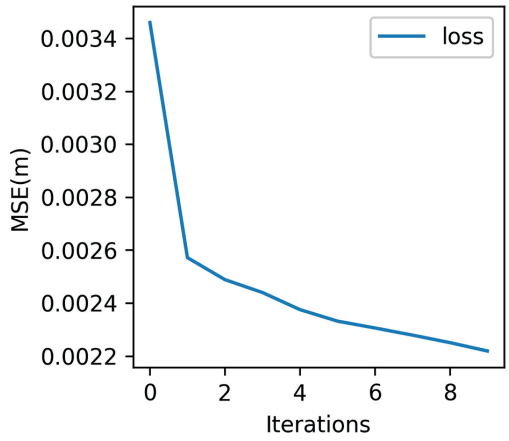


Fig. 6. Loss curve obtained from Conv2D training process shows the decreases in Mean square error (MSE) as the iterations increased.

5. Results

5.1. Correlation between input parameters

To examine the influence of the input parameters (variables in Eq. (2)) on target DT variable and with each other, spatial Pearson correlation maps were created (Fig. 7). The results compared DT with other inputs displayed distinct patterns. For instance, comparison of DT with wind components (u and v), showed both negative and positive correlations with values varying from −0.3 in southern Baltic Sea to 0.35 in eastern and northern areas (Fig. 7 a-b). This observation is realistic for winds, known to be influenced by the Coriolis effect and different coastline shapes and orientations, often triggering diverse marine dynamics (as an example, coastal upwellings), resulting in DT changes (Delpeche-Ellmann et al., 2017). A comparison of DT and pressure also had an essential influence, ranging from a substantial negative value of −0.48 for most sites to −0.05 near the Danish straits. The results of this comparison are also practical since the pressure is known to be influenced by NAO index, therefore, the sites located on the western sections of the Baltic Sea and thus closer to the North Atlantic shall be influenced more. With respect to the comparison of DT with SST, most of the gridpoints did not display any consistent relationship. Instead, SST appears to influence DT at specific regions in the middle parts of the Baltic Sea.

The interconnection between other input parameters was also examined (i.e. excluding DT, which was described above) in order to identify if any collinearity existed. Recall that in ML, too-high collinearity amongst inputs is not advised. Results (Fig. 7(e–j)) show that correlation of pressure and u-wind had the highest absolute value of 0.26. As a result, all suggested inputs were preserved in this study.

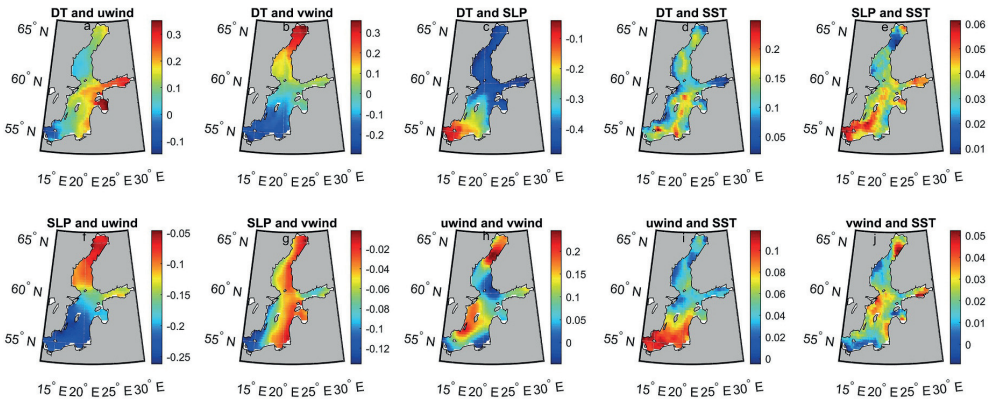


Fig. 7. Pearson correlation coefficients between meteorological features (DT, uwind, vwind, SLP, and SST).

5.2. Model set-up

Given the complexity of modelling ocean dynamics and the sometimes noisy environment, handling noisy data is an essential aspect of ML, and there are various techniques available to deal with it. Some of the most commonly used techniques for handling possible noise in the data for DL models are: normalization, data augmentation, dropout regularization, early stopping, and noise reduction techniques (Krizhevsky et al., 2017; Srivastava et al., 2014). In this study, the input data was normalized, this assists in reducing the impact of outliers and prepares the data to be more amenable to ML process. Moreover, it ensures that the activations within the network have similar scales and distributions, which makes the training process more stable and efficient. An early stopping procedure is also utilized to prevent overfitting and improve the generalization of the model to noisy data. This can help prevent the model from memorizing the noisy data and improve its ability to generalize to new unseen data. Given the importance of data normalization step in AI prediction task, the data was standardized using the StandardScaler (Pedregosa et al., 2011) function from python 3.9 Scikit-Learn (sklearn) package. Since this study examines three years of data (2017–2019) i.e. 1095 days, considering that there exist 731 spatial gridpoints for each day, altogether, there are 800445 sample points. The first 80% of the data (876 days, i.e., 640356 sample points) was utilized for model training, whereas the rest 20% (219 days, or 160089 sample points) was applied to assess its efficiency. The results of daily-averaged DT over the entire BS for train and test days are presented in Fig. 8. A seasonal trend that is highest in the summer with peak values of around 34.5–35 cm (e.g. July) and low in the winter months with values of 32–33 cm (e.g. January) is observable.

Table 4
Prediction performance of Conv2D model on training and test data for the Baltic Sea.

Model	Training data		Test data	
	R^2	RMSE (cm)	R^2	RMSE (cm)
Conv2D	0.95	4.7	0.91	4.7

5.3. Conv2D forecasting performance

5.3.1. Statistical evaluation of Conv2D model using R^2 and RMSE

Table 4 presents the predictability of the Conv2D model for training and test datasets in the Baltic Sea, where the achieved $R^2 = 0.95$ and 0.91 for train and test sets implies that the forecasted model for DT performs well. Also, the Conv2D model forecasted the training and test data by *RMSE* of 4.7 cm. This similarity supports the robustness of the adopted methodology. To investigate the proposed model's capabilities for spatial predictions, spatial *RMSE* plot for test data was shown in Fig. 9, from which most errors are between −4 cm and +4 cm. Some locations with larger values (*RMSE* of over 10 cm) exist, particularly in the BS's southeastern sections. Examination of the inputs strongly correlated with DT (Fig. 7) shows that both *u*-wind and SST may influence these locations. This study used a different SST and wind source than the Nemo-HDM simulation. Thus, these larger values may represent such differences in environmental sources or it could also be some input parameter that was not considered.

5.3.2. Performance at selected locations (actual vs forecasted)

To visually inspect the spatial performance of the proposed model, four arbitrary gridpoints were selected at stations C1 to C4 in the Baltic Sea (Fig. 10). The results showed that at all four stations: (i) the forecasted data (blue line) captured the behaviour of the actual DT values

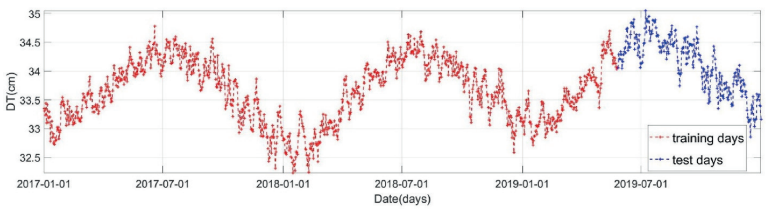


Fig. 8. Plot of daily averaged DT for the timespan 2017–2019 over the entire Baltic Sea, based on daily DT averaging of 731 sample points. Training (red) and test (days include 80% and 20% of data, respectively.

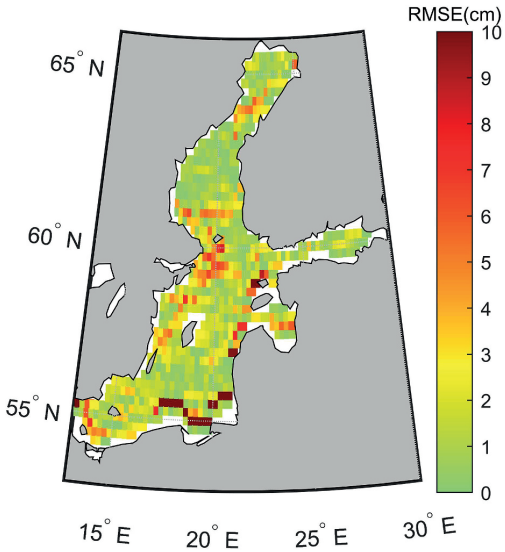


Fig. 9. Spatial 2D *RMSE* between the Conv2D predicted vs. actual DTs for the test data in the Baltic Sea.

reasonably well; (ii) the residuals (red line) varied from −5 to +5 cm and (iii) the highest residuals appeared to occur at C2 for the months July, September, and November. This may hint at either problem with one or several of the input sources for these months or with another input source (e.g. sea ice) that should have been considered.

The forecasted model's potential to replicate the pattern of observed DT throughout the entire Baltic Sea was examined through boxplots of Fig. 11. The box represents the third quartile, median value and lower quartile. The whiskers reflect the lowest and highest values. Both

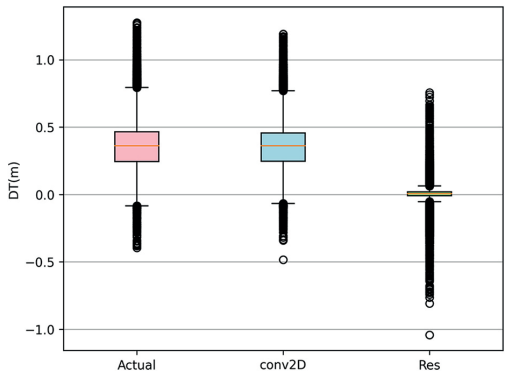


Fig. 11. Boxplot for comparison of proposed Conv2D model and observed DT along with the residuals for test set in the Baltic Sea.

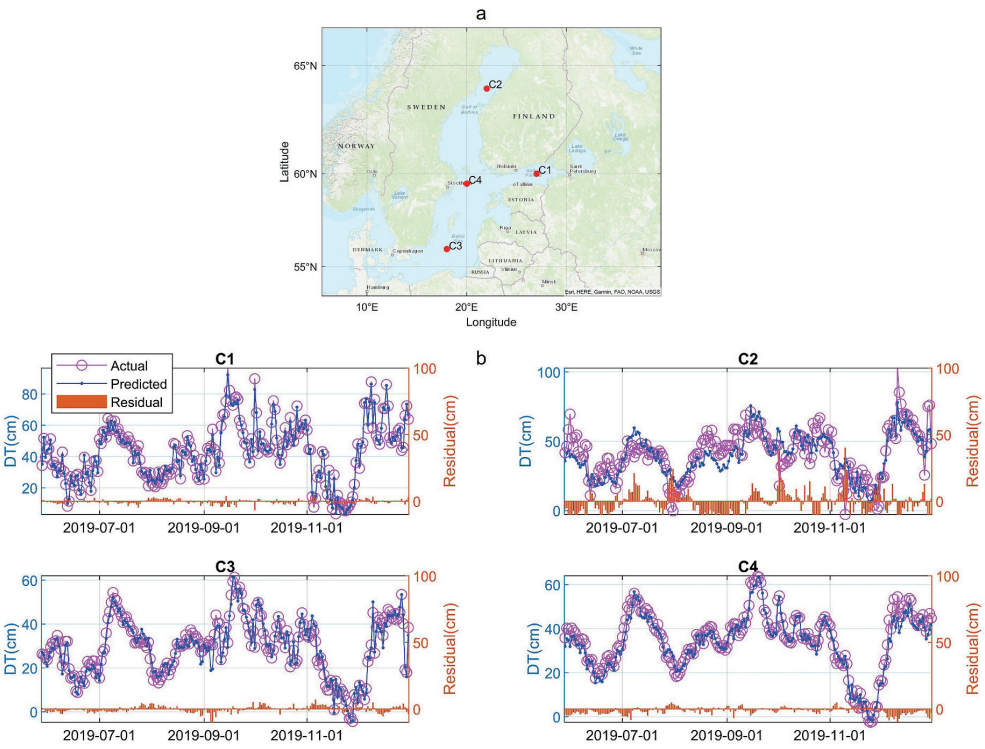


Fig. 10. a) Selected points to visually inspect the time series of actual vs. predicted DTs. b) Conv2D prediction on test data against the actual values for selected cells 1 to 4 from 27 May 2019 to 31 Dec 2019. C1 (60°N, 27°E -Gulf of Finland), C2 (64°N, 22°E -Bothnian Bay), C3 (56°N,18°E -Eastern Gotland Basin), and C4 (59.5°N, 20°E -Northern Baltic Proper). Test data points were also converted to their daily means to have such a pairwise comparison.

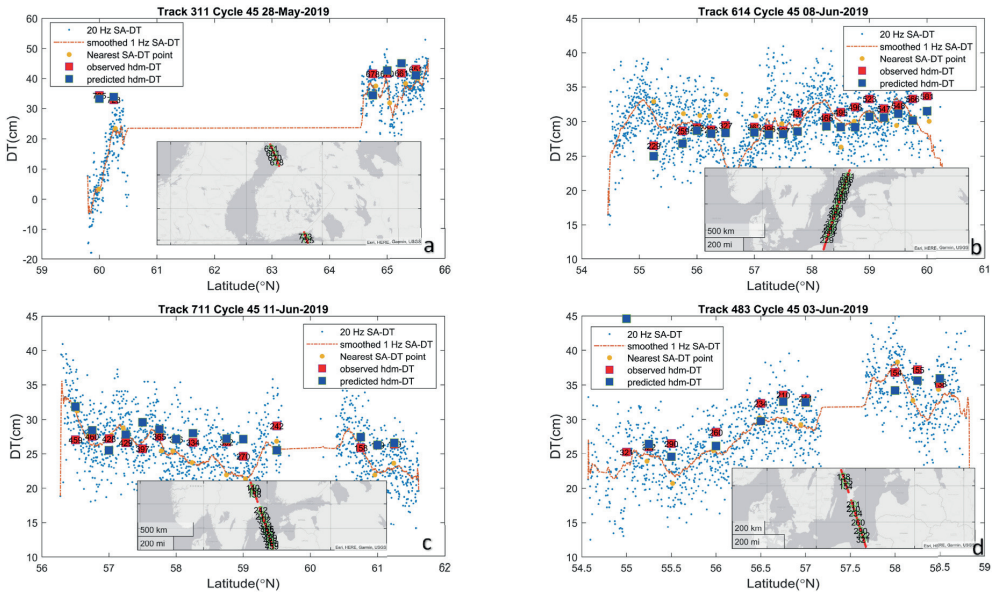


Fig. 12. Validation of the forecasted DT with the test DT and SA-DT for cycle 45 and a) track 311, 28-May 2019 b) track 614, 08-Jun 2019 c) track 711, 11-June 2019 d) track 483, 03-Jun 2019. The blue points show the original 20 Hz SA-DT and the red dashed lines are the smoothed 1 Hz SA-DT using a moving average method with a sample rate of 5%.

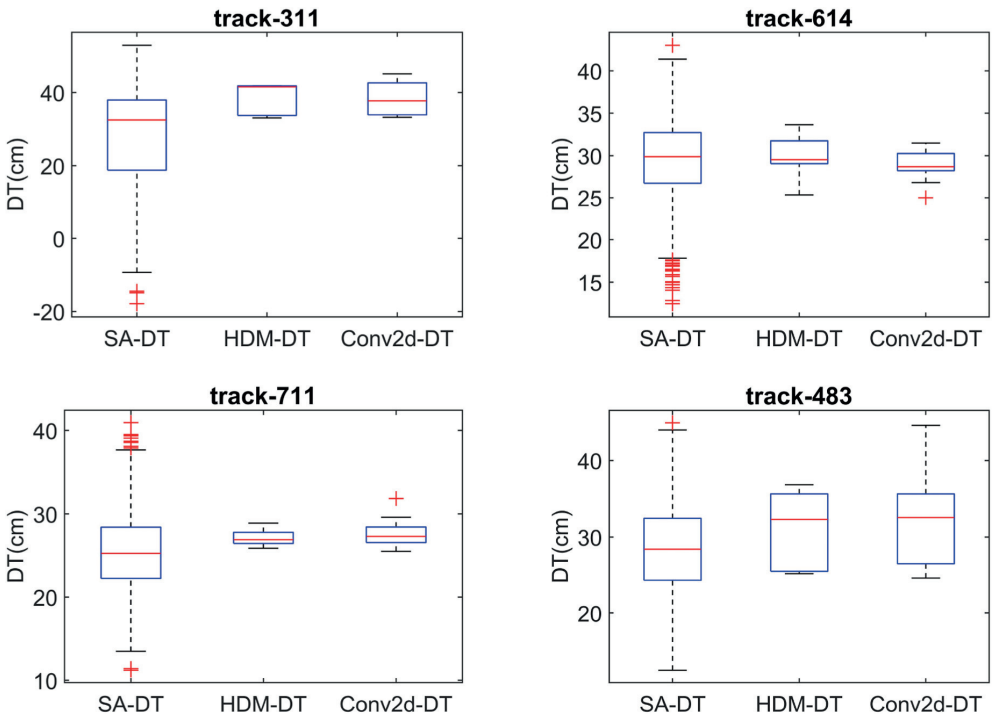


Fig. 13. Boxplot of the 1Hz SA data, HDM-DT, and the forecasted Conv2d-DT for tracks 311, 614, 711 and 483 in the Baltic Sea. Red pluses show the outlier points by each source.

observed and forecasted boxplot shapes are symmetric. The median value was around 36 cm for observed and forecasted DT. Moreover, the lower and upper quartiles, minimum and maximum values for the observed DT were attained as 24, 46.5, −40, and 128 cm, while for the forecasted DTs, such values are computed as 24, 46, −39, and 134 cm, respectively. We also computed the boxplot for the residuals; its median value fluctuates around zero. Very narrow lower and upper quartiles of 1 cm for residuals show an encouraging performance. Although the model produced some outliers in some cases ranging to 1 m discrepancy, considering the volume and size of the test data (160089), we found that their frequency, e.g., those larger than 15 cm, is around 1.96% of total samples. However, about 90% of the predictions are within ± 5 cm. From Fig. 11, it can also be seen that the Conv2d also predicted upper and lower quartiles and whiskers well. In general, the performance of the Conv2d model was reasonable in predicting the statistics.

5.3.3. *External validation of the proposed model by SA*

To spatially evaluate the effectiveness of the suggested Conv2d model, we compared the forecasted DT with HDM-based DT as well as DT values acquired via SA (Sentinel-3A). Four random tracks were selected within the test period, including the 311, 483, 614, and 711 (cycle 45). Fig. 12(a–d) compare the outcomes of several DT measurements, including our forecasted DT, observed DT from the corrected-HDM model, 20 Hz SA-DT, and also smoothed 1Hz SA-DT highlighted by nearest points to the gridpoints. Fig. 12a shows that, for track 311, except for gridpoint 725, which was presumably contaminated by lands

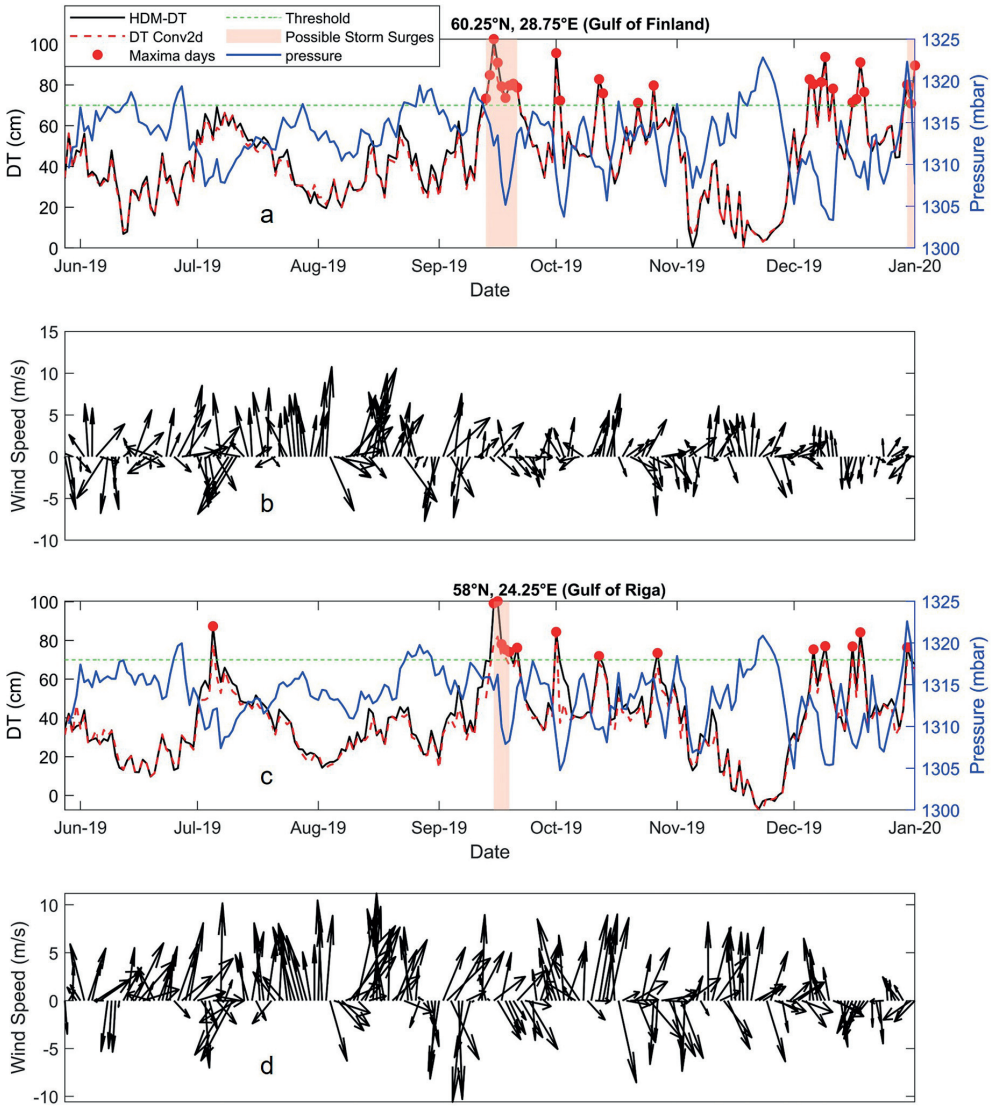


Fig. 14. The performance of the model on detection of sea level maxima events at two different locations in Baltic Sea during test days. Figures (a &b) and (c & d) show the time series of HDM-DT, modeled Conv2d DT, pressure and wind directions for points of interest in Gulf of Finland and Gulf of Riga, respectively. The red points, green dashed lines and shaded areas show the sea level maxima dates, considered threshold and possible storm surge events, respectively.

due to its closeness to the coastlines, forecasted DT points reflected both observed DT and the SA-DT trends with deviations of less than 10 cm.

In both Fig. 12b and c, for tracks 614 and 711, the DTs were forecasted with less than a 5 cm difference compared to SA measurements.

According to Fig. 12d, track 483 estimated DTs with less than 5 cm compared to 1 Hz SA; however, for gridpoint 321, the discrepancy between SA-DT and Conv2d-DT was around 20 cm. Considering Fig. 9, it can be evident that this gridpoint was likewise in the crucial zones that exhibited increased disparity for the whole test period, implying that this location may have been subject to a systematic inaccuracy.

It should be noted that SA offers instantaneous observations, whereas the present study utilized daily-averaged DTs from HDM. To investigate the applicability of SA data for validation of the proposed forecast model, the boxplot for the 1 Hz SA-DT, as well as the gridpoints of HDM-DT and the forecasted Conv2d-DT were computed. To avoid the impacts of distant regions, we only examined gridpoints within a radius of 0.1° of the SA tracks. For previously mentioned tracks, we visually compared several statistics such as extreme values, median, and range of variations. According to Fig. 13, the boxplot widths (data range) are more comparable for track 483, which encompasses middle Baltic Sea to south-eastern areas. For track 311, the boxplot revealed a distinct pattern for SA-DT than other sources since only a limited number of gridpoints were available. There is more consistency across the sources in the upper quartile (ranging from 39 to 40 cm), representing uncontaminated DTs. Although the median values for data types for tracks 311 and 614 are relatively similar (ranging from 25 to 27 cm and 28.5–30 cm, respectively), the smaller boxplot widths with a reduced data range and extremes suggest smoothed daily values compared to instantaneous SA observations, which may imply certain drawbacks when using the daily DT forecasts in such comparisons.

5.3.4. Model application on sea level maxima

In order to assess the effectiveness of proposed model in forecasting sea level maxima, we examined two distinct locations within the eastern Baltic Sea (Fig. 14). These locations consisted of: (i) Pärnu Bay located in the Gulf of Riga and known for having storm surges as high as +275 cm and (ii) the northern Gulf of Finland known for having storm surges as high as 200 cm (Kowalewski and Kowalewska-Kalkowska, 2017; Wolski et al., 2014). Based on previous studies performed in these areas by Wolski et al. (2014), a criterion of DT > 70 cm has been used to determine sea level maxima events.

From Fig. 14 (a & c), the maxima values can clearly be identified in both locations (red points), with maxima values ranged from 70 cm to 100 cm. These maxima values do not necessarily identify storm surges but instead can be due to a combination of different physical mechanisms. For instance, changes in the volume of sea water in the Baltic Sea, NAO, local storms etc. (Pindsoo and Soomere, 2020). These identified sea level maxima events occurred for both basins around the same time, with the most intense maxima occurring for September–October with a height of 100 cm and December of 2019 with a height of 90 cm. This agrees well with previous studies (Bellinghausen et al., 2023; Wolski and Wiśniewski, 2023). Atmospheric pressure and winds have been known to be indicators of possible storm surges (Kowalewski and Kowalewska-Kalkowska, 2017), thus examination of these additional data with the sea level maxima were also conducted for the same dates and locations. As Fig. 14 a and c represent, in some days, the peaks pattern in DT coincides well with depression in atmospheric pressure which could suggest possible storm surge events (shaded dates in Table 5 and Fig. 14).

Regarding the Gulf of Finland, during the test period, 27 sea level maxima days in HDM-DT were identified, of which 25 could be detected by the Conv2d model. The majority of these events lasted at least 2 days and the most significant one occurred from 13 to 21 September 2019. We also observed that from December 05, 2019 to end of the month, there were 12 maxima days lasting in shorter period (1 or 2 days) which can be as a result of sub-daily extreme events (e.g. possible storm surges) usually occurs in this month (Wolski and Wiśniewski, 2022). The modeled DT peaks reached up to 1 m and the Conv2D model showed a value of 97.6 cm, thus a difference of 4.9 cm exists (cf. Table 5 and Fig. 14).

Similarly, in Pärnu Bay, 15 sea level maxima days were identified in HDM-DT, of which 11 days were captured by the Conv2d model. Most of these events lasted 1 day, and the most significant one occurred from 15 to 19 September 2019, with the modeled DT values corresponding well to the observed high DTs.

Table 5 displays the dates of sea level maxima, the corresponding maximum DT records for HDM-DT and Conv2d DT, and the residuals for the two locations under study. The bold font dates indicate the maximum events that were captured by the Conv2d model. The proposed model successfully detected nine of the ten maximum events (bold dates) in the Gulf of Finland and eight of the eleven maxima events in the Gulf of Riga. For Gulf of Finland, the achieved results of maximum DT

Table 5
Information on sea level maxima dates, the corresponding maximum DT records for HDM-DT and Conv2d DT, the difference in DTs, pressure and wind speed ranges for two locations under study. The bold font and shaded dates highlight the maxima sea levels and possible storm surge events captured by the Conv2d model, respectively.

	Gulf of Finland						Gulf of Riga				
Sea level maxima events by HDM-DT (2019)	max HDM-DT (cm)	max Conv2d-DT (cm)	Diff in max-DT (cm)	Pressure range (mbar)	wind speed range (m/s)	Sea level maxima events by HDM-DT (2019)	max HDM-DT (cm)	max Conv2d-DT (cm)	Diff in DT (cm)	Pressure range (mbar)	wind speed range (m/s)
13-21 Sep	102.5	97.61	-4.89	1314.8-1305.2	0.88-8.56	**05 Jul**	87.36	78.32	-9.04	1312.2	7.92
01-02 Oct	95.60	88.85	-6.75	1305.3-1303.7	1.42-4.60	**15-19 Sep**	100.26	82.01	-18.25	1316.2-1307.9	5.06-8.44
12-13 Oct	82.79	77.36	-5.43	1309.3-1305.7	4.68-5.71	**21 Sep**	76.22	71.74	-4.48	1315.9	7.85
22 Oct	71.21	68.65	-2.56	1314.8	7.23	**01 Oct**	84.42	87.5	3.08	1304.8	4.64
26 Oct	79.73	77.88	-1.85	1312.2	7.05	12 Oct	71.95	67.95	-4	1309.9	8.13
05-06 Dec	82.77	79.70	-3.07	1309.4-1308.5	3.69-4.74	27 Oct	73.49	66.17	-7.32	1311.4	4.32
08-09 Dec	93.68	90.15	-3.53	1304.6-1303.4	2.26-2.95	**06 Dec**	75.39	70.79	-4.6	1311.7	6.96
11 Dec	78.17	76.91	-1.26	1311.1	3.11	**09 Dec**	77.05	71.58	-5.47	1305.4	4.24
16-19 Dec	91.08	90.30	-0.78	1310.3-1307	3.05-3.42	**16 Dec**	76.9	71.96	-4.94	1307.4	5.21
30-31 Dec	89.54	89.51	-0.03	1318.3-1307.7	2.25-6.12	**18 Dec**	84.19	78.02	-6.17	1311.6	5.76
						30 Dec	76.4	69.85	-6.55	1319.9	3.61

values and residuals (being up to 6.75 cm) indicate the successful performance of the proposed model in capturing maximum events.

As mentioned above, the possibility of storm surge events in the studied locations were performed by examining the pressure and wind speed patterns. Based on Table 5, we found that on two occurrences (13-21 Sep 2019 and 30-31 Dec 2019), the increase in DT corresponded with a sudden drop in pressure (~10 and 11 mbar) and a rise in wind speed magnitudes (from 0.9 to 8.6 m/s and 2.25–6.1 m/s) with a wind direction of (perpendicular to the coastline), which may suggest two storm surge events that the proposed model successfully tracked for Gulf of Finland (see shaded areas in Fig. 14 a).

For the Gulf of Riga, there was some indication of one possible storm surge event (15-19 Sep 2019) by negative pressure gradients (~8 mbar) and winds blowing to the south with speed of 8.4 m/s. The residuals mainly remain within 6 cm, but peaked at 9 and 18 cm (during possible storm surge event), signifying the complex structure of maximum events in the region and that there is still room for improvements in the Conv2D model. Furthermore, from Table 5 and Fig. 14 b and d, it was observed that the end of a maximum (or surge) event was accompanied by a sudden decrease in wind magnitudes, thereby confirming its contributing role in the continuation of a maximum event, in both regions.

In general, although the Conv2D model was able to accurately predict maximum sea level values up to 80 cm, it showed a smoothed pattern when capturing DT values larger than 80 cm. In the example of September–October, the Conv2D model underestimated the sea level in Gulf of Riga. This suggests that the DL has simulated a more complex structure of storm surge due to the input drivers that were given or even that perhaps another input driver was missing (e.g. waves, river discharge). Overall, the study demonstrates the promising potential of using DL to forecast sea level maxima with possible application for storm surges. It also shows that for better prediction of exteme events (e.g storm surges), additional input drivers should be considered e.g. waves. Such an application has not been extensively addressed yet.

6. Discussion

Most of the previous studies on sea-level forecasting utilized a limited number of points, often at the location of tide gauge stations. Instead, given the availability of realistic HDM-derived DT, this study employed a DL-based CNN method that considered various inputs (lagged DT, winds, pressure, SST, and DOY) and forecasted at 731 gridpoints throughout the whole Baltic Sea.

We obtained an *RMSE* of 4.7 cm for Conv2D model both for training and test sets. This similarity indicates promising results for the method employed. Concerning spatial predictive performance, a spatial *RMSE* was calculated (Fig. 9) with the majority of the discrepancies between −4 cm and +4 cm. Larger values (*RMSE* of >10 cm) are present in south-eastern parts of Baltic Sea. Reasons for this could be due to the input sources utilized. Based on investigation toward the identification of the relationship of DT with other inputs (Fig. 7), it was found that DT had a strong correlation with both the *u*-wind and SST. Since the original HDM used a different SST and wind source for simulation than that used as inputs in this study (e.g., Multi-scale Ultra-high Resolution SST and Cross-Calibrated Multi-Platform wind), these larger RMSE values and their locations possibly explain such differences. It could also be due to some other inputs that were not considered (e.g. river discharge, waves etc.).

The developed model's behavior was then examined at selected Baltic Sea stations, and this produced residuals within 5 cm. A seasonal trend of high residuals appears to have occurred in July, November and December, indicating possible problems with input sources or another source that should have been considered. In this study, limited inputs (what was known to be the most relevant) were utilized. In the future, however, other parameters (e.g., sea ice, river discharge, and waves) may also be utilized.

Examination of sea level maxima was performed at the Gulf of Finland and Gulf of Riga located in the eastern Baltic Sea. In most cases, the Conv2D model reproduced these maxima events (using a threshold of >70 cm), with the values reaching up to 100 cm. For the Gulf of Finland, 9 of total 10 maxima events were identified with two possible storm surges events. Whilst for Gulf of Riga, 8 of 11 events were identified with one possible occurrence of storm surge. Comparison of the original HDM model with the Conv2D model estimated the residuals in the Gulf of Finland being less than 7 cm and in the Gulf of Riga less than 18 cm. For the higher residuals, the Conv2D model showed a more smoothed pattern and underestimated the sea level. This suggests that other inputs sources should also be considered especially that of waves when exploring sea level maxima and storm surges and that the Conv2D model can still be improved.

External validation of the Conv2D forecasted model was also performed using along-track SA data. The comparison results showed discrepancies of less than 5 cm, confirming the accuracy of sea-level data from both HDM and SA. The SA data also showed more data points and variations than HDM (Fig. 12), indicating that the HDM model is more or less capable of simulating sea-level. The HDM sea-level data are still unrealistically smooth, whereas the SA tends to show more realistic variation that may indicate small-scale dynamics. This aspect can also be explored for future studies to incorporate the SA data in the Conv2D model, especially in missing data areas.

One important issue while developing ML forecasting models is the occurrence of hidden biases in the model. Here, we have attempted to address potential raises of hidden bias in different ways. One potential source of bias is the training data itself, which may be biased towards certain times of day if only certain measurements are used (Pour et al., 2020). To address this, we used a diverse set of features (e.g winds, atmospheric pressure, surface temperature etc.). Also, for validation of the output model results, an external source of sea level data from SA was utilized. So that if the biases existed, they would be revealed. Additionally, we addressed hidden bias in the algorithm by regularizing with early stopping during model training, and performing careful data pre-processing and feature engineering. Another useful tool for diagnosing bias is identifying residuals. If the residuals show a pattern or trend, it may indicate that the model is systematically under or over-estimating the target variable, thus forming a bias. However, we did not observe this in the results.

An important limitation of this study however is the temporal and spatial gridding strategy employed, for whilst HDM-DT was available at a better resolution of hourly-one nautical mile, actual grid used was that of 0.25° grid and daily averages. This was because of two main reasons. First, this study intended to explore the Conv2D with the accurate HDM-DT and use a lower gridding scheme. This combination still yielded promising results. Second, as our objective was to forecast DT field in Baltic Sea's selected gridpoints in the short term, we forecasted DT one day ahead using a multivariate multiple-location model. To generate such a model, we required a computationally efficient method. Accordingly, final resolutions were chosen to correspond with the alternative input sources of daily SST and 0.25° pressure and wind data. In the future, when addressing longer time-horizons, it would be advantageous to use higher spatiotemporal resolutions (e.g. the original HDM data points and hourly resolutions) by merging CNN algorithms with Long Short-Term Memory Networks-based models, which appear to perform better at longer temporal dependencies. Data-driven forecasting of sea levels are still in progress and of course there are room for improvements, to produce more interpretable models for example, using physic-informed models (Tausía et al., 2023) along with ensemble techniques, transfer learning and attention-based mechanisms.

7. Concluding remarks and future work

In this study, a DL technique under a Conv2d strategy was utilized to forecast daily DT of entire Baltic Sea. Several key inputs, such as winds, SST, pressure, DT and DOY were used. The training and test data both

obtained *RMSE* of 4.7 cm. Spatial analysis showed *RMSE* with most errors between −4 cm and +4 cm. Some locations of higher values were calculated in the Baltic Sea's south-eastern parts (*RMSE* of >10 cm). The reasons for these higher values may be linked to the source of wind and SST data utilized or other inputs that were not considered. Examination of sea level maxima in the eastern section of the Baltic Sea at the Gulf of Finland and the Gulf of Riga showed that the Conv2d model reproduced most of the maxima events, with residuals that were less than 7 cm for the Gulf of Finland and less than 18 cm for the Gulf of Riga. In the cases of higher residuals, the Conv2d tended to underestimate the sea level results. This implies that when examining storm surges, other input sources should also be considered e.g waves and that the Conv2d can still be improved. External validation of forecasted data was performed using SA, which showed a difference within 5 cm between the forecasted model and SA. SA data, however, showed more variations; Thus, possible future studies of DL can explore SA data in forecast solutions, especially for locations with missing data. Also, the proposed method has the potential to be combined with in-situ and SA data since this has the potential to analyze mesoscale processes e.g. eddy structures, resonant oscillations, and seiches.

Authorship statement

"SR; Conceptualization, Methodology & Analysis, Software, Visualization, Validation, Writing - Original Draft, ND; Conceptualization, Writing – review & editing, Validation, Supervision, AE; Conceptualization, Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition. All authors have reviewed and approved the published version of the work."

Code availability section

The source codes are available for download at the link: https://github.com/Saeed-Rajabi/CNN-DT.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The study was funded by the Estonian Research Council grants PRG1129 and PRG1785 "Development of continuous DYNAmic vertical REFerence for maritime and offshore engineering by applying machine learning strategies/DYNAREF/". The authors are thankful to Mr. V. Jahanmard for providing access to the corrected HDM model data. Two anonymous reviewers are also thanked for their constructive comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2023.105406.

References

Agha Karimi, A., Bagherbandi, M., Horemuz, M., 2021. Multidecadal Sea level variability in the Baltic Sea and its impact on acceleration estimations. Front. Mar. Sci. 8 https://doi.org/10.3389/fmars.2021.702512.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., 2016. Tensorflow: A system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp. 265–283.

Ågren, J., Strykowski, G., Bilker-Koivula, M., Omang, O., Märdla, S., Forsberg, R., Ellmann, A., Oja, T., Liepiņš, I., Paršeliūnas, E., Kaminskis, J., Sjöberg, L.E., Valsson, G., 2016. The NKG2015 gravimetric geoid model for the Nordic-Baltic region. In: 1st Joint Commission 2 and IGFS Meeting International Symposium on Gravity, Geoid and Height Systems, pp. 8–9.

Ali Ghorbani, M., Khatibi, R., Aytek, A., Makarynskyy, O., Shiri, J., 2010. Sea water level forecasting using genetic programming and comparing the performance with Artificial Neural Networks. Comput. Geosci. 36, 620–627. https://doi.org/10.1016/j.cageo.2009.09.014.

Bai, G., Wang, Z., Zhu, X., Feng, Y., 2022. Development of a 2-D deep learning regional wave field forecast model based on convolutional neural network and the application in South China Sea. Appl. Ocean Res. 118, 103012 https://doi.org/10.1016/j.apor.2021.103012.

Balogun, A.-L., Adebisi, N., 2021. Sea level prediction using ARIMA, SVR and LSTM neural network: assessing the impact of ensemble Ocean-Atmospheric processes on models' accuracy. Geomatics, Nat. Hazards Risk 12, 653–674. https://doi.org/10.1080/19475705.2021.1887372.

Bellinghausen, K., Hünicke, B., Zorita, E., 2023. Short-term prediction of extreme sea-level at the Baltic Sea coast by random forests. Nat. Hazards Earth Syst. Sci. Discuss. https://doi.org/10.5194/nhess-2023-21 [preprint] in review.

Braakmann-Folgmann, A., Roscher, R., Wenzel, S., Uebbing, B., Kusche, J., 2017. sea level anomaly prediction using recurrent neural networks. In: Proceedings of the 2017 Conference on Big Data from Space.

Bruneau, N., Polton, J., Williams, J., Holt, J., 2020. Estimation of global coastal sea level extremes using neural networks. Environ. Res. Lett. 15, 074030 https://doi.org/10.1088/1748-9326/ab89d6.

Chen, B., Liu, H., Xiao, W., Wang, L., Huang, B., 2020. A machine-learning approach to modeling picophytoplankton abundances in the South China Sea. Prog. Oceanogr. 189, 102456 https://doi.org/10.1016/j.pocean.2020.102456.

Chollet, F., 2015. https://github.com/fchollet/keras.

Delpeche-Ellmann, N., Giudici, A., Rätsep, M., Soomere, T., 2021. Observations of surface drift and effects induced by wind and surface waves in the Baltic Sea for the period 2011–2018. Estuar. Coast Shelf Sci. 249, 107071 https://doi.org/10.1016/j.ecss.2020.107071.

Delpeche-Ellmann, N., Mingelaitė, T., Soomere, T., 2017. Examining Lagrangian surface transport during a coastal upwelling in the Gulf of Finland, Baltic Sea. J. Mar. Syst. 171, 21–30. https://doi.org/10.1016/j.jmarsys.2016.10.007.

Filippo, A., Rebelo Torres, A., Kjerfve, B., Monat, A., 2012. Application of Artificial Neural Network (ANN) to improve forecasting of sea level. Ocean Coast Manag. 55, 101–110. https://doi.org/10.1016/j.ocecoaman.2011.09.007.

Guillou, N., Chapalain, G., 2021. Machine learning methods applied to sea level predictions in the upper part of a tidal estuary. Oceanologia 63, 531–544. https://doi.org/10.1016/j.oceano.2021.07.003.

Hieronymus, M., Hieronymus, J., Arneborg, L., 2017. Sea level modelling in the Baltic and the North Sea: the respective role of different parts of the forcing. Ocean Model. 118, 59–72. https://doi.org/10.1016/j.ocemod.2017.08.007.

Hordoir, R., Axell, L., Höglund, A., Dieterich, C., Fransner, F., Gröger, M., Liu, Y., Pemberton, P., Schimanke, S., Andersson, H., Ljungemyr, P., Nygren, P., Falahat, S., Nord, A., Jönsson, A., Lake, I., Döös, K., Hieronymus, M., Dietze, H., Löptien, U., Kuznetsov, I., Westerlund, A., Tuomi, L., Haapala, J., 2019. Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas – research and operational applications. Geosci. Model Dev. (GMD) 12, 363–386. https://doi.org/10.5194/gmd-12-363-2019.

Hünicke, B., Zorita, E., 2008. Trends in the amplitude of Baltic Sea level annual cycle. Tellus Dyn. Meteorol. Oceanogr. 60, 154–164. https://doi.org/10.1111/j.1600-0870.2007.00277.x.

Imani, M., Kao, H.-C., Lan, W.-H., Kuo, C.-Y., 2018. Daily sea level prediction at Chiayi coast, Taiwan using extreme learning machine and relevance vector machine. Global Planet. Change 161, 211–221. https://doi.org/10.1016/j.gloplacha.2017.12.018.

Jahanmard, V., Delpeche-Ellmann, N., Ellmann, A., 2022. Towards realistic dynamic topography from coast to offshore by incorporating hydrodynamic and geoid models. Ocean Model., 102124 https://doi.org/10.1016/j.ocemod.2022.102124.

Jahanmard, V., Delpeche-Ellmann, N., Ellmann, A., 2021. Realistic dynamic topography through coupling geoid and hydrodynamic models of the Baltic Sea. Continent. Shelf Res. 222, 104421 https://doi.org/10.1016/j.csr.2021.104421.

Jang, E., Kim, Y.J., Im, J., Park, Y.-G., Sung, T., 2022. Global sea surface salinity via the synergistic use of SMAP satellite and HYCOM data based on machine learning. Remote Sens. Environ. 273, 112980 https://doi.org/10.1016/j.rse.2022.112980.

Janik, M., Bossew, P., Kurihara, O., 2018. Machine learning methods as a tool to analyse incomplete or irregularly sampled radon time series data. Sci. Total Environ. 630, 1155–1167. https://doi.org/10.1016/j.scitotenv.2018.02.233.

Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., Pender, G., 2020. A deep convolutional neural network model for rapid prediction of fluvial flood inundation. J. Hydrol. 590, 125481 https://doi.org/10.1016/j.jhydrol.2020.125481.

Kamangir, H., Collins, W., Tissot, P., King, S.A., Dinh, H.T.H., Durham, N., Rizzo, J., 2021. FogNet: a multiscale 3D CNN with double-branch dense block and attention mechanism for fog prediction. Mach. Learn. with Appl. 5, 100038 https://doi.org/10.1016/j.mlwa.2021.100038.

Kang, Y., Kim, M., Kang, E., Cho, D., Im, J., 2022. Improved retrievals of aerosol optical depth and fine mode fraction from GOCI geostationary satellite data using machine learning over East Asia. ISPRS J. Photogrammetry Remote Sens. 183, 253–268. https://doi.org/10.1016/j.isprsjprs.2021.11.016.

Karimi, A.A., Andersen, O.B., Deng, X., 2021. Mean sea surface and mean dynamic topography determination from Cryosat-2 data around Australia. Adv. Space Res. 68, 1073–1089. https://doi.org/10.1016/j.asr.2020.01.009.
Karimi, S., Kisi, O., Shiri, J., Makarynskyy, O., 2013. Neuro-fuzzy and neural network techniques for forecasting sea level in Darwin Harbor, Australia. Comput. Geosci. 52, 50–59. https://doi.org/10.1016/j.cageo.2012.09.015.
Kow, P.-Y., Wang, Y.-S., Zhou, Y., Kao, I.-F., Issermann, M., Chang, L.-C., Chang, F.-J., 2020. Seamless integration of convolutional and back-propagation neural networks for regional multi-step-ahead PM2.5 forecasting. J. Clean. Prod. 261, 121285 https://doi.org/10.1016/j.jclepro.2020.121285.
Kowalewski, M., Kowalewska-Kalkowska, H., 2017. Sensitivity of the Baltic Sea level prediction to spatial model resolution. J. Mar. Syst. 173, 101–113. https://doi.org/10.1016/j.jmarsys.2017.05.001.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. https://doi.org/10.1145/3065386.
Kulikov, E.A., Medvedev, I.P., Koltermann, K.P., 2015. Baltic sea level low-frequency variability. Tellus Dyn. Meteorol. Oceanogr. 67, 25642 https://doi.org/10.3402/tellusa.v67.25642.
Kumari, P., Toshniwal, D., 2021. Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. Appl. Energy 295, 117061. https://doi.org/10.1016/j.apenergy.2021.117061.
Kurniawan, A., Ooi, S.K., Babovic, V., 2014. Improved sea level anomaly prediction through combination of data relationship analysis and genetic programming in Singapore Regional Waters. Comput. Geosci. 72, 94–104. https://doi.org/10.1016/j.cageo.2014.07.007.
Le, N.Q.K., Yapp, E.K.Y., Ou, Y.-Y., Yeh, H.-Y., 2019. iMotor-CNN: identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. Anal. Biochem. 575, 17–26. https://doi.org/10.1016/j.ab.2019.03.017.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.
Leppäranta, M., Myrberg, K., 2009. Physical Oceanography of the Baltic Sea. Springer.
Liang, S.X., Li, M.C., Sun, Z.C., 2008. Prediction models for tidal level including strong meteorologic effects using a neural network. Ocean Eng. 35, 666–675. https://doi.org/10.1016/j.oceaneng.2007.12.006.
Liu, J., Jin, B., Wang, L., Xu, L., 2022. Sea surface height prediction with deep learning based on attention mechanism. Geosci. Rem. Sens. Lett. IEEE 19, 1–5. https://doi.org/10.1109/LGRS.2020.3039062.
Makarynska, D., Makarynskyy, O., 2008. Predicting sea-level variations at the Cocos (Keeling) Islands with artificial neural networks. Comput. Geosci. 34, 1910–1917. https://doi.org/10.1016/j.cageo.2007.12.004.
Makarynskyy, O., Makarynska, D., Kuhn, M., Featherstone, W.E., 2004. Predicting sea level variations with artificial neural networks at Hillarys Boat Harbour, Western Australia. Estuar. Coast Shelf Sci. 61, 351–360. https://doi.org/10.1016/j.ecss.2004.06.004.
Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. Ecol. Model. 411, 108815 https://doi.org/10.1016/j.ecolmodel.2019.108815.
Mostafavi, M., Delpeche-Ellmann, N., Ellmann, A., 2021. Accurate Sea surface heights from sentinel-3A and jason-3 retrackers by incorporating high-resolution marine geoid and hydrodynamic models. J. Geod. Sci. 11, 58–74. https://doi.org/10.1515/jogs-2020-0120.
Mostafavi, M., Delpeche-Ellmann, N., Ellmann, A., Jahanmard, V., 2023. Determination of accurate dynamic topography for the Baltic Sea using satellite altimetry and a marine geoid model. Rem. Sens. 15, 2189. https://doi.org/10.3390/rs15082189.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp. 807–814.
Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251–267. https://doi.org/10.1016/j.geoderma.2019.06.016.
Nieves, V., Radin, C., Camps-Valls, G., 2021. Predicting regional coastal sea level changes with machine learning. Sci. Rep. 11, 7650. https://doi.org/10.1038/s41598-021-87460-z.
Pashova, L., Popova, S., 2011. Daily sea level forecast at tide gauge Burgas, Bulgaria using artificial neural networks. J. Sea Res. 66, 154–161. https://doi.org/10.1016/j.seares.2011.05.012.
Passaro, M., Cipollini, P., Benveniste, J., 2015. Annual sea level variability of the coastal ocean: the Baltic Sea-North Sea transition zone. J. Geophys. Res. Ocean. 120, 3061–3078. https://doi.org/10.1002/2014JC010510.
Passaro, M., Müller, F., Dettmering, D., Abulaitijiang, A., Scarrott, R., Chalençon, E., Sweeney, M., 2021. Baltic+ SEAL: Product Handbook. Frascati. https://doi.org/10.5270/esa.BalticSEAL.PH1.1.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pegliasco, C., Chaigneau, A., Morrow, R., Dumas, F., 2021. Detection and tracking of mesoscale eddies in the mediterranean sea: a comparison between the Sea Level anomaly and the absolute dynamic topography fields. Adv. Space Res. 68, 401–419. https://doi.org/10.1016/j.asr.2020.03.039.
Pindsoo, K., Soomere, T., 2020. Basin-wide variations in trends in water level maxima in the Baltic Sea. Continent. Shelf Res. 193, 104029 https://doi.org/10.1016/j.csr.2019.104029.
Pour, S.H., Wahab, A.K.A., Shahid, S., 2020. Physical-empirical models for prediction of seasonal rainfall extremes of Peninsular Malaysia. Atmos. Res. 233, 104720 https://doi.org/10.1016/j.atmosres.2019.104720.
Primo de Siqueira, B.V., Paiva, A. de M., 2021. Using neural network to improve sea level prediction along the southeastern Brazilian coast. Ocean Model. 168, 101898 https://doi.org/10.1016/j.ocemod.2021.101898.
Qiao, F., Wang, G., Khokiattiwong, S., Akhir, M.F., Zhu, W., Xiao, B., 2019. China published ocean forecasting system for the 21st-Century Maritime Silk Road on December 10, 2018. Acta Oceanol. Sin. 38, 1–3. https://doi.org/10.1007/s13131-019-1365-y.
Rajabi-Kiasari, S., Hasanlou, M., 2020. An efficient model for the prediction of SMAP sea surface salinity using machine learning approaches in the Persian Gulf. Int. J. Rem. Sens. 41, 3221–3242. https://doi.org/10.1080/01431161.2019.1701212.
Shao, Q., Li, W., Hou, G., Han, G., Wu, X., 2022. Mid-term simultaneous spatiotemporal prediction of Sea Surface height anomaly and Sea Surface temperature using satellite data in the south China sea. Geosci. Rem. Sens. Lett. IEEE 19, 1–5. https://doi.org/10.1109/LGRS.2020.3042179.
Slobbe, D.C., Klees, R., Gunter, B.C., 2014. Realization of a consistent set of vertical reference surfaces in coastal areas. J. Geodyn. 88, 601–615. https://doi.org/10.1007/s00190-014-0709-9.
Song, T., Han, N., Zhu, Y., Li, Z., Li, Y., Li, S., Peng, S., 2021. Application of deep learning technique to the sea surface height prediction in the South China Sea. Acta Oceanol. Sin. 40, 68–76. https://doi.org/10.1007/s13131-021-1735-0.
Soomere, T., Eelsalu, M., Kurkin, A., Rybin, A., 2015. Separation of the Baltic Sea water level into daily and multi-weekly components. Continent. Shelf Res. 103, 23–32. https://doi.org/10.1016/j.csr.2015.04.018.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Stock, A., 2022. Spatiotemporal distribution of labeled data can bias the validation and selection of supervised learning algorithms: a marine remote sensing example. ISPRS J. Photogrammetry Remote Sens. 187, 46–60. https://doi.org/10.1016/j.isprsjprs.2022.02.023.
Suursaar, Ü., Sooäär, J., 2007. Decadal variations in mean and extreme sea level values along the Estonian coast of the Baltic Sea. Tellus Dyn. Meteorol. Oceanogr. 59, 249. https://doi.org/10.1111/j.1600-0870.2006.00220.x.
Taherisadr, M., Asnani, P., Galster, S., Dehzangi, O., 2018. ECG-based driver inattention identification during naturalistic driving using Mel-frequency cepstrum 2-D transform and convolutional neural networks. Smart Heal 9 (10), 50–61. https://doi.org/10.1016/j.smhl.2018.07.022.
Tausía, J., Delaux, S., Camus, P., Rueda, A., Méndez, F., Bryan, K.R., Pérez, J., Costa, C.G. R., Zyngfogel, R., Cofiño, A., 2023. Rapid response data-driven reconstructions for storm surge around New Zealand. Appl. Ocean Res. 133, 103496 https://doi.org/10.1016/j.apor.2023.103496.
Tur, R., Tas, E., Haghighi, A.T., Mehr, A.D., 2021. Sea level prediction using machine learning. Water 13, 3566. https://doi.org/10.3390/w13243566.
Wang, G., Wang, X., Wu, X., Liu, K., Qi, Y., Sun, C., Fu, H., 2022. A hybrid multivariate deep learning network for multistep ahead Sea Level anomaly forecasting. J. Atmos. Ocean. Technol. 39, 285–301. https://doi.org/10.1175/JTECH-D-21-0043.1.
Wolski, T., Wiśniewski, B., 2023. Characteristics of seasonal changes of the Baltic Sea extreme sea levels. Oceanologia 65, 151–170. https://doi.org/10.1016/j.oceano.2022.02.006.
Wolski, T., Wiśniewski, B., 2022. Characteristics of seasonal changes of the Baltic Sea extreme sea levels. Oceanologia. https://doi.org/10.1016/j.oceano.2022.02.006.
Wolski, T., Wiśniewski, B., 2014. Long-term, seasonal and short-term fluctuations in the water level of the southern Baltic Sea. Quaest. Geogr. 33, 181–197. https://doi.org/10.2478/quageo-2014-0041.
Wolski, T., Wiśniewski, B., Giza, A., Kowalewska-Kalkowska, H., Boman, H., Grabbi-Kaiv, S., Hammarklint, T., Holfort, J., Lydeikaitė, Ž., 2014. Extreme sea levels at selected stations on the Baltic Sea coast. Oceanologia 56, 259–290. https://doi.org/10.5697/oc.56-2.259.
Zhou, Y., Lu, C., Chen, K., Li, X., 2023. Sea surface height anomaly prediction based on artificial intelligence. In: Artificial Intelligence Oceanography. Springer Nature Singapore, Singapore, pp. 63–82. https://doi.org/10.1007/978-981-19-6375-9_3.
Žust, L., Fettich, A., Kristan, M., Ličer, M., 2021. HIDRA 1.0: deep-learning-based ensemble sea level forecasting in the northern Adriatic. Geosci. Model Dev. (GMD) 14, 2057–2074. https://doi.org/10.5194/gmd-14-2057-2021.

Appendix 2

Publication II

Rajabi-Kiasari, S., Ellmann, A., & Delpeche-Ellmann, N. (2025). Sea level forecasting using deep recurrent neural networks with high-resolution hydrodynamic model. Applied Ocean Research, 157, 104496, doi: doi.org/10.1016/j.apor.2025.104496.

Applied Ocean Research 157 (2025) 104496

Contents lists available at ScienceDirect

Applied Ocean Research

journal homepage: www.elsevier.com/locate/apor

Sea level forecasting using deep recurrent neural networks with high-resolution hydrodynamic model

Saeed Rajabi-Kiasari [a,*], Artu Ellmann [a], Nicole Delpeche-Ellmann [b]

[a] Department of Civil Engineering and Architecture, Tallinn University of Technology, Ehitajate road 5, 19086 Tallinn, Estonia
[b] Department of Cybernetics, School of Science, Tallinn University of Technology, Ehitajate road 5, 19086 Tallinn, Estonia

ARTICLE INFO

Keywords:
Sea level forecasting
LSTM
GRU
Baltic sea
Deep learning
Recurrent neural networks
Spatiotemporal prediction
Gulf of Finland
Hydro-geodesy

ABSTRACT

Changes in climate, along with increasing marine activities in coastal and offshore regions, highlight the need for effective sea level forecasting methods. In recent years, forecasting techniques, especially those utilizing machine learning/deep learning methods (ML/DL), have shown promising capabilities. However, sea level forecasting is often limited in accuracy and spatiotemporal coverage, primarily due to the challenges posed by available observational data, which complicates the assessment of existing ML/DL techniques in complex and dynamic regions like the Baltic Sea. This study addresses these challenges by utilizing a high-resolution spatiotemporal framework that integrates high-resolution hydrodynamic and marine geoid models available to Baltic countries, enabling further capabilities to be explored in terms of sea level accuracy and validation. Specifically, it examines short-term sea level forecasting in the eastern Baltic Sea and the potential of utilizing two recurrent neural network-based models such as the Long Short-Term Memory Networks (LSTMs), and the Gated Recurrent Unit (GRU) along with high-resolution input data sources. These models were specifically chosen, due to their expected capabilities with time series data and their ability to learn both short and long-term connections of the input datasets.

To achieve this, a multivariate multistep-ahead (3, 6, 9, 12, and 24 h) forecasting framework was developed. The DL models' input components are high-resolution sea level data obtained from a bias-corrected hydrodynamic model, wind speed, surface pressure, and sea surface temperature. Results for various time steps (from 3 h to 24 h ahead), during the test period, revealed that the two DL models generally showed similar performance, with slightly superior results with the GRU model. For instance, GRU and LSTM showed an averaged root mean square error (RMSE) of 4.96 cm and 5.3 cm and a coefficient of determination (R^2) of 0.93 and 0.92, respectively. Investigations of the time series forecasting performance at selected locations, also demonstrated the superiority of the GRU model, for all time steps, with Willmott's index (WI) values generally above 0.9 and high reliability as reflected in Prediction Interval Coverage Probability (PICP) values mostly exceeding 90 %. The results, however, weren't always perfect; both the GRU and LSTM models encountered limitations with forecasting the sea level maxima. Further examination of the spatial discrepancies also reveals some problematic areas in the eastern Gulf of Finland. This may have been influenced by the exclusion of some input components such as river discharge, salinity and meridional winds, further enhanced by complex hydrodynamics, extreme sea level variations, strong local currents, resonance-induced seiches and seasonal ice cover. In addition, an external validation of the GRU results was performed using along-track satellite altimetry from Sentinel 3A and 3B missions. For most of the satellite tracks, the discrepancy was better than 5 cm, proving the capabilities of the model generalization capabilities. These findings hold significant implications for advancing our comprehension of oceanic dynamics, enhancing maritime safety, and benefiting a wide range of applications that are dependent on accurate sea level forecasting.

* Corresponding author.
E-mail address: saeed.rajabi@taltech.ee (S. Rajabi-Kiasari).

https://doi.org/10.1016/j.apor.2025.104496
Received 17 January 2024; Received in revised form 17 January 2025; Accepted 27 February 2025
Available online 10 March 2025

1. Introduction

The effects of our changing climate show several signs, such as unexpected floods, increased intensity and frequency of extreme events, changing ocean dynamics, etc. (IPCC, 2023). As a result, accurate sea level forecasting is a vital component that allows safe and efficient planning of maritime activities but also supports present and futuristic advancements (e.g. ocean digital twins). This is especially relevant for a wide range of applications (e.g. navigation, marine engineering, coastal protection). Different approaches, such as data assimilation and machine learning/deep learning methods (ML/DL) have been applied to forecast sea levels. The main problem stems from that whilst data assimilation methods have shown reasonable forecasting capabilities, they often suffer from potential inaccuracies due to physical formulation, numerical approximations, and unresolved small-scale processes (Primo de Siqueira and Paiva, 2021). In recent years, however, ML/DL methods have been widely used due to their forecasting being offline (contrastingly to data assimilation), increased computing capabilities and their ability to identify reliable patterns, specifically for shorter-term predictions (Qin et al., 2023). This demonstrates their potential to be even more efficient (an obtainable accuracy of a few centimeters, for a time scale of a few hours to a few days/weeks) while also being less computationally demanding (Rajabi-Kiasari et al., 2023).

The challenge remains in that various ML approaches exist (see Table 1) and determining which approach is the most optimum, the best input variables to be included and also identifying the limitations, so that it can be used effectively and wisely. It is known that the effects of climate change are usually not generated independently but are instead influenced by various inter-connected components (e.g. tides, winds, sea surface temperature, etc.). The main message is that it seems both practical and important to utilize methods that consider interconnections in sea level forecasting, as this is expected to increase its capabilities (Rajabi-Kiasari et al., 2023) but also indirectly allows us to have a better understanding of the connectivity of various drivers.

It is first important to note that sea level forecasting can be performed for different timescales, including both short-term (hours to days) and long-term (weeks to years), each serving specific purposes and applications (Imani et al., 2014; Karimi et al., 2013). Short-term predictions are often applicable to maritime navigation safety, storm monitoring, and coastal management (Accarino et al., 2021). On the other hand, long-term predictions support climate studies, trend analysis, understanding of ocean circulation patterns and also coastal management. As mentioned above, the challenge remains in determining the best forecasting method for the specific purpose and knowledge of its limitations.

Various ML/DL approaches can be utilized in both univariate (i.e. it considers only the target variable) and multivariate frameworks (where several variables with respect to the target parameter are also considered). Moreover, different sources of sea level data often exist (e.g. tide gauges, satellite altimetry, hydrodynamic models), whereas each of these sources often has its limitations concerning spatial and temporal resolution and accuracy (Jahanmard et al., 2022, 2023a). There exist numerous studies on utilizing ML/DL approaches for sea level forecasting. Table 1 presents a summarized list of the most related earlier studies. A wide range of univariate studies are available using traditional ML models such as linear regression, regression tree, ensemble, Artificial Neural Networks, and Gaussian Process Regression (Hazrin et al., 2023), Adaptive Neuro-Fuzzy Inference System (Primo de Siqueira and Paiva,

Table 1
A summary of most related sea level forecasting studies used ML/DL models.

Authors	Study area	Model(s)	Input(s)	Sea level source
(Hazrin et al., 2023)	four different locations in Malaysia	MLR, SVM, regression tree, ensemble, ANN, GPR	Sea level	TG
(Altunkaynak and Kartal, 2021)	3 stations located in different regions of the Bosphorus Strait, Turkey	SVR, KNN, decision tree with discrete wavelet transforms	Sea level	TG
(Wang et al., 2020)	Two coastal tide stations in New Jersey and Louisiana, USA	ANN, Wavelet ANN, ANFIS, Wavelet-ANFIS	SLA and wind-shear velocity	TG
(Sabililah and Adytia, 2023)	Pangandaran Beach, Indonesia	Transformer, LSTM, RNN	Sea level	TG
(Song et al., 2021)	Bohai Sea, China	LSTM, GRU and SRU	SSH anomaly	Gridded SA
(Ishida et al., 2020)	Osaka gauging station in Japan	LSTM	WS, MSLP, and AT, annual global mean AT, relative positions of the moon and the sun near the target	TG
(Balogun and Adebisi, 2021)	West Peninsular, Malaysia	ARIMA, SVR, LSTM	SST, SST, density, P, WS, total cloud cover, PCP and SLA	Gridded SA
(Accarino et al., 2021)	five stations in the SANI area (Otranto, Bari, Taranto, Vieste and Crotone) in Italy	Multi-modal LSTM	Sea level	TG
(Zhou et al., 2022)	South China Sea	MLFrnn	SSH anomaly	Gridded SA
(Wang et al., 2022)	on the open sea and coastal regions of the South China Sea	HMnet	SST anomaly, WS anomaly, and SLA	Gridded SA
(Liu et al., 2022)	South China Sea	attention-based LSTM	DT	Gridded SA
(Ayinde et al., 2023)	Gulf of Guinea	MLPR, MLR, RF, GBR, and LSTM	MSLA, SST, SSS, AT, WS, PCP, EVP, P, radiation and flux	Gridded SA
(Zhang et al., 2020)	North Sea	Generative Adversarial Networks	Sea level	TG and Numerical model: Atlantic Margin Model-7 (AMM7)
(Bellinghausen et al., 2023)	Seven TG stations in the Baltic Sea	RF classification	P, WS, prefilling	TG
(Rajabi-Kiasari et al., 2023)	Baltic Sea	2d CNN	DT, WS, P, SST, day of the year	Bias-corrected Hydrodynamic model

Abbreviations are SSH: Sea Surface Height, SSS: Sea Surface Salinity, SST: Sea Surface Temperature, AT: Air Temperature, WS: Wind Speed, P: Atmospheric Pressure, MSLP: Mean Sea Level Pressure, SLA: Sea Level Anomaly, MSLA: Mean Sea Level Anomaly, DT: dynamic topography, PCP: Precipitation, EVP: Evaporation, HF: Heat Flux, LR: Linear Regression, SVM: Support Vector Machine, KNN: k-nearest Neighbor, RVM: Relevance Vector Machine, ANN: Artificial Neural Networks, GPR: Gaussian Process Regression, ANFIS: Adaptive Neuro-Fuzzy Inference System, LSTM: Long Short-Term Memory, GRU: Gated Recurrent Unit, SRU: Simple Recurrent Unit, DLNN: Deep Neural Networks, ARIMA: Autoregressive Integrated Moving Average, SVR: Support Vector Regression, MLFrnn: Multilayer Fusion Recurrent Neural Network, MLP: Multi-Layer Perceptron, CNN: Convolutional Neural Network, MLPR: Multilayer Perceptron Regression, MLR: Multiple Linear Regression, RF: Random Forest, GBR: Gradient Boosting Regressor, HMnet: Hybrid Model network, TG: tide gauge.

2021), Support Vector Machine and k-nearest Neighbor (Altunkaynak and Kartal, 2021) and Relevance Vector Machine (RVM) (Fu et al., 2019; Imani et al., 2018).

The literature review in Table 1 yields two conclusions. First, multivariate forecasting methods generally outperform univariate models. This is because incorporating meteorological factors (such as wind speed, sea level pressure, and steric components like sea surface temperature and salinity) enhances the accuracy and performance of the models (Balogun and Adebisi, 2021). It is also worth noting that in some regions, the sea level changes can also be influenced by other factors such as river runoff and climatic effects (Tsimplis and Woodworth, 1994). Other studies have also combined empirical mode decompositions with ML models to enhance the performance of sea level prediction accuracy (Fu et al., 2019; Jin et al., 2023; Raj et al., 2022; Song et al., 2022; Zhao et al., 2021). Secondly, recent advancements in deep neural networks, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN-RNN models have demonstrated superior performance in oceanography, especially for spatio-temporal forecasting (Sun et al., 2022). These models leverage automatic feature extraction and memory-storing capabilities to produce more accurate and higher-resolution spatio-temporal maps, making them well-suited for forecasting dynamic marine parameters such as sea levels (Braakmann-Folgmann et al., 2017; Liu et al., 2022; Song et al., 2021; Wang et al., 2022; Zhang et al., 2020; Zhou et al., 2022), surface temperature (Xiao et al., 2019), wind speed (Gao et al., 2023), and significant wave height (Zilong et al., 2022).

Among the various DL approaches, Recurrent Neural Networks (RNNs)—specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures—are particularly promising for sea level forecasting (see Section 2.2 for more details). These models excel in processing sequential data, allowing them to capture long-term dependencies and learn from temporal patterns effectively (Accarino et al., 2021; Ishida et al., 2020; Liu et al., 2022; Sabililah and Adytia, 2023). Recurrent DL models, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures offer distinct advantages over other models in their processing of sequential (e.g. temporal) data. One notable advantage is their ability to capture and retain long-term dependencies within sequences, making them well-suited for time-series forecasting. Moreover, LSTMs and GRUs provide a solution to the vanishing gradient problem encountered in standard RNNs, allowing for more effective learning of temporal patterns (see Section 2.2 for more details). The key difference between these two methods is in their internal architecture, for LSTMs use a more complex memory cell structure whilst GRUs employ a simpler architecture, striking a balance between computational efficiency and model expressiveness. As a result, this study examines the performance of these two popular methods (GRU and LSTM) for sea level forecasting.

The study area for this examination is the Baltic Sea, which is surrounded by nine countries and is one of the busiest shipping routes in the world (Baran and Neumann, 2023). In the Baltic Sea, the short-term sea-level variations are mainly driven by atmospheric forcings, especially that of the winds which have anisotropic characteristics (Soomere, 2003). Thus, making sea level forecasting more challenging, but very important for practical operations. For the Baltic Sea, several studies have explored ML/DL methods using tide gauges (Bellinghausen et al., 2023; Tiggeloven et al., 2021). However, Rajabi-Kiasari et al. (2023) performed the first case study of utilizing hydrodynamic model data to forecast one-day-ahead sea level by employing a two-dimensional CNN model with a spatial resolution of $1/4^{\circ} \times 1/4^{\circ}$. In their study, several inputs were considered such as the sea level from a corrected hydrodynamic model (referred to Dynamic Topography), wind speed and direction, surface air pressure, sea surface temperature and day of the year. The results were promising yielding RMSE within 4 cm. There were, however, some challenges that emerged in the south-eastern sections of the Baltic Sea with larger RMSE values (~10 cm), and this hinted to be associated with the source of inputs utilized. Given the rapid short-term fluctuations in sea levels that exist in the Baltic Sea and the necessity for higher spatial and temporal sea level forecasting required for the marine operations model (Kowalewski and Kowalewska-Kalkowska, 2017), this study now explores such improvements through the utilization of higher spatial and temporal resolution input data with the two DL methods, GRU and LSTM. This was necessary for our study region where sea level dynamics can vary substantially over short distances due to factors such as the small internal Rossby radius (e.g. 2–4 km in the eastern Baltic Sea) and the rapid changes in local storm conditions (Soomere et al., 2008; Weisse et al., 2021). Studies suggest that a resolution of 0.5–1 nautical mile (1/3–1/2 of the Rossby radius) is better for capturing these dynamic patterns (Soomere et al., 2008; Kowalewski and Kowalewska-Kalkowska, 2017). Such a scheme is expected to produce better sea level forecasting results.

It should be also noted that accurate sea level forecasting does not only depend on the ML/DL algorithms utilized but also on the quality of input data. Most of the previous studies have predominantly focused on utilizing different sources of sea level data such as tide gauges (TG) and gridded satellite altimetry (SA) products (cf. Table 1). These sources, however, have their own limitations. For example, TGs, commonly used for sea level forecasting, are land-based and typically measure relative sea levels (i.e., with respect to the land unless corrected for vertical land displacement), focusing only on the sea level near their location (De Biasio et al., 2020). This leaves offshore areas underexamined, despite their importance for navigation, marine engineering, and other critical activities. Daily averaged gridded SA data products, considered as absolute sea level, for which, the reference datum is an ellipsoid (mathematical representation of the shape of the earth) are also utilized in the literature for spatiotemporal forecasting. However, depending on the application, they lack the necessary validations, for parameters such as sea surface height (SSH) and sea level anomaly (SLA) often utilized. Additionally, such data sets are often limited from providing high spatio-temporal resolutions.

On the other hand, hydrodynamic models have been scarcely utilized for sea level forecasting using ML/DL. They, however, can provide better spatial and temporal resolutions than that of SA and TG. The limitation of usage is that their vertical reference datum may be undisclosed (Jahanmard et al., 2021; Slobbe et al., 2014), thus a vertical bias that varies spatially and temporally exists when compared to other sea level sources (e.g. TG and SA). Using the geoid as the vertical reference for all sea level data ensures accuracy and consistency from coastal to offshore, which is particularly important for the Baltic Sea region, where nine countries with differing vertical references exist. To address this challenge, recent studies in the Baltic Sea region have employed a dense network of geoid (equipotential surface of the earth) referred TG data and interpolation techniques (Jahanmard et al., 2021, 2022; Varbla et al., 2022). These methods quantified the vertical bias that existed in the HDM and then a corrected HDM model was produced. Their results have been successfully verified by along-track satellite altimetry observations. This approach demonstrated improvements in HDM on average from 6.5 to 4.1 cm across the Baltic Sea region. This improved HDM dynamic topography product now refers to the geoid and more realistically represents sea level variations (see Section 2 for more details). As a result, this study shall now utilize a bias-corrected HDM, whereby the vertical datum refers to the geoid.

It is also vital to validate ML models to ensure an effective DL model. This is commonly performed using different cross-validation techniques (Imani et al., 2014; Rajabi-Kiasari et al., 2023), which consider the internal validation (using a subset of the original data). Instead, a more rigorous approach is by external validation using an entirely independent dataset (Vuolio et al., 2020). This external validation is critical for evaluating the model's practical applicability, helping to mitigate overfitting risks, and potentially ensuring a more unbiased assessment. This study uses along-track satellite altimetry observations to further validate the DL results.

Hence, the objective of this research is to develop a high-resolution

spatio-temporal framework for short-term sea level forecasting utilizing two recurrent neural network methods (LSTM and GRU). This research offers several key contributions: (i) utilization of a HDM corrected for spatial and temporal biases by using geoid-referred tide gauges, thus providing a realistic and compatible representation of sea level data across both coastal and offshore regions; (ii) implementation of a high spatial (1 nautical mile) and temporal (hourly) resolution scheme, that so far is unprecedented in ML/DL-based sea level forecasting studies, addressing challenges posed by small-scale phenomena like the small Rossby radius and rapid changes caused by local storms; (iii) to enhance the predictive capabilities, a multivariate approach using HDM data (winds, sea surface temperature, pressure, sea levels) combined with LSTM and GRU models, improving forecasting accuracy up to 24 h ahead; (iv) perform a careful seasonal EDA (exploratory data analysis) to find the most related meteorological parameters; and (v) external validation of the forecasted results with along-track satellite altimetry data from Sentinel 3A and 3B missions. The DL methods were examined in the eastern Baltic Sea for the period 2017 to 2019 (85 % for train and 15 % for test). To the best of the authors' knowledge, this is the first study toward spatio-temporal multistep-ahead sea level forecasting using hydrodynamic models within the Baltic Sea region.

The structure of the paper is as follows: Section 2 describes the spatio-temporal DT forecasting problem and the background structure of proposed models. Section 3 shows the steps used in preparing the input data and description of the study area. Section 4 explains the used architecture for the two recurrent neural network methods along with feature selection, model training process, and hyperparameter selection. Section 5 presents the results of this study. Implications of our research results are discussed in Section 6. In Section 7, the conclusions and future work suggestions are provided.

2. Background and methodology

2.1. Spatio-temporal dynamic topography forecasting problem

Dynamic Topography (DT) is an absolute sea level measurement that refers to a geoid surface (which represents an equipotential surface of the earth). DT can be captured from different sources including satellites, hydrodynamic models, Global Navigation Satellite System (GNSS), etc. The formulation to extract *DT* from satellite sea surface height (*SSH*) observations can be defined as

$$DT_{(\phi_s,\lambda_s,\ t)} = SSH_{(\phi_s,\lambda_s,\ t)} - N_{(\phi_s,\lambda_s)} \tag{1}$$

where *N* denotes the geoid height at different spatial grid points (ϕ_s, λ_s).

Note, the geoid is the equipotential surface that closely approximates mean sea level, considering the irregular distribution of Earth's mass and the gravitational forces acting upon it. It can be determined via in-situ (ship and/or airborne) gravity and remote sensing measurements (using satellite-derived gravity data). As a first-order approximation, the geoid can be assumed stationary (it does not change with time) (Morrow et al., 2023).

As mentioned earlier, variations in DT/SSH can be influenced by various factors. In this study, the inclusion of relevant input variables (based on previous studies and exploratory data analysis), namely historical measurements of DT, pressure, uwind (zonal wind component), vwind (meridional wind component), SST, and SSS are utilized. The rationale behind the selection of these variables is provided in Sections 3.2 and 4.1 of the paper. Accordingly, based on the meteorological data collected from *n* grid points, we aim to utilize a multi-faceted forecasting approach, which is both multi-task and multivariate, to forecast future conditions at each of the grid points simultaneously. For instance, considering a specific grid point *s* with its geographical coordinates (ϕ_s, λ_s), we have a sequence of previous input data spanning *w* time steps $(p_{t-w},\ \ldots,\ p_t)$ up to the current time step (t) for the observable variables and the intention is to forecast future DT values at the same point. The model output is an estimate of the target DT $(\widehat{DT})$ for the following time steps after t. The mathematical expression for this forecasting task can be formulated as follows:

$$\widehat{DT}_{(\phi_s,\lambda_s,\ t+(1:\Delta))} = f\left(Pressure_{(\phi_s,\lambda_s)}, uwind_{(\phi_s,\lambda_s)},\ vwind_{(\phi_s,\lambda_s)},\ SST_{(\phi_s,\lambda_s)}, SSS_{(\phi_s,\lambda_s)},\ DT_{(\phi_s,\lambda_s)}\right)_{(t-w:t)} \tag{2}$$
$$, where\ \ s = 1 : number\ of\ grid\ points$$

Here, the parameters *w*, Δ, and *f* define the temporal lag, the lead time (also called forecast horizon) and the mapping function, respectively. Notably, (ϕ_s, λ_s) represent the geographical coordinates corresponding to each grid point. The flowchart in Fig. 1 has been proposed to better understand the methodology.

Several sequential steps are involved in the methodology (Fig. 1). The first step requires collecting all essential input data (dynamic topography, sea surface temperature, pressure, etc.). The second stage is followed by a pre-processing stage, which involves tasks such as (i) data gridding (since data sets used for inputs may be at different resolutions), and scaling (data normalization to overcome the impacts of larger data), (ii) feature selection (determining the most relevant inputs), and (iii) data structuring (to make data ready for deep learning models). The third step involves the configuration and implementation of each deep learning model (LSTM and GRU) and establishing the necessary hyperparameters (e.g. topology and units of neural networks, loss function, optimizer) (see Section 4.2 for more details). Following this, the model training is implemented for LSTM, and GRU model (details to follow). The fourth step then involves an evaluation of the models' performance on the test data using different performance metrics (e.g., RMSE, R^2, WI and PICP) (see Section 2.3). In the last step, an externally independent validation of the results is performed using along-track satellite altimetry sea level data (see Section 3.2.4).

2.2. Deep neural network models

Machine learning (ML) models, a subset of artificial intelligence, involve creating statistical models that enable computers to perform tasks without explicit programming. The basic concept is that the data set is composed of training and test data. ML algorithms gain insights from data during training, gradually improving their performance. This is followed by hypothesis maps that are used to compute predictions of a variable. The adaptation or improvement of the hypothesis is based on the disparity between the observed and predicted values, measured by a loss function. As a result, the ML procedure consists of three main components: (i) input data, (ii) the ML model that consists of its hypothesis map and (iii) the loss function that is used to measure the quality of a particular hypothesis. These three components consist of design choices by the user, based on the data and the purpose of the study (Jung, 2022). The general steps include collecting and identifying relevant data, preparing it, selecting a model, and iteratively adjusting parameters to minimize errors (i.e. ideally to obtain minimum loss).

Deep neural network models represent an advanced category within ML, that use deeper frameworks to train larger models on large datasets where they automatically learn hierarchical features from raw data and

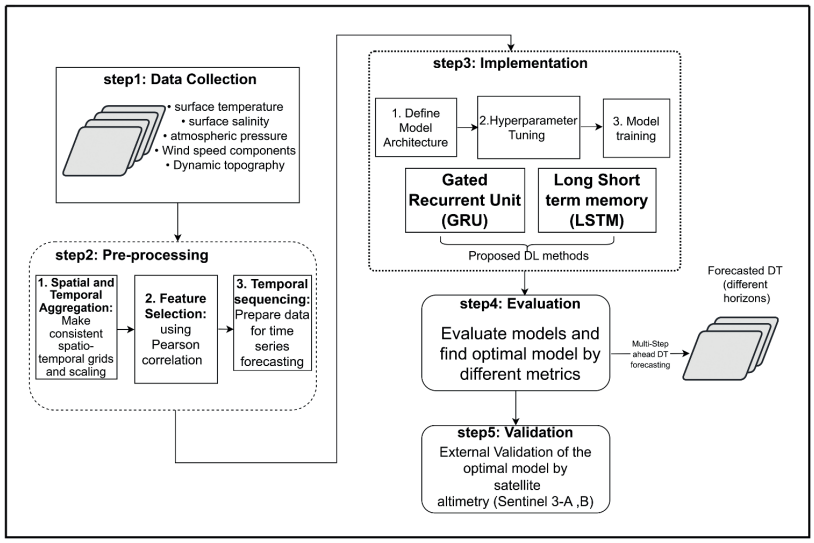


Fig. 1. Flowchart of the proposed strategy for spatio-temporal DT forecasting.

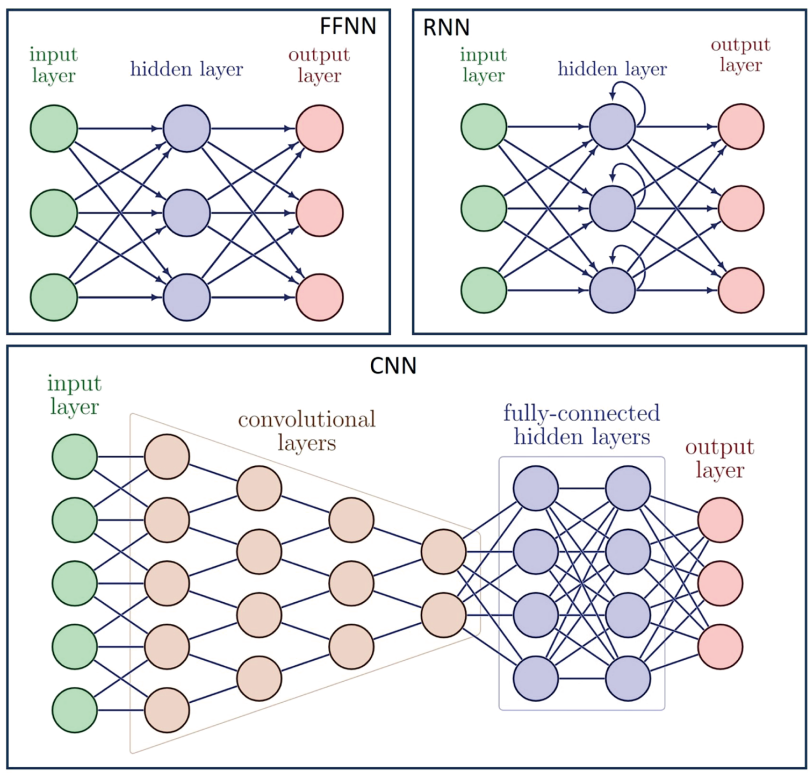


Fig. 2. Comparison between Feed-Forward neural networks (FFNNs), recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

have a higher number of parameters.

Feed Forward neural networks (FFNNs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are the three main types of DL models used in sea level forecasting problems. They, however, differ in their architectural design and applications (see Fig. 2). FFNNs are the basic structure of DL models, from input through

hidden layers to output. They are well-suited for tasks where sequential order is less relevant. CNNs also use the same structure as FFNN but they also include convolutional layers before hidden neurons that process fixed-size local receptive fields with automatic feature extraction, making them more suitable for grid-like data, e.g., images.

Recurrent neural networks (RNNs) (Elman, 1990; Rumelhart et al., 1986; Werbos, 1988) are specialized DL models for tasks involving dynamic temporal dependencies, such as natural language processing or time series forecasting (Zou et al., 2023). RNNs have found extensive applications in various forecasting tasks. Compared to other models, RNNs are favored for their memory to handle time series tasks. Another distinguishing characteristic of RNNs is that they share parameters across each layer of the network, allowing the model to capture temporal patterns and dependencies. Compared to FFNNs which often assume inputs and outputs are independent of each other, RNNs consider one element at a time, allowing them to maintain memory through recurrent connections (see Fig. 2).

RNNs, however, come with several drawbacks, specifically the challenge of exploding and vanishing gradients (Bengio et al., 1993). During the training, RNNs use gradient descent to adjust their parameters (weights and biases) to minimize the prediction error. To accomplish this, the RNN uses back-propagation through time that computes the gradients from the end of the sequence to the start. However, the multiplication of gradients in back-propagation, especially in sequences with many time steps, can lead to the exponential decay of gradients (i.e. the gradients approach zero), making them particularly susceptible to vanishing/exploding gradient problems (Bengio et al., 1994). Accordingly, when gradients during backpropagation become extremely small/large, possible impacts can be: (i) hindering the algorithm from converging to an optimal solution, (ii) causing numerical instability during training, and (iii) making it challenging for the network to effectively learn long-term dependencies and adjust its parameters. These limitations lead to the development of LSTMs and GRUs (by introducing gating mechanisms), used in this research and will be explained in the following sub-sections.

2.2.1. LSTM model

The Long Short-Term Memory (LSTM) network emerges as a specialized solution and robust version of RNN models, which is particularly well-suited for the analysis of time series and sequential data, aimed at mitigating the long-term memory limitations that are inherent in conventional RNN networks (Chen et al., 2021). A key feature of the LSTM network is its utilization of internal mechanisms known as gates. These gates regulate information transfer between the current and previous states, allowing them to learn both short-term and long-term dependencies in time series data (Espinosa et al., 2021).

Compared to RNN structure that only use one hidden recurrent layer, The architecture of LSTMs includes a memory cell that helps prevent the issues of gradient explosion or vanishing (see Fig. 3). Consequently, the LSTM network effectively propagates crucial information along the network to yield the desired output. The LSTM network presented in Fig. 3 (dotted lines) comprises three primary gates: the forget gate, input gate, and output gate, in addition to a cell state, explained below.

Forget Gate: The function of the forget gate is to recognize the information to be preserved and disregarded. It combines the input data of the current time step and the hidden state of the previous time step and then subjects this composite input to the sigmoid function. The output of this function, constrained between 0 and 1, signifies whether the information should be retained (closer to 1) or forgotten (closer to 0). In the following Eqs. (3)–(6), σ and Tanh are the sigmoid and hyperbolic tangent activation functions; $\cdot$ is the elementwise multiplication product; x_t and h_{t-1} refer to the input at time t and the hidden state of the previous timestep, respectively; W_* *and* b_* show the weight and bias term associated with the gates; f_t, i_t, o_t denote the forget, input, output of LSTM gates, respectively; and c_t is the cell state (He et al., 2023).

$$f_t = \sigma\left(W_f[x_t, h_{t-1}] + b_f\right) \tag{3}$$

Input Gate: The input gate is responsible for updating the values within the cell state. It receives the input data from the current time step and the hidden state information from the previous time step. These inputs undergo processing through the sigmoid function, determining which data should be updated (close to 1) or discarded (close to 0). Additionally, the input data from the current time step, coupled with the hidden state information from the previous time step, are subjected to the Tanh function, scaling their values within the range of −1 to 1. The output of the Sigmoid and Tanh functions is then multiplied to enable the Sigmoid function to decide which values of the Tanh function's output should be retained.

$$\begin{aligned} i_t &= \sigma(W_i[x_t, h_{t-1}] + b_i) \\ \widehat{c}_t &= tanh(W_c[x_t, h_{t-1}] + b_c) \end{aligned} \tag{4}$$

Cell State: The cell state is updated through a pointwise

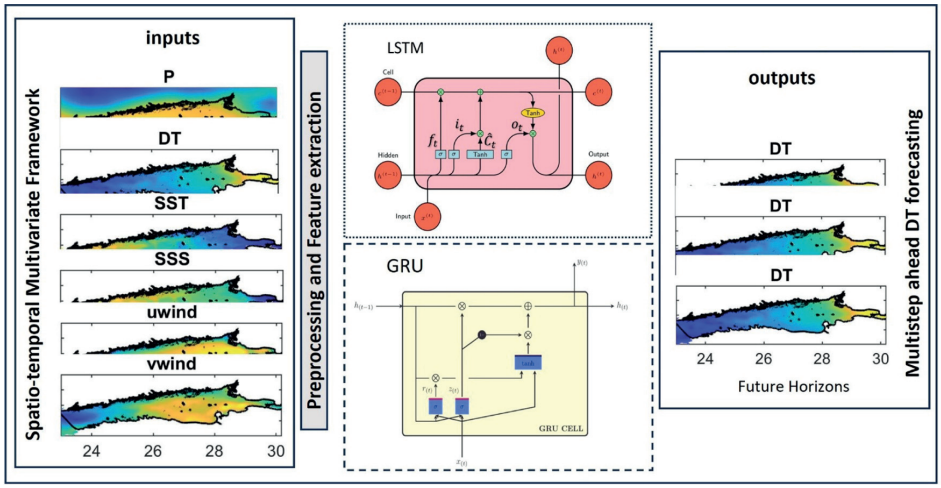


Fig. 3. Schematic architecture of GRU (dashed line), LSTM (dotted line) models and the proposed framework for spatio-temporal multivariate multistep ahead DT forecasting (solid line).

multiplication of the output from the forget and input gates. This step involves removing or discarding any values in the cell state that are multiplied by output values close to zero from the forget gate. Subsequently, the cell state is augmented with the output from the input gate through pointwise addition. Following this step, the cell state values are updated and deemed as new.

$$c_t = f_t \cdot c_{t-1} + i_t * \widehat{c}_t \tag{5}$$

Output Gate: The output gate determines the composition of the next hidden state for the subsequent time step. The hidden state retains information from prior inputs. The input data from the current time step and the hidden state information from the previous time step are fed into a sigmoid function and a Tanh function, respectively. The resulting outputs from these functions are multiplied to determine which information the hidden state carries forward to the next time step. Consequently, the newly updated cell state and the updated hidden state are passed to the subsequent time step in the sequence.

$$\begin{aligned} o_t &= \sigma(W_o[x_t, h_{t-1}] + b_o) \\ h_t &= o_t \cdot tanh(c_t) \end{aligned} \tag{6}$$

2.2.2. Gated recurrent unit (GRU)

GRU model was introduced by Cho et al. (2014), and it also solves the gradient vanishing problems of RNNs using a simplified gate mechanism compared to the LSTMs. Their internal structure closely resembles that of the LSTM layer, combining the input gate and forget gate into a single update gate. GRU uses only the hidden state for information transfer (see Fig. 3), resulting in a simpler model structure (Bai and Tahmasebi, 2023), with fewer parameters (Mao et al., 2023), and mitigating the risk of overfitting, yet with comparable performance (Sarveswararao et al., 2023). This makes GRU a preferred choice in scenarios with limited computational resources (Ahmed et al., 2023).

As depicted in Fig. 3 (dashed lines), the GRU deep learning model is similar to the LSTM network with reset and update gates, in contrast to the three gates in LSTM. In the following, we discuss how these two gates work.

Update Gate:

The Update Gate mirrors the LSTM's Forget and Input Gates, deciding how much of the past information needs to be incorporated into the network. It multiplies the new input (x_t) by the previous hidden state (h_{t-1}) using weight values adaptively adjusted during training through iterative processes; applies a sigmoid function for an output between 0 and 1; and adjusts these weights iteratively during training. The output of the Update Gate is then combined with the previous hidden state via element-wise multiplication.

Reset Gate:

The Reset Gate determines the extent to which past information should be forgotten. It combines the new input (x_t) and previous hidden state (h_{t-1}) with different weights; passes them through a sigmoid function; and outputs values between 0 and 1. The Reset Gate's output is then dot-multiplied with the hidden state from the previous step. Additionally, the new input is initially weighted, added to the reset gate's output, and processed through a Tanh function.

In the final step, the Update Gate output is element-wise multiplied by the output of the Tanh function, aimed to preserve pertinent new information in the new hidden state. This output is added to the dot product of the Update Gate output and the previous hidden state, leading to the creation of the new hidden state. The related equation for different gates is defined as follows (Sun et al., 2023)):

$$z_t = \sigma(W_u[x_t, h_{t-1}] + b_t) \tag{7}$$

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r) \tag{8}$$

$$\begin{aligned} \widetilde{h}_t &= tanh(W_h[r_t * x_t, h_{t-1}] + b_h) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t \end{aligned} \tag{9}$$

In the above Eqs. (7–9), the terms z_t *and* r_t represent the outputs of the update and reset gates, respectively. W_u and W_r are the weight matrices associated with these gates at time step t. x_t corresponds to the input at the current time step t, while h_{t-1} denotes the previous hidden state at time step $t-1$. The notation b_* signifies the bias term, and σ refers to the sigmoid activation function. Finally, h_t stands for the updated hidden state at time step t and $\widetilde{h}_t$ represents the candidate memory content at the same time step.

2.3. Statistical metrics for evaluation of the models

To assess the predictive performance of the proposed models, we split the dataset into training and test sets. The training set is used for estimating model weights and the test set, which remains unseen during training, serves as the ultimate benchmark for model performance evaluation. Evaluation metrics include:

Root Mean Squared Error (RMSE): RMSE provides a measure of the typical prediction error, with higher weight given to larger errors. The formulation is defined as below:

$$\mathrm{RMSE}_{(\phi_s,\lambda_s)} = \sqrt{\frac{\sum_{t=1}^{n}\left(\widehat{DT}_{(\phi_s,\lambda_s,t)} - DT_{(\phi_s,\lambda_s,t)}\right)^2}{m}} \tag{10}$$

Coefficient of Determination (R-squared): R-squared quantifies the proportion of variance in the sea level data that are captured by the model predictions. The formulation is defined as below:

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{s=1}^{n}\sum_{t=1}^{m}\left(\widehat{DT}_{(\phi_s,\lambda_s,t)} - DT_{(\phi_s,\lambda_s,t)}\right)^2}{\sum_{s=1}^{n}\sum_{t=1}^{m}\left(\widehat{DT}_{(\phi_s,\lambda_s,t)} - \overline{DT}_{(\phi_s,\lambda_s)}\right)^2}, \\ \overline{DT} &= \frac{1}{m}\sum_{t=1}^{m} DT_{(\phi_s,\lambda_s,t)} \end{aligned} \tag{11}$$

Willmott's Index (WI): WI is a statistical measure used to evaluate the accuracy of predictive models. It compares the observed and predicted values, considering both the mean and the distribution of errors. WI ranges from 0 to 1, where 1 indicates perfect agreement between the predicted and actual values, and 0 means no agreement. The formula can be defined as:

$$WI = 1 - \frac{\sum_{t=0}^{m-1}\left(\widehat{DT}_{(\phi_s,\lambda_s,t)} - DT_{(\phi_s,\lambda_s,t)}\right)^2}{\sum_{t=0}^{m-1}\left(\left|\widehat{DT}_{(\phi_s,\lambda_s,t)} - \overline{DT}_{(\phi_s,\lambda_s)}\right| + \left|DT_{(\phi_s,\lambda_s,t)} - \overline{DT}_{(\phi_s,\lambda_s)}\right|\right)^2} \tag{12}$$

In Eqs. (10)–(12), t pertains to the time series length, extending from 1 to m, while s identifies the gridpoint among n total gridpoints. m specifies the number of training and test days at specific locations identified as (ϕ_s, λ_s). $DT_{(\phi_s,\lambda_s,t)}$ represents the observed DT at each time and location, and $\overline{DT}$ is the average of these values. Furthermore, $\widehat{DT}_{(\phi_s,\lambda_s,t)}$ describes the predicted DT over the time span t by the applied model. In addition to these quantitative statistics, we also visually inspect the model's predictions against ground truth satellite altimetry data, evaluating the model performance for generalization.

Prediction Interval Coverage Probability (PICP): PICP is a metric used here to assess the reliability of the model's uncertainty estimates. PICP is determined by the ratio of observed values that lie within the prediction interval. Typically, a higher PICP value (closer to $(1-\alpha)\%$) indicates more reliable uncertainty estimates (Yu et al., 2024). The formula for the PICP is given by:

$$PICP_{(\phi_s,\lambda_s)} = \frac{1}{m}\sum_{t=1}^{m} I\left(DT_{(\phi_s,\lambda_s,t)} \in \left[\widehat{DT}^{lower}_{(\phi_s,\lambda_s,t)}, \widehat{DT}^{upper}_{(\phi_s,\lambda_s,t)}\right]\right) \tag{13}$$

where, m is the total number of observations, and $\mathrm{DT}_{(\phi_s,\lambda_s,t)}$ is the observed value for the t-th observation. $\widehat{\mathrm{DT}}^{\mathrm{lower}}_{(\phi_s,\lambda_s,t)}$ and $\widehat{\mathrm{DT}}^{\mathrm{upper}}_{(\phi_s,\lambda_s,t)}$ are the lower and upper bounds of the prediction interval for the t-th observation. I is the indicator function, which equals 1 if the observed value

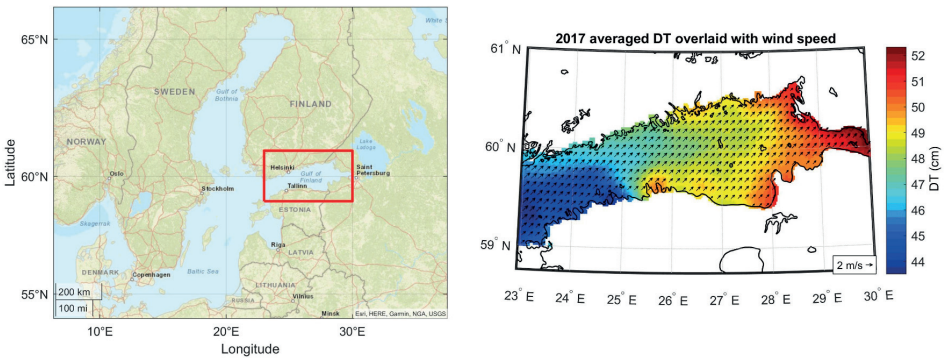


Fig. 4. Geographical location of the study area: Gulf of Finland, the eastern Baltic Sea (red rectangle) along with plotted 2017 averaged DT overlaid with annual average wind directions.

$DT_{(\phi_s,\lambda_s,t)}$ lies within the prediction interval, and 0 otherwise.

3. Case study

3.1. Study area

The Baltic Sea (BS) is one of the largest semi-enclosed estuarine sea areas on the Earth. Located in northern Europe and surrounded by nine countries (see Fig. 4), this marine area is very popular with intensified industrial and recreational activities (e.g. shipping, ports, beaches, etc.). Nevertheless, this also positions the BS as a sensitive marine environment subject to extreme anthropogenic pressure. The BS has an average depth of 54 m, and its coastline is quite diverse over 8000 km length (e.g. consisted of cliffs and sandy or stony beaches) and a coastal population of around 15 million within 10 km of the coast. The BS estuarine conditions are due to its connection to the Atlantic Ocean via the Danish Straits (which is a narrow and shallow channel) from which, salt water is sourced. The freshwater in the BS primarily originates from the numerous rivers in the surrounding lands. Due to this estuarine environment, a stratified multi-layer structure (both horizontally and vertically) often exists (Leppäranta and Myrberg, 2009). This often leads to a decreasing salinity/density from west to east and south to north. Such characteristic permanently affects the sea level with higher sea level values in the north and east than in the south and west. A change of density of 1kg/m^3 could also lead to a change in sea level of 5 cm. So, changes in temperature or salinity are expected to affect the sea level in annual and semi-annual time frames.

Tides are also relatively small in the BS with only a few centimeters ranges. Due to this estuarine environment, several oceanographic processes often take on a seasonal pattern and are quite prevalent. For example, in the summer and autumn months, coastal upwellings are common in the Baltic Sea, and this often brings cooler nutrient-laden bottom waters to the surface (Delpeche-Ellmann et al., 2017). For the winter months, some sections, especially in the northern and eastern parts tend to be ice-covered and this also affects the waves and coastal processes (Leppäranta and Myrberg, 2009). The BS is divided into several sub-basins based on bathymetry and geomorphology. These sub-basins include Bothnian Bay, Bothnian Sea, Baltic Proper, Gulf of Finland, Gulf of Riga, Bornholm Basin and Arkona Basin, each with its own unique characteristics and factors influencing sea level measurements.

Several factors influence the sea level based on different time scales. For instance, with respect to long-term effects, the global sea level change (due to thermal sea-water expansion and melting of glaciers) will influence the BS sea level. Whilst for shorter time scales, the sea level can be influenced by several main factors such as (i) the variations in the water balance caused by water exchange in the Danish Straits, (ii) river discharge, (iii) sea ice (iv) changes in density of seawater and (v) the atmospheric factors that include air pressure and winds that are known as the primary cause for short-term (weeks, days, hours) sea-level fluctuations in the Baltic Sea (Medvedev et al., 2021). The winds are known to be anisotropic in the Baltic Sea (Soomere, 2003), thus making it challenging for forecasting. Nevertheless, the prevalent wind direction is usually south-west, however, northerly winds are also common (Alenius et al., 1998; Keevallik and Soomere, 2003). It should be noted that the influence of winds can also trigger some marine processes to occur that may influence its density, water balance, etc.

This study explores the eastern section of the Baltic Sea, known as the Gulf of Finland (hereafter referred to gulf, Fig. 4). The gulf is characterized as narrow (with width varying from 48−135 km) and elongated (with a length of approximately 400 km). The mean water depth is around 37 m (maximum depth is 123 m). The source of water exchange in the gulf is different from the Baltic Sea for it is located at the extreme eastern section, thus the furthest from the North Sea. Hence, the water exchange mostly occurs with the adjacent basin (i.e. the Baltic Proper) through an interplay of estuarine and wind-driven processes and also the riverine input from one of the largest river sources such as the Neva River.

It is also important to emphasize that the primary forces responsible for sea level variations in this area are the winds and the depression of atmospheric pressure (Alenius et al., 1998; Rosentau et al., 2021). Since air pressure change of 1 mbar corresponds to 1 cm in sea level elevation, air pressure can contribute as much as ±50 cm and is expected to influence the study area (Leppäranta and Myrberg, 2009; Pellikka et al., 2018). Also, the western part of the gulf tends to be mainly influenced by winds prevailing in the Baltic Proper and the eastern part is surrounded by the mainland. Thus, different wind regimes and ocean dynamics are expected from west to east of the gulf.

The dynamics of surface layer movements within the gulf however appear to be significantly influenced by wind speed (Delpeche-Ellmann et al., 2016, 2021; Skriptunov and Gorelits, 2001). Also, strong westerly winds are known to increase water levels, especially in the eastern section, where they create a tilt in the water surface (Pindsoo and Soomere, 2020). Tidal oscillations of the sea level are of minor importance in the dynamics of the Gulf (Pellikka et al., 2018). Some studies have suggested that a change in the wind regime (wind direction and storminess) may influence the mean sea level and extremes found in the easternmost sections of the gulf and that extreme events such as storm surges can be more prominent in the coastal areas e.g. St. Petersburg and Pärnu Bay areas and the low-lying areas such as the western Estonian bays (Suursaar and Sooäär, 2007).

3.2. Data collection and preprocessing

3.2.1. Target dynamic topography: corrected HDM model

The dependent variable (DT) utilized in this research was sourced from a bias-corrected Nemo Nordic Hydrodynamic Data Model (HDM) within the context of the Baltic Sea region (Jahanmard et al., 2021, 2022). The original Nemo-Nordic (hereinafter Nemo-HDM) is a three-dimensional coupled ocean-sea ice model of the Baltic and North Sea (Hordoir et al., 2019), developed by the Swedish Meteorological and Hydrological Institute (SMHI), which simulated sea level data at hourly temporal resolution and a horizontal resolution of one nautical mile. The bias-corrected model used here was constructed through an extensive network of geoid-referred tide gauges. It is noteworthy that this corrected model exhibited a notable reduction in RMSE when compared with the original Nemo-HDM Sea level dataset. Validation was performed using tide gauges thus considering the most realistic ground truth, and satellite altimetry data, that allows coverage for both coastal and offshore regions (e.g., on average RMSE of 4 cm and correlation of 0.98 with tide gauges). The data utilized in this study was made available at a spatial and temporal resolution identical to that of the Nemo-HDM model from '01.01.2017' to '31.12.2019', which has a hourly temporal and at one nautical mile spatial resolution. DT has been referred to the European Vertical Reference System (EVRS), to which the tide gauges and geoid model are also referenced. A comprehensive description of the algorithm employed to correct the Nemo-HDM bias and generate the dataset can be found in (Jahanmard et al., 2021, 2022).

3.2.2. Meteorological features

The meteorological input data for this study was also sourced from the same Nemo HDM model (Table 2). Variables such as wind speed (in both east-west and north-south directions), surface temperature, and surface salinity were collected for the same period as DT (2017–2019). Hence, both of spatial and temporal resolutions of these data sets were similar to the DT.

Atmospheric air pressure that was reduced to sea surface was downloaded from ERA5 (https://cds.climate.copernicus.eu/) with hourly, 1/4°×1/4° resolutions. Thus, to make it consistent with the rest of our data, a bilinear interpolation method was applied.

3.2.3. Geoid model

This study employs the NKG2015 geoid model (Ågren et al., 2016) to convert the satellite-derived SSH into the DT values. This geoid model is tailored to the Baltic region countries and offers a stable reference framework, ensuring consistency and precision in our analyses. For the whole Nordic-Baltic region, the NKG2015 geoid model agrees very well with GNSS/levelling control points. It showed an improvement with a standard deviation of 3 cm after a 1-parameter fit and 1–2 cm for regions with smooth gravity field when compared with other geoid models (NKG2004, EGM2008, EGG08, and EIGEN6C4). Since the NKG 2015 geoid model used here is originally in zero tide system, it is needed to convert it to the mean tide system (that is conventionally used for the satellite altimetry data products) using the equation (Varbla et al., 2022):

$$N_m = N_z + (0.29541\left(\sin\varphi^2 - \sin{\varphi_{NAP}}^2\right) + 0.00042\left(\sin\varphi^4 - \sin{\varphi_{NAP}}^4\right) \quad (14)$$

where φ_{NAP} is the latitude of NAP.

3.2.4. Satellite altimetry sea surface height

For further validation of forecasted sea levels, an external validation was performed using 20 Hz along-track satellite altimetry data from both Sentinel 3A and Sentinel 3B missions. This data was sourced from the EUMETSAT website (https://www.eumetsat.int/). Eq. (1) was used to extract DT from the satellite SSH measurements. For obtaining accurate altimetry data, some atmospheric and geophysical corrections are also needed, introduced as follows:

$$SSH = H_{orbit} - (R + iono + DTC + WTC + SSB + SET + PT) \quad (15)$$

In the above equation, R and H_{orbit} denote the range and altitude, while *iono* (ionospheric correction), *DTC* (dry tropospheric correction), *WTC* (wet tropospheric correction), *SSB* (sea state bias), *SET* (solid earth tide), *PT* (pole tide) are the geophysical correction terms, and *SSH* is the resulting sea surface height with respect to the ellipsoid (e.g., GRS 80). The along-track SA measurements and DT values do not have similar spatial and temporal resolution. SA data has a spatial resolution of 300 m, and the satellite passes the same track every 27 days (see Table 2). So, to ensure compatibility between the sources, and minimize the delay with hourly model forecasts, tracks and cycles were carefully selected. To mitigate known coastal errors, SA measurements within 5 km of the coast were excluded, addressing potential land contamination issues (Mostafavi et al., 2021). Moreover, outliers in the SA measurements were identified and removed using the Median Absolute Deviation (MAD) method.. For further details about the SA data used, please consult https://www.eumetsat.int/.

3.2.5. Data preprocessing

Data preprocessing involves several steps to ensure data quality control and suitability. These steps include data validation, gridding and scaling, feature engineering and data structuring (see Fig. 1). These steps are required for training and evaluation of the model. The dataset comprises multiple historical records in the gulf, obtained from various sources. To prepare the data for model training, we perform a series of preprocessing steps as follows.

Spatial and Temporal Aggregation: The high-resolution data are aggregated into consistent spatio-temporal grids (for pressure data, we used bilinear interpolation).

Feature Scaling: Normalization is applied to ensure that all input features have similar scales, preventing numerical issues during training (Rajabi-Kiasari and Hasanlou, 2020). Here, we used a Standard scalar function of python sklearn package (Pedregosa et al., 2011).

Data Structuring: To prepare data for training our proposed models (according to the Eq. (2)), we structure them into series with fixed time steps. The input shape consists of five dimensions: (samples/batch, lag window, latitude grid, longitude grid, features). However, to make it compatible with the applied recurrent models, we used an input shape of (samples/batch, lag window, latitude grid × longitude grid × features).

Table 2
Description of different variables used in this research.

Variable	Spatial resolution	Temporal resolution	Source
Wind speed (u and v)	1 nm	Hourly	Nemo Nordic
Surface Pressure	0.25°×0.25°	Hourly	Era5
Sea Surface temperature	1 nm	Hourly	Nemo Nordic
Sea Surface Salinity	1 nm	Hourly	Nemo Nordic
Dynamic Topography	1 nm	Hourly	Corrected Nemo Nordic (Jahanmard et al., 2023b)
Sea Surface Height	300 m	27 days revisiting time, 20 Hz data at each pass	Along-track Sentinel 3A and 3B (EUMETSAT)

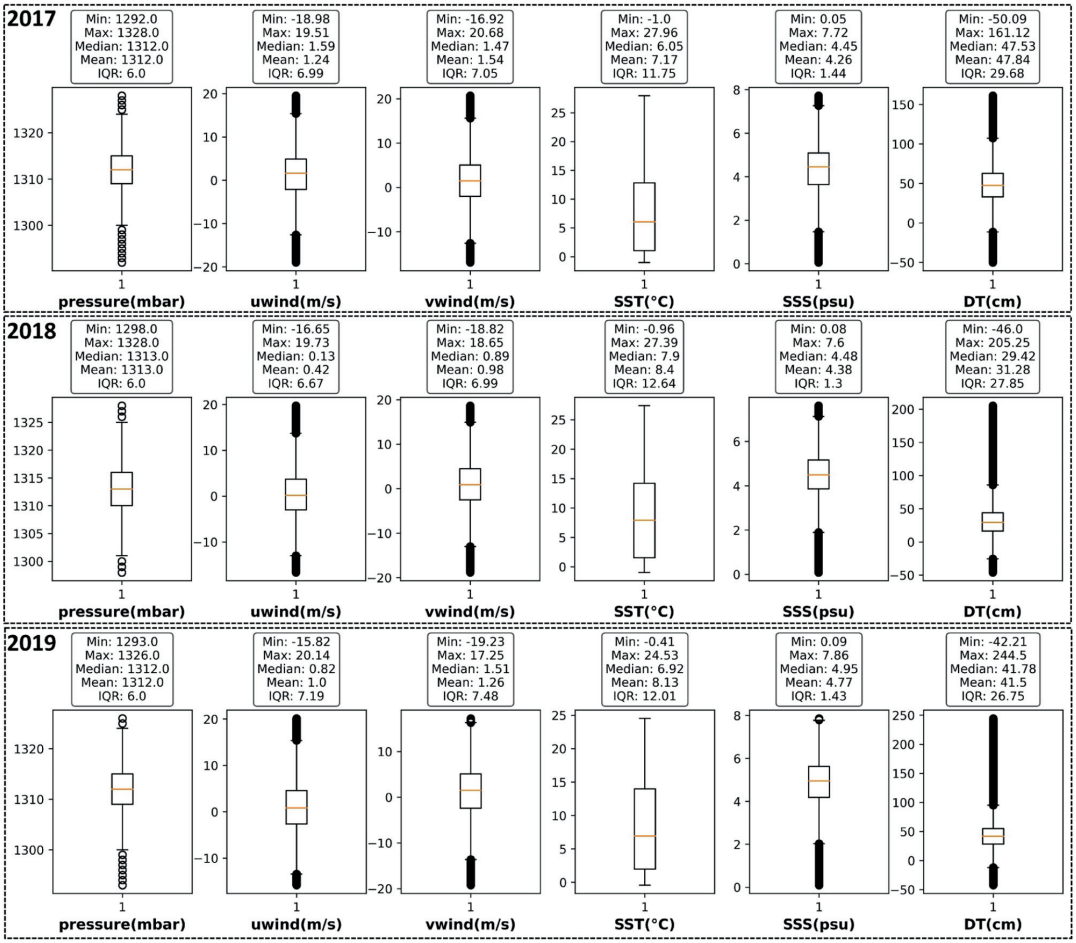


Fig. 5. Boxplot of statistics for input components over 2017, 2018 and 2019 in the Gulf of Finland.

4. Model configuration

4.1. Choosing influential input variables

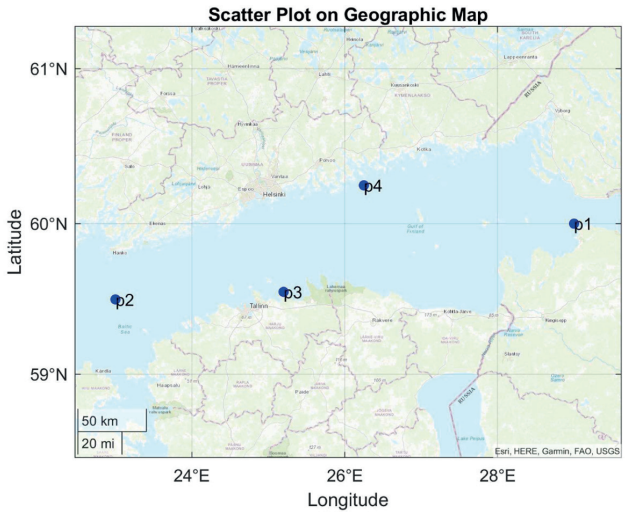
DL models are renowned for their ability to automatically extract influential features and patterns from raw data, making them suitable for complex tasks such as time series analysis when compared to ML models (Hazrin et al., 2023). Despite these automatic feature extraction capabilities of DL models, it is first essential to understand and visualize the temporal patterns within the various data sets for the gulf. This approach aids in discovering the connections between features and assists in knowing how these relationships vary over time, specifically on a seasonal scale. As a result, to explore each input component used in the DL model and how they vary seasonally and annually, statistical boxplots were depicted in Fig. 5. Each feature's boxplot graphically illustrates the mean, median, interquartile range (IQR), and extremes.

Furthermore, to explore the correlation between different input components, a Pearson correlation analysis was performed. Pearson correlation serves as a robust tool for quantifying linear associations among variables by finding the correlations between the components and also for dimensionality reduction. This also reduces the overfitting risk by preventing the model from becoming excessively complex (Hawkings et al., 2004). As a result, Pearson correlation coefficients were calculated for all input components across distinct seasons (spring, summer, winter, autumn) in 2017. Fig. 6 shows the results of the Pearson correlation analysis at four selected locations (located in the eastern to western Gulf (P1 to P4)).

The lower the atmospheric pressure, the higher the sea level (known as inverse barometer effect). For the three years examined (2017−2019) in the box plots (Fig. 5), the minimum and maximum ranged from 1293 to 1328 millibar within the gulf.

For the Pearson correlation, pressure consistently demonstrates a strong negative correlation with dynamic topography (DT) in all seasons and locations (mostly around −0.5, see Fig. 6) except in summer, when the correlation between pressure and DT is around −0.3. This suggests that changes in atmospheric pressure are closely related to variations in DT.

Changes in surface salinity may also cause changes in water density, affecting the sea level values. Salinity minimum and maximum values generally range around 0 to 8 psu in the gulf. Whilst annual changes in salinity are not expected to vary that much, they may vary for different seasons and depending on the geographic location. It is known that from west to eastward of gulf, there is a general decrease in salinity. Also, at the eastern end of the gulf is the Neva River, which contributes the



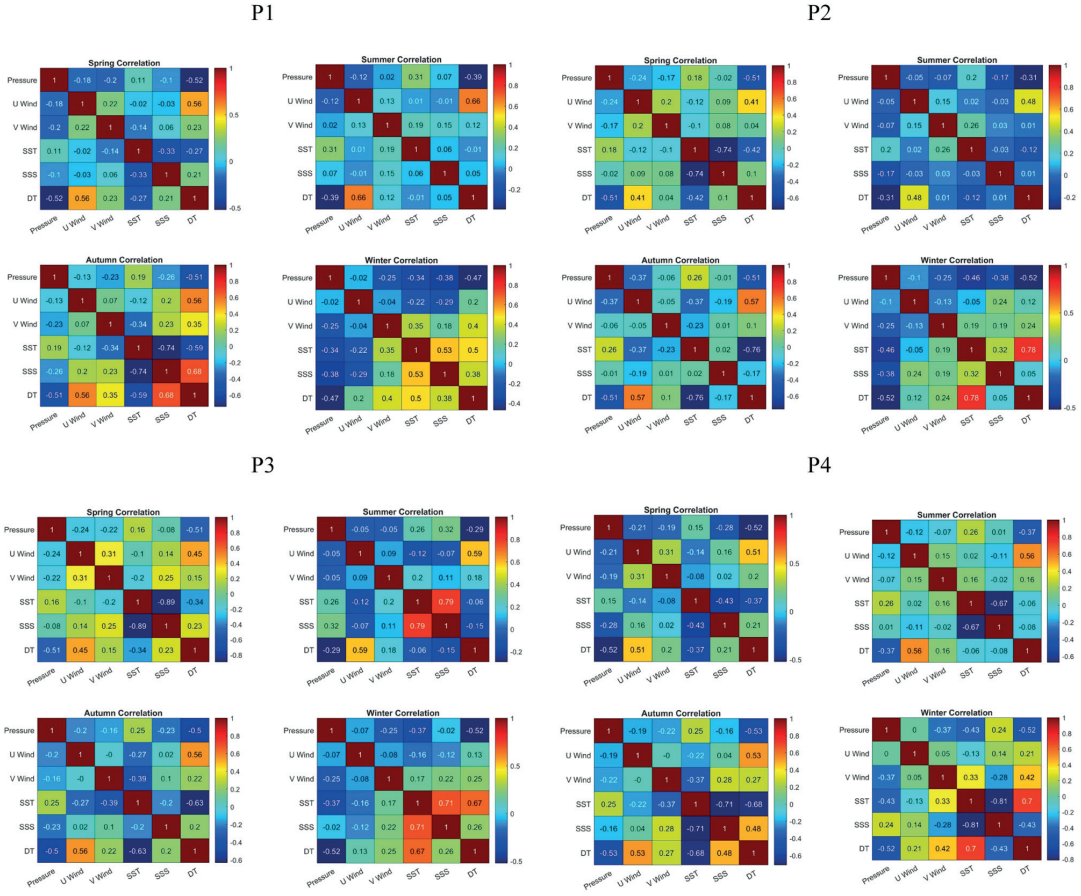


Fig. 6. Seasonal Pearson correlation results were computed at four randomly selected grid points (P1 (60 °N, 29 °E), P2 (59.5 °N, 23 °E), P3 (59.55 °N, 25.2 °E) and P4 (60.25 °N, 26.25 °E) in the Gulf of Finland, for the year 2017.

largest portion (40 %) of the total flow into the Baltic Sea Catchment area (Helcom, 2021).

From the Pearson correlation graph (Fig. 6), SSS varies geographically. At P1 (easternmost location), the highest correlation was observed in autumn (0.6), while the lowest occurred in summer (0.1). At P4, positive correlations occurred in spring (0.2) and autumn (0.4), whereas negative correlations occurred in summer and autumn. For P3, all seasons had positive values of 0.2 except for summer, whilst P2 showed weak correlations for all other seasons (0.01). Accordingly, these results indicate a generally weak correlation for salinity at most selected locations except for the easternmost station P1 in the spring and autumn, which may influence the DT.

Conversely, wind speed (u and v components) shows yearly and seasonal variations, which are reflected in shifts in median and mean values and a wide range of wind conditions (IQR 6.7 − 7.5 m/s). The maximum uwind varied from 19.73 to 20.14 m/s and the minimum uwind from −15.83 to −18.98 m/s. This demonstrates the west and east components of the wind regime and shows the zonal westerly winds (positive) are dominating stronger. The vwind varied from a maximum of 17.25 to 20.69 m/s and the minimum was −19.23 to −16.92 m/s, showing that both northerly and southerly winds may affect the study areas. These wide variations in wind speed and direction reflect the anisotropic wind patterns present throughout the years and the complexities the DL models may encounter in successfully forecasting possible effects on the sea level.

The relationship between DT and uwind is also marked by a consistent and strong positive correlation across all four points and throughout various seasons. In spring, all points displayed notably positive correlations, with values of 0.51 for P1, 0.57 for P2, 0.44 for P3, and 0.51 for P4. This trend continues into the summer and autumn seasons, with P1 and P4 showing strong positive correlations, respectively. For winter, there was a decrease in correlation with values of 0.12 for P2 and 0.13 for P3, whilst P1 and P4 still exhibit positive correlations of 0.20. This consistent positive correlation throughout all seasons and across various points identifies the influence of zonal (west-east) wind speeds (uwind) on DT across the locations. The vwind had less correlation than uwind with DT (values of around 0.1 to 0.2 and was also less consistent across all points and seasons). The highest correlation of vwind and DT occurred in the winter (0.42, 0.4) at P1 and P4.

Sea surface temperature (SST) exhibited seasonal variation, with maximum values within the range of 24.53 to 27.97 °C and minimum values fluctuating around −1 to 0 °C. This reflects the seasonal influence, where in some years, the SST may be warmer/colder than in others, which may influence the sea level variations.

As expected, the Pearson correlations between SST for the four points (P1, P2, P3, P4) vary across different seasons. In winter, all stations showed a positive correlation that varied from 0.5 to 0.7. For all other seasons, a negative correlation existed especially in the autumn with values of −0.76 at P2.

DT displayed wide variations with maximum values within the range of 161.12 to 244.50 cm and minimum values of −50.09 to −42.22 cm. In 2019, a maximum of 244.50 cm was observed. This is expected to be connected with the wind speed and direction that also showed a wide variation range (higher IQR).

As expected, the Pearson correlations for the four points (P1, P2, P3, P4) vary across different seasons. The highest correlation at all stations appears to be with uwind (positive correlation) and air pressure (negative correlation) for most stations at all seasons. The uwind correlation tends to be high (0.5) for most stations and seasons except for winter when the correlation was low (around 0.2). The vwind correlation for all seasons and stations was low (around 0.2) except for station P4 in winter where it increased to 0.4. The air pressure negative correlation (0.5) was consistent for most seasons but slightly lower in the summer.

Based on the above Pearson correlation analysis performed and due to the limited computer processing capabilities, all examined parameters couldn't be utilized, so a selection process was performed. The final parameters chosen as input components to be used in the DL models were pressure, uwind, SST and DT. Thus, vwind and SSS were excluded from further model processing. This, however, may affect the results at particular locations and seasons.

4.2. Model training

For the model set-up, 85 % of the data were used for training ('2017–01–01' to '2019–07–20') and the remaining 15 % were used as test data ('2019–07–21' to '2019–12–30'). The Nemo HDM model was at one nautical mile resolution and a temporal resolution of one hour. So, for three years this calculates to be (24 × 365 × 3) =26,280 hourly time steps. The study area grid points were (120×259) =31,082 for each feature (i.e., uwind, SST, pressure and DT). The training set is used to train the model, and the test set is to check the models' performance and generalization capabilities.

The selection of the hyperparameters influences the model's architecture, training process, and overall performance. This is one of the most important steps to prevent models from overfitting and having better generalization capabilities. An explanation of common hyperparameters used in the recurrent DL models is displayed in Table 3.

For the two models tested in this study, a trial-and-error method was implemented to determine the optimal hyperparameters. The list of chosen hyperparameters used in the two models is presented in Table 3. In addition, the lag window parameter was set as 12 h and the forecast

Table 3
Description of LSTM and GRU models' parameters and chosen hyperparameters.

Model parameters	Description	Chosen hyperparameter
LSTM/GRU Units	Specifies the dimensionality of the model's internal state.	512
Activation Functions	The activation function is applied after each layer in the model to add nonlinearity. The common choice for RNNs is 'Tanh' and Sigmoid.	'default'
Batch Size	Determines the number of samples used in each forward and backward pass during training.	128
Number of Training Epochs	Specifies how many times the model will be exposed to the entire training dataset during training.	50
Loss Function	Determines the objective function that the model is trying to minimize during training.	'MSE'
Optimizer	The optimizer determines the specific algorithm used to update the model's weights during training. Common optimizers include Adam, RMSprop, and Stochastic Gradient Descent (SGD).	'Adam'
Dropout Rate Regularization	A regularization technique that helps prevent overfitting. It specifies the proportion of neurons or units that are randomly dropped out during training, forcing the model to be more robust.	0.1
Kernel Regularization	Technique used to limit the model's weights with certain values. It adds a penalty term to the loss function based on the magnitude of the weights. Common regularization techniques include L1 and L2 regularization. The regularization Strength hyperparameter controls the strength of the kernel regularization	L2, 0.01

Table 4
Training results of LSTM and GRU models.

Model	Train	
metric	RMSE (cm)	R^2
LSTM	5.81	0.93
GRU	5.63	0.94

horizons include 3, 6, 9, 12, 24 h ahead timesteps (parameters *w* and Δ in Eq. (2)).

5. Evaluation results

5.1. General performance of the models

The LSTM and GRU models' performance was assessed across various time horizons (3, 6, 9, 12, 24 h) using two metrics, RMSE and R^2 as outlined in Eqs. (10) and (11). Higher R^2 values and lower RMSE values indicate a better fit of the model to the data. Based on the training results, for all three years averaged (Table 4), the GRU model demonstrates a better performance for the training process with an averaged R^2 of 0.94 and RMSE of 5.63 cm compared to the LSTM with 0.93 and 5.81 cm. Furthermore, for the test period, as shown in Fig. 7 and Table 5, both GRU and LSTM models show good performance. This demonstrates the applied LSTM and GRU models' robustness.

Results of Fig. 7 and Tables 4 and 5 demonstrate that the two models showed similar behavior given their comparable structures. As expected, the performance slightly deteriorates in both models as the time horizon increases. The GRU model, however, slightly outperformed LSTM with better results at all horizons, with higher R^2 and lower RMSE (Fig. 7). For instance, at 3 h horizon, the GRU model achieved an average RMSE of 3.55 cm, indicating that, on average, its predictions were approximately 3.55 cm different from the true values, whilst the LSTM yields RMSE as of 4.13 cm. The GRU R^2 value for this 3h was around 0.96, meaning the model explained 96 % of the variance in the data, demonstrating a strong predictive accuracy, whilst for the LSTM, it was 0.95.

At a 24-h horizon, both models' performance faced the most significant challenge, the GRU had an average RMSE of 5.99 cm, whilst the RMSE for LSTM values was 6.17 cm. Both LSTM and GRU attained R^2 of 0.89 indicating that they still had a good overall fit to the data and could explain 89 % of the variance.

Overall, both the LSTM and GRU models appear to be strong choices for sea level forecasting, providing almost similar and accurate forecasts within the specified time horizons. On average, the GRU model produced R^2 of 0.93 and a RMSE of 4.96 cm for daily predictions, while the LSTM model attained an R^2 of 0.92 and an RMSE of 5.3 cm over the test data. So, the GRU model outperformed the LSTM model when considering all time horizons.

5.2. Models' time-series performance

The temporal performance of the two models during the test period was also evaluated. This evaluation focused on the same four grid points (as depicted in Fig. 6) and generated time series plots for these locations, along with the predicted values provided by the LSTM and GRU models (see Fig. 8).

Fig. 8 illustrates the comparison of sea-level time series forecasting results using LSTM and GRU models for 3 h and 12 h prediction horizons across four test points (P1 to P4). The GRU model includes a 90 % confidence interval (CI), and its reliability is assessed using Prediction Interval Coverage Probability (PICP) values (see Eq. (13)). The PICP is a metric used here to assess the reliability of the model's uncertainty estimates. Both GRU and LSTM models generally follow the observed sea level trends well. However, the GRU model demonstrates slightly better accuracy in capturing rapid changes and variations, particularly for longer forecasting horizons like 12 h. The GRU's 90 % CI effectively encompasses the observed values, as reflected in high PICP values exceeding 90 % in most cases. For example, at P1, the PICP is 92.17 % for the 3 h horizon and 90.71 % for the 12 h horizon, showing the GRU model's strong reliability. The performance remains consistent across all points, although P2 at the 3-hour horizon has a slightly lower PICP of 89.64 %, indicating marginally reduced robustness at that location. Similarly, for the 12-hour horizon, the PICP values for P1, P2, P3, and P4 range between 90.25 % and 91.02 %, all indicating strong performance in uncertainty quantification. These values reflect the model's robust ability to maintain reliable and consistent predictions across different forecasting horizons, with the PICP consistently staying close to or above the target of 90 %. The GRU model's confidence intervals widen for the 12 h horizon compared to the 3 h horizon, reflecting the expected increase in forecast uncertainty over longer time frames. Despite this, the intervals still capture most of the observed values, demonstrating the model's strong uncertainty quantification.

However, for both models, there were instances where the results were less favorable, specifically for grid point P1 (mostly during Nov-Dec, when the sea level extremes occurred). Thus, both models appear to have problems with forecasting the extreme sea level values at all stations. Possible reasons for this will be discussed in more detail in

Table 5
Results of the performance of the DL models on different time horizons for the selected test period ('2019–07–21 to '2019–12–30).

Horizons (hours)	Models			
	GRU		LSTM	
	R^2	RMSE (cm)	R^2	RMSE (cm)
3	0.96	3.55	0.95	4.13
6	0.95	4.41	0.94	4.85
9	0.92	5.16	0.92	5.47
12	0.91	5.67	0.90	5.87
24	0.89	5.99	0.89	6.17
average	0.93	4.96	0.92	5.3

Fig. 7. General performance of the LSTM, and GRU models during test period at different time horizons (3, 6, 9, 12, 24 h ahead) using spatially averaged R^2 and RMSE.

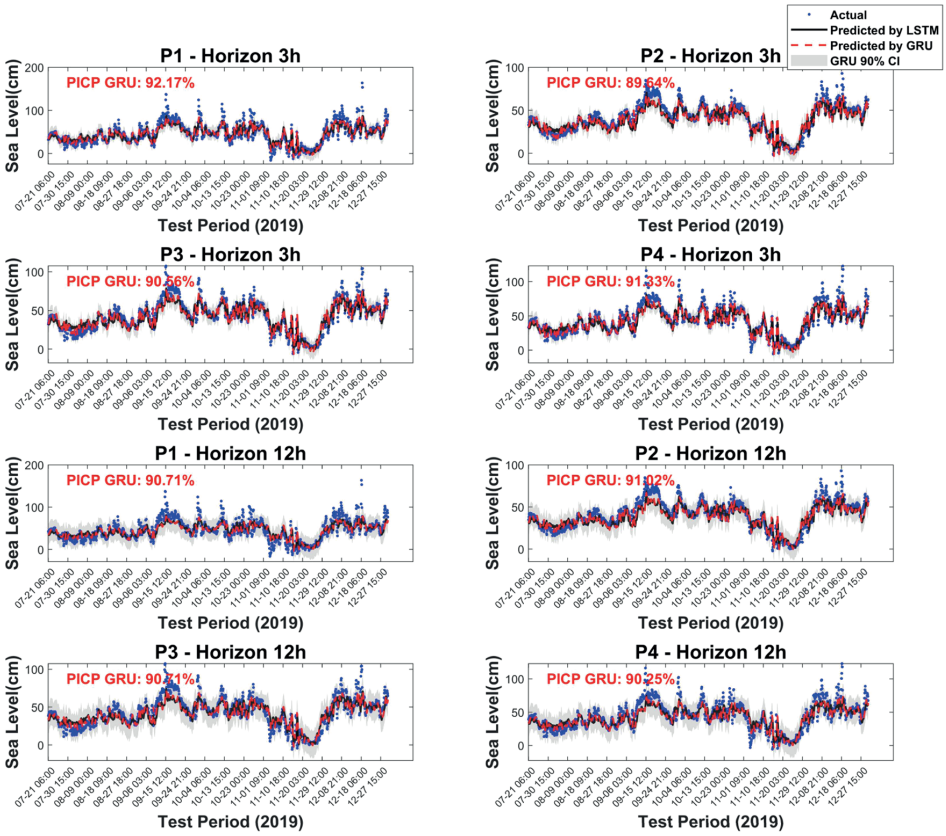


Fig. 8. Time series forecasting of LSTM and GRU models at different gridpoints (P1 to P4) for horizons 3 h and 12 h ahead, shaded by 90 % confidence interval predictions by GRU model and PICP values.

Section 6.

Figs. 9 and 10 illustrate the scatter density plots to assess the models' predictive performance for grid points P1 to P4 at various time horizons. Each point's color denotes the density of data at specific location, aiding in the identification of data concentration. The scatter plots mostly display a balanced distribution of data points both above and below the $y = x$ line, indicating equitable treatment between overestimation and underestimation. The positive regression "a" coefficient further highlights the model's accurate agreement with actual DT values.

The results depicted in Fig. 9 indicate that the LSTM model performed notably well for grid points P2 and P3 located on western and southern sections of gulf, yielding RMSE values ranging from 7 to 9.8 cm at time horizons of 3 and 24 h. Additionally, the model's predictions for DT values at these grid points exhibited a strong correlation, with R values (the root of R-squared in Eq. (11)) of 0.93 and 0.83 at the same horizons. LSTM model also showed relatively good performance in forecasting DTs in P4, with RMSE and R ranging 8 − 13 cm and 0.92−0.75 for 3 h and 24 h horizons. In contrast, grid point P1, located in the eastern part of gulf, experienced less favorable predictions, with RMSE values mostly exceeding 10 cm and R values lower than 0.8. These findings reflect the inherently complex nature of DT predictions in this specific geographical area. In terms of the WI index, the LSTM model demonstrated strong performance, particularly for P2 and P3, where WI values were generally above 0.9 across all forecasting horizons. Performance remained satisfactory for P4, with WI values exceeding 0.84. The lowest performance was observed for P1 at the 24 h horizon, where the WI value reached 0.78.

Fig. 10 also presents the scatter density plots for P1 to P4 for the GRU model. Observe that for all grid points P1 to P4, the GRU model had better performance than LSTM at all horizons. However, like the LSTM model, it had also poorer results for P1 (most easterly station), as compared to other grid points.

5.3. Spatial performance of the GRU model

This section presents the findings concerning the spatial performance of the GRU model, given its superior performance over the LSTM model. As illustrated in Fig. 11, an analysis of the GRU model's average performance during the test period spanning from '2019–07–21 06:00:00′ to '2019–12–30 24:00:00′ was performed. Examining the error plots (predicted minus observed DTs), it is evident that the model generally exhibits strong performance, with the majority of errors within a range of ±2cm. Notably, the model's performance remains relatively consistent across different time horizons.

To further understand the model's instantaneous performance, we also examined its spatial performance at a specific time instant ('2019–11–08 00:00:00′), as shown in Fig. 12. At this time point, the model demonstrated a strong ability to simultaneously forecast various grid points, with errors averaging around 3 cm and a standard deviation within 2 cm for all horizons. However, some larger errors were observed in the easternmost gulf (around −4 cm).

We also evaluated the performance of the GRU model's prediction

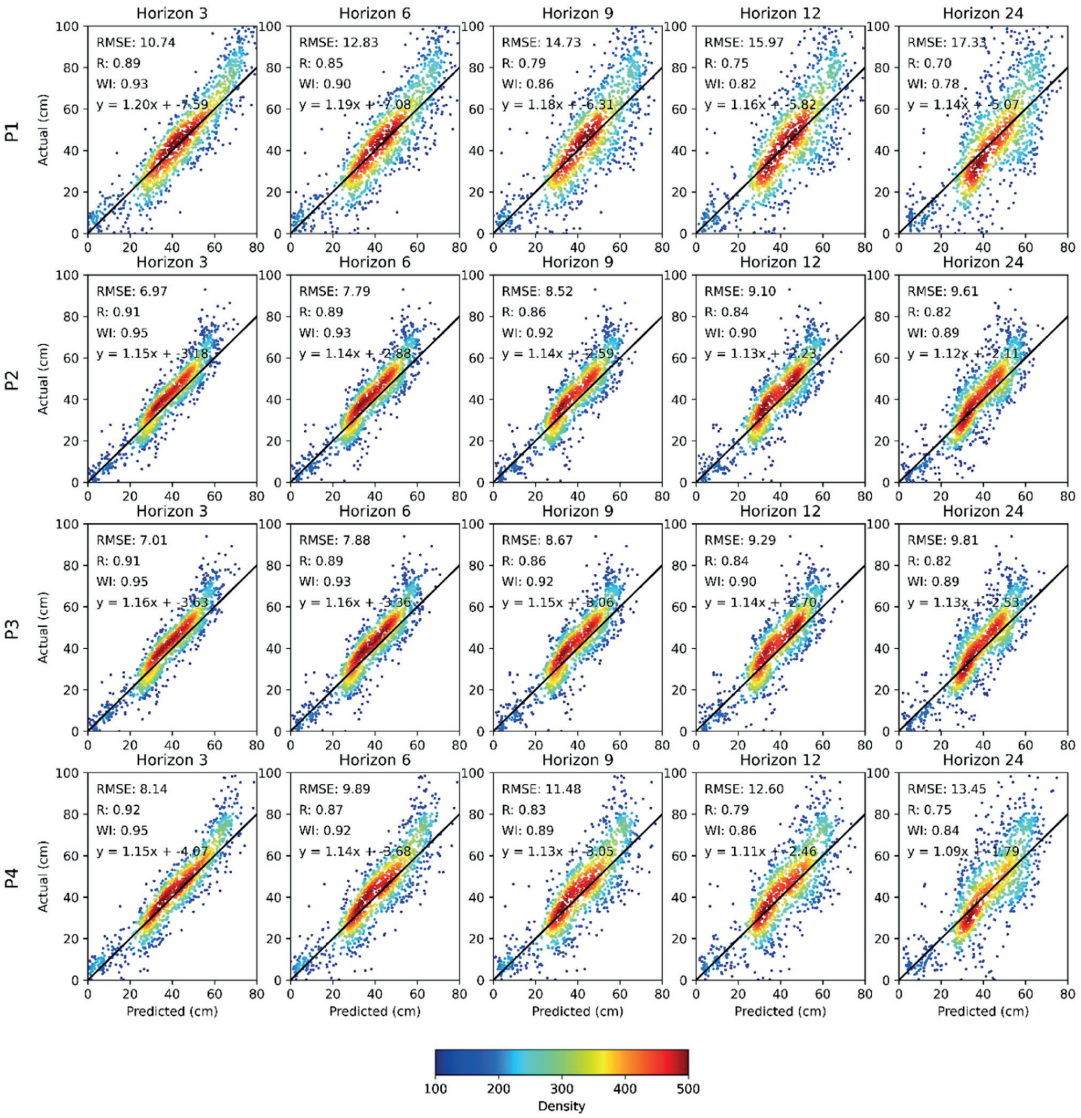


Fig. 9. Scatter density plot between actual and predicted DT values at grid points P1 to P4 for LSTM model during test period at different horizons (3, 6, 9, 12, 24 h).

errors for horizon 3 h at specific grid points (P1 to P4), as illustrated in Fig. 13. Overall, the model performed well during the summer months of July and August, with errors typically ranging from −3 to 3 cm at all points. A similar error range of −5 to 5 cm was observed for points P2 and P3 in October and November. However, the model's predictions were less accurate at P1 and P4 during September and December. Specifically, at P1, the model exhibited larger errors, likely due to an underestimation of peak sea levels in the eastern Gulf during these months.

Fig. 14 also provides an intercomparison of the residuals at four gridpoints (P1, P2, P3, P4) at various forecast horizons, segmented by three quantiles: the 25th percentile (Q1), the median (50th percentile or Q2), and the 75th percentile (Q3), giving insight into how the model's predictions vary in terms of accuracy, bias, and uncertainty. The median (Q2) residuals across all points (P1, P2, P3, P4) are relatively stable and do not exhibit large variations over the forecast horizon, indicating that the model's predictions for the median values are more consistent and reliable over time. Such consistent performance suggests that the model is fairly accurate in capturing the central tendency of the sea level data, with minimal bias, across all forecast horizons. The difference between the 75th percentile (Q3) and 25th percentile (Q1) residuals and the widening gap suggests an increase in uncertainty as the forecast horizon increases, which is a typical behaviour in predictive models, particularly in time series forecasting.

The analysis of the residual histograms also reveals that the model performs best at shorter forecast horizons (3 to 9 h), with errors concentrated around zero and minimal variability. As the forecast horizon extends to 12 and 24 h, the spread of the residuals increases, particularly for grid points P1, indicating greater uncertainty, suggesting that the model's ability to predict sea levels decreases over time, especially at P1 with more complex dynamics.

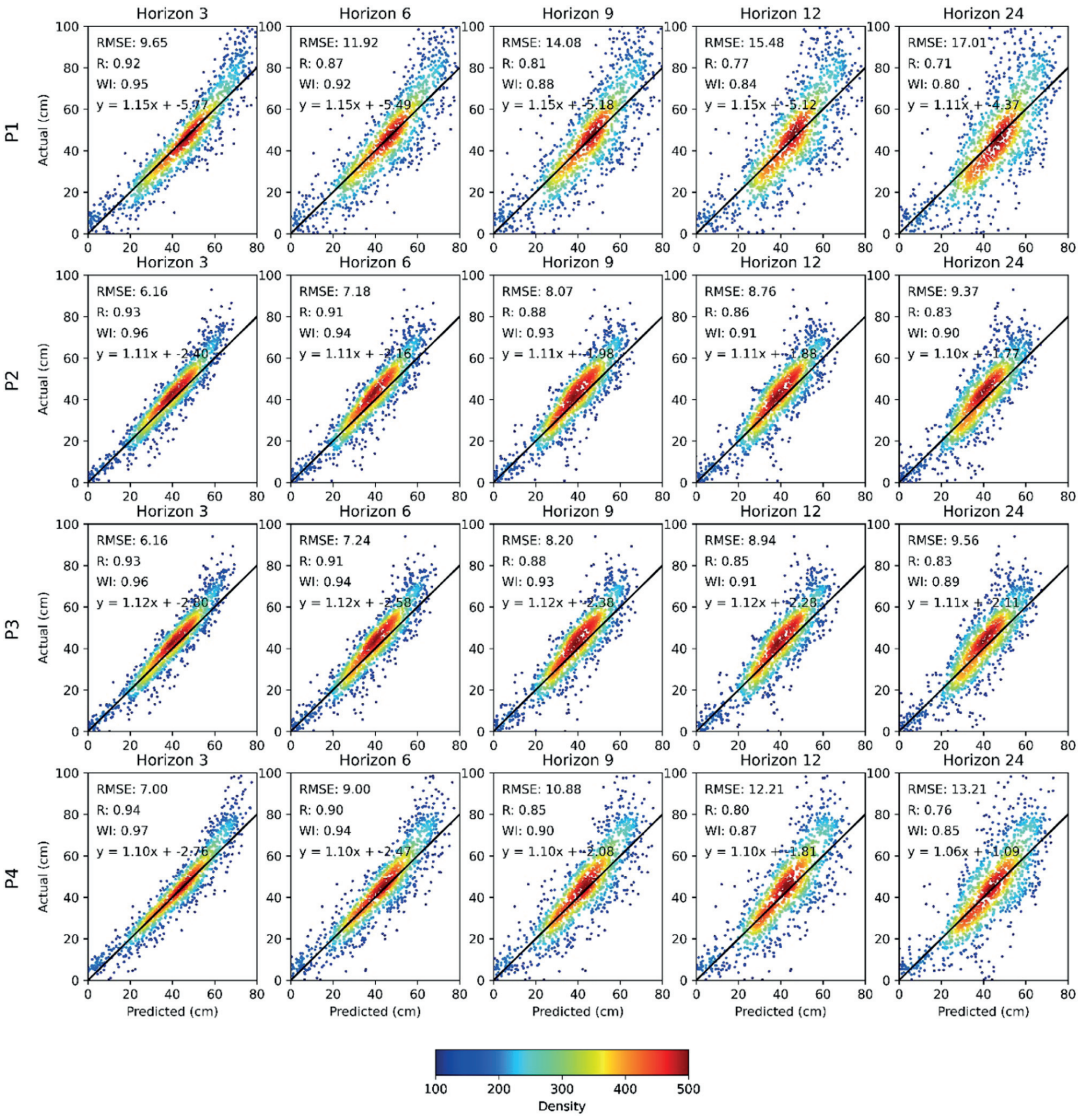


Fig. 10. Scatter density plot between actual and predicted DT values at gridpoints P1 to P4 for GRU model during test period at different horizons (3, 6, 9, 12, 24 h).

5.4. Model validation with satellite altimetry data

As both GRU and LSTM models demonstrated nearly identical performance, with GRU slightly outperforming, only the GRU model is employed for external validation by the SA data. Therefore, a comparison of the GRU forecasted DT results for horizon 3 h, with the along-track satellite altimetry observations and the actual DTs from corrected HDM is performed. For this purpose, numerous satellite tracks of Sentinel 3A and Sentinel 3B missions (using the 1 Hz and 20 Hz along-track data) were selected during the test period at different satellite cycles.

Based on previous studies in this region that validated along-track SA measurements against HDM model, a systematic bias of up to ±20 cm can be expected between both S3A and S3B with HDM sea level source, specifically within the eastern part of gulf, mainly due to the geoid problems with insufficient coverage of marine gravity data (Ågren et al., 2016; Jahanmard et al., 2023a). This may affect the selected SA tracks in the central part of gulf.

Fig. 15 shows the selected tracks and cycles within the Gulf of Finland also the comparison of different DTs for these tracks. Examination of the figure shows that both SA 20 Hz and smoothed 1 Hz measurements generally show more high-frequency variations compared to the HDM DT and GRU DT. These higher-frequency observations are expected for the SA measurements, as they capture the actual reality of instantaneous sea surface (including some background noise). Whilst the HDM model was based on mathematical ocean modelling equations that more or less represent the lower frequency of the sea level (Jahanmard et al., 2023a; Mostafavi et al., 2023).

Furthermore, Fig. 15 shows that the GRU-forecasted DTs and the HDM DT are on most occasions in good agreement with SA DT values,

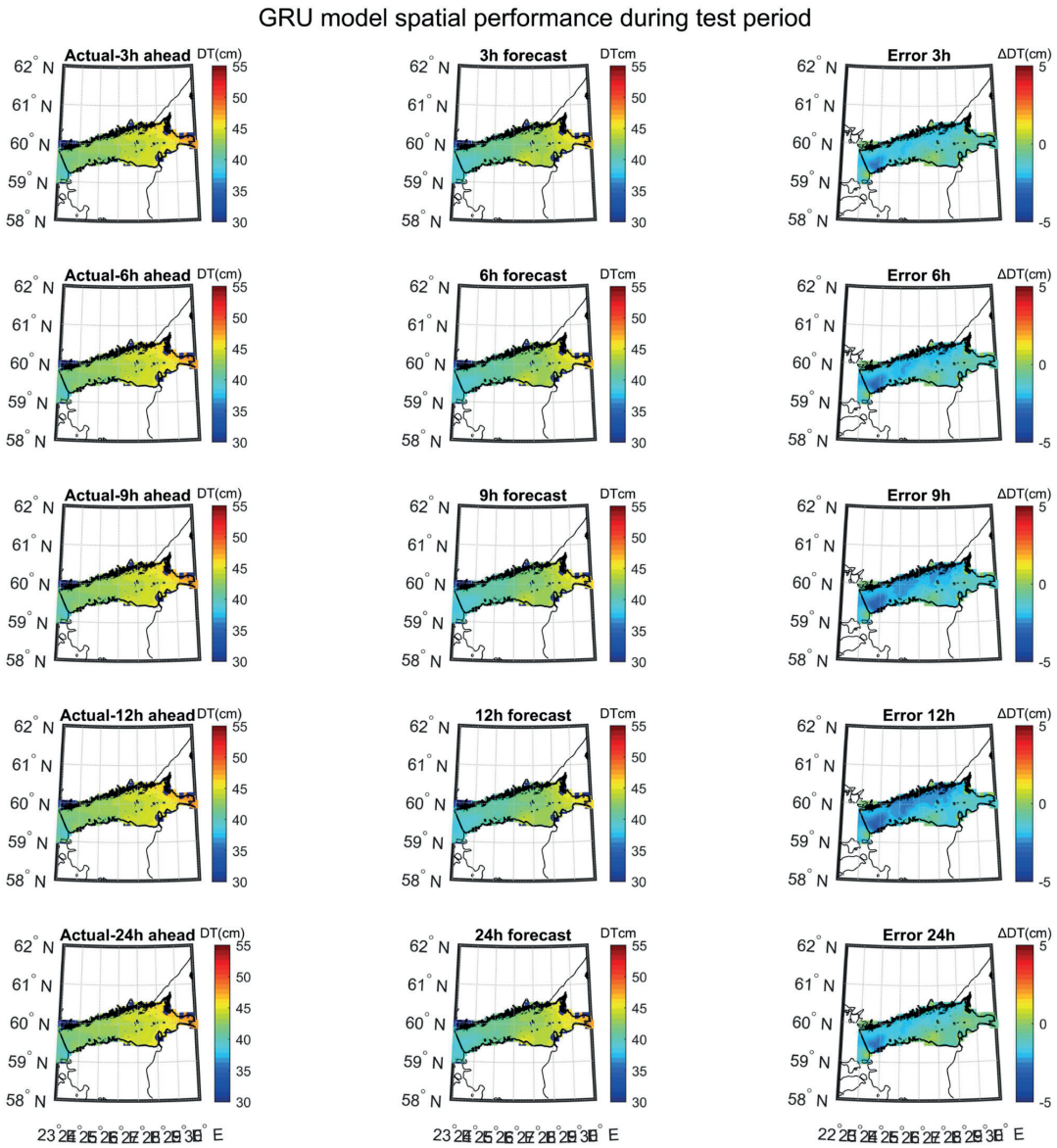


Fig. 11. Spatial performance of the GRU model during the test period comparing the actual, predicted and the error (predicted minus actual) at time horizons 3, 6, 9, 12, 24 h during entire test period.

with discrepancies of lower than 5 cm for tracks S3A-83, S3A-300, S3A-414, S3B-83 and S3B-197. However, the GRU model had poorer validation results for tracks S3A-739, S3A-186, S3B-739, and S3B-299 (10–15 cm). The reason for these larger discrepancies may be due to the HDM model not accurately capturing the observed ocean dynamics. This indicates that the DL model outcome is only as good as the initial data and criteria that were used. So, if there are some inadequacies in the initial data, the DL model would be affected. The reason for these higher discrepancies at these locations should be examined in future studies. Additionally, the DT forecasts and HDM corrected DTs in general had better consistency with Sentinel 3A tracks compared to the Sentinel 3B, which is also in agreement with previous studies (Mostafavi et al., 2023).

6. Discussion

The results of this study indicate that both Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models can be effectively applied in sequence-to-sequence sea-level forecasting tasks. The feature selection results of Fig. 6 showed that the zonal east-west wind speed

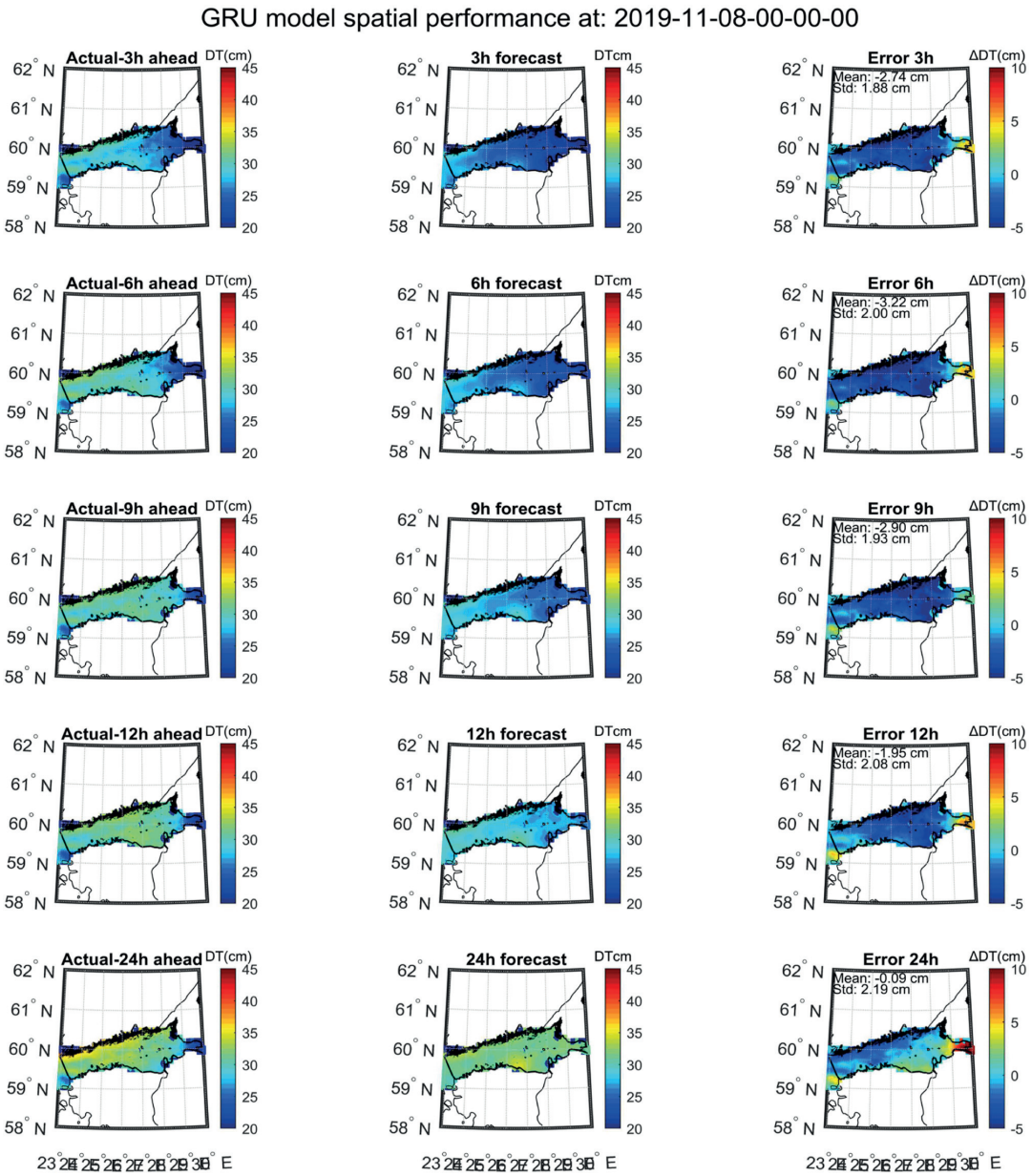


Fig. 12. Spatial performance of the GRU model during the test period comparing the actual, predicted and the error (predicted minus actual) at time horizons 3, 6, 9, 12, 24 h for the arbitrary time instant '2019-11-08 00:00:00'.

(uwind), air pressure and SST have higher correlations with DT variations. Similar studies with short-term sea-level predictions in the Baltic Sea have also shown these input components to be the most applicable variables (Bellinghausen et al., 2023; Rajabi-Kiasari et al., 2023). Results showed that both DL models have almost similar performance (Tables 4 and 5 and Fig. 7). However, the GRU model slightly outperformed the LSTM model by better training and test accuracies (with R^2 and RMSE of 0.93 and 4.96 cm, as compared to that of LSTM 0.92 and 5.3 cm, respectively). The main difference between the LSTM and GRU models was that the GRU model has a simpler architecture in storing and updating the connections between the different gates (see Section 2.2.2 and Fig. 3). The reasons behind the superior results of the GRU model compared to the LSTM model can be due to its simpler architecture, which uses only two gates (update and reset) compared to the three gates in LSTMs (input, forget, and output). This simplicity makes GRUs faster to train, less prone to overfitting, and more stable. The fewer parameters in GRUs help reduce the risk of overfitting and improve generalization to unseen data. The GRU's superior results compared

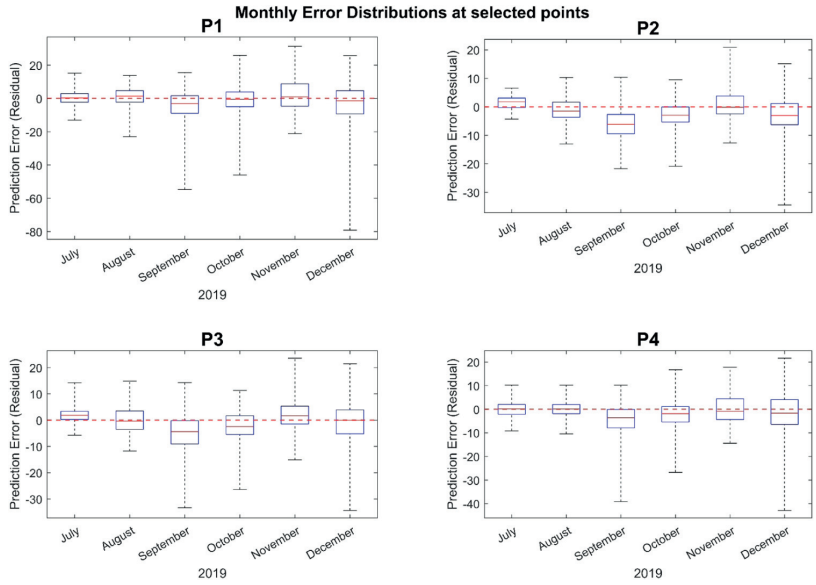


Fig. 13. Boxplot of GRU model errors (predicted minus actual) at selected spatial grid points for horizon 3 h, during different months in the test period.

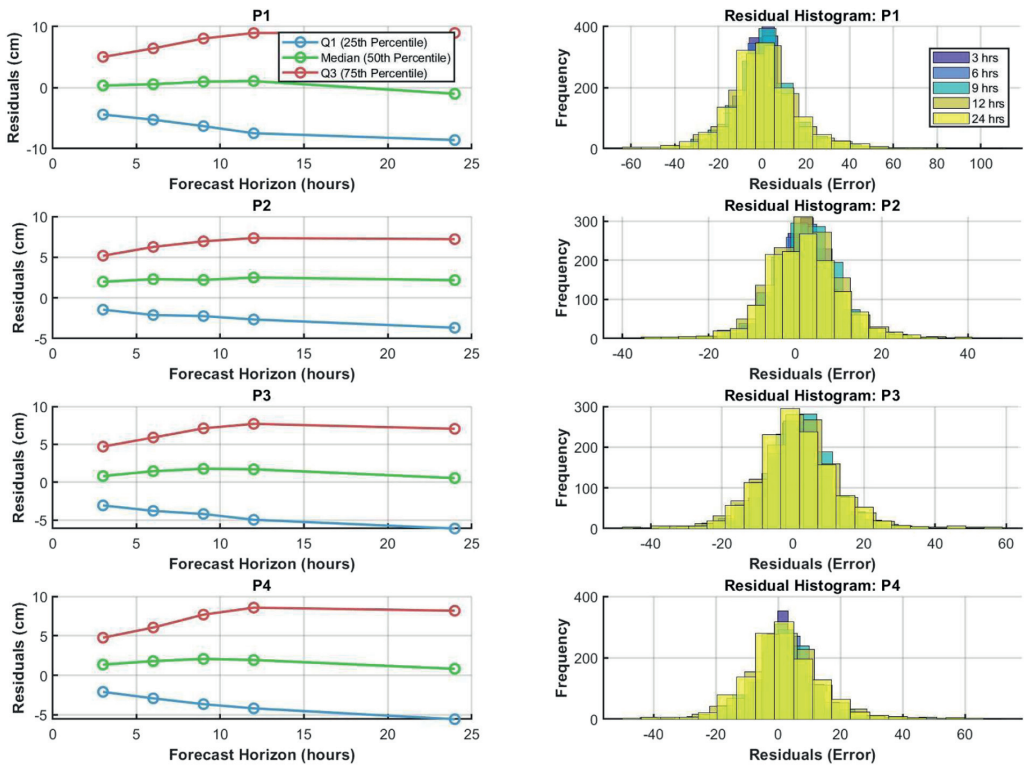


Fig. 14. Intercomparison of GRU model's residual analysis based on 25 %, 50 % and 75 % quantiles and histogram of errors at different grid points (P1 to P4) and forecast horizons (3 h to 24 h).

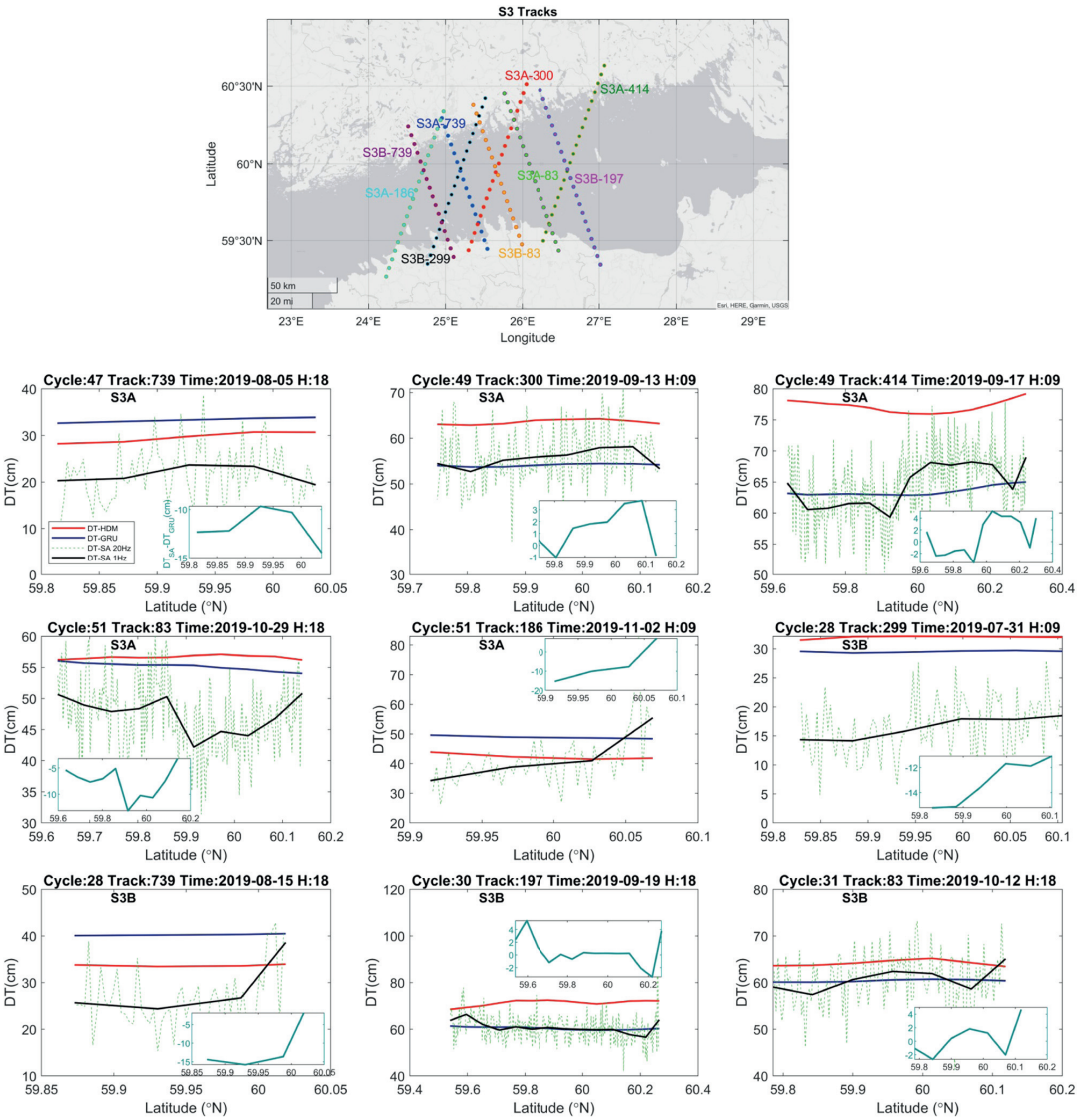


Fig. 15. Validation of DT forecast results by GRU model against HDM DT and SA-derived (20 Hz and smoothed 1 Hz) DT at different tracks and cycles within the test period in Gulf of Finland. The inset figures show the discrepancy between SA and GRU forecasted DTs.

with LSTM for short-term time series forecasting have been also reported in previous studies (Chia et al., 2022; Jiang et al., 2024; Li et al., 2022). Now, this is also confirmed for the Baltic Sea level forecasting.

The performance of the DL models was also analyzed by time series data at four selected grid points within the region. Notably, the models demonstrated superior performance at grid points P2, P3, and P4, situated in the northern, southern, and western Gulf regions, in comparison to P1, which is located in the eastern gulf and showed more discrepancies. This observed difference can be attributed to several factors. One of the key factors contributing to this variation is the selection of the testing period, which mostly falls within the winter months of 2019. Seasonal correlation plots in Fig. 6 revealed that during the winter months, SST and atmospheric pressure exhibited a more pronounced influence on DT variations. These influential factors were incorporated into the models, aligning with the prevailing conditions during the test period at grid points P2, P3, and P4, thus enhancing their predictive accuracy. Conversely, grid point P1 in the eastern gulf, located near the Neva River, presents a distinct scenario. In this location, the river discharge and lower salinity values may play a significant role in regulating the sea level variations (Alenius et al., 1998). Also, the vwind component had a negative correlation (around −0.3) with DT in the winter months at stations P1 and P4 (see Fig. 6). In addition, for P1, higher correlation was determined for salinity and vwind with DT (higher than 0.3) for autumn and winter. Incorporating river discharge, salinity, and vwind components could potentially improve sea level models in the eastern Gulf of Finland, as these factors might influence sea level dynamics (Alenius et al., 1998). However, these additions do not necessarily address modeling challenges, as data-driven machine

learning (ML) models often struggle with capturing complex patterns such as resonance oscillations and extreme sea levels. These input components (vwind, sea surface salinity, river discharge) were not included in the final variable selection to manage the demanding computational load on the high-resolution grids. Future studies should further explore how incorporating these factors into physics-based models could address current limitations.

The time series performance comparison in Fig. 8 showed that both LSTM and GRU models closely track observed sea level trends, with the GRU model achieving slightly higher accuracy. The GRU model demonstrates robust uncertainty quantification, as indicated by high Prediction Interval Coverage Probability (PICP) values, exceeding 90 % in most cases. For instance, at P1, the PICP is 92.17 % for the 3-h horizon and 90.71 % for the 12-h horizon, highlighting strong reliability. While P2 shows marginally lower PICP at the 3-h horizon (89.64 %), overall performance remains consistent across points and horizons. The GRU model's confidence intervals effectively capture observed values, although they widen slightly for longer horizons.

Studies of scatter density plots in Figs. 9 and 10 using the WI index showed strong performance for both models, specifically for the GRU, with WI values generally above 0.9 for P2 and P3, satisfactory performance for P4 (above 0.84), and the lowest performance for P1 at 24-h horizon (0.78). Furthermore, spatial performance plots (Figs. 11 and 12) for the GRU model showed a consistent error at different horizons with mean and standard deviations of errors within 2 cm for the test period. Some larger discrepancies occur in the eastern gulf. The possible reasons for the eastern Gulf of Finland experiencing complex hydrodynamic conditions that challenge accurate modeling can be due to several reasons; (i) Salinity gradients are particularly sharp in this region, where freshwater from rivers like the Neva mixes with the brackish Baltic Sea, especially during winter (Soomere et al., 2008); (ii) The area also experiences extreme sea level variations, particularly in winter, due to forced progressive waves and resonance-induced seiches, which complicate predictions (Alenius et al., 1998; Jahanmard et al., 2023a); (iii) The significant river discharge and the small Rossby radius contribute to strong local currents, further complicating sea level modeling (Soomere et al., 2008, 2009); (iv) Additionally, specific oceanographic conditions and seasonal ice cover influence water levels, adding to the region's modeling challenges (Alenius et al., 1998; Sukhachev et al., 2014).

As shown in Figs. 8–10, the model underestimated sea level maxima values in most cases. Previous studies have also demonstrated the MLs' tendency to under-predict extreme sea levels (Qin et al., 2023). This may be attributed to several factors. One key consideration is the insufficient representation of maximum/extreme events in the training dataset (skewness towards normal sea levels than extremes) (Ramos-Valle et al., 2021), limiting the model's ability to learn the complex patterns and interactions associated with sea level peaks. Previous studies in the same region have suggested using Peak-above-threshold techniques to differentiate between extreme stormy and non-stormy events using classification ML models (Bellinghausen et al., 2023). Hence, future studies can consider such techniques in more detail to address sea levels in extreme cases.

Additionally, the model's sensitivity to extreme conditions, such as storm surges and high wind speeds, may not be adequately captured in the present study, leading to systematic underestimations. These issues can be addressed in future studies by introducing the contributing variables for extreme sea levels. For instance, in the gulf, the wind climate closely influences the sea level, and the extreme wind speed tends to occur with two maxima corresponding to SW and north (occasionally NNW) winds (Keevallik and Soomere, 2003). In this study, only the uwind was considered. So, future studies may consider both u and vwind components. To improve forecasting of maxima events, another future improvement can be customizing the model's loss function to be penalized for extreme events (Man et al., 2023) as well as employing optimization algorithms to improve the model's performance in forecasting extreme sea level events.

In Fig. 13 for boxplots of errors at selected grid points, it is demonstrated that the model struggles with seasonal and spatial patterns in the eastern Gulf of Finland (P1), especially during winter when extreme sea level variations occur. These extremes, driven by strong winds, pressure systems, and complex interactions between freshwater inflows and salinity gradients, are often underestimated by the model due to the region's unique hydrodynamic conditions. The model's poorer predictions at P1 and P4 in September remain unclear and require further investigation. These locations have distinct salinity levels—lower at P1 and higher at P4—which might affect the model's accuracy, unlike the smoother salinity ranges at P2 and P3. Additionally, Alenius et al. (1998) noted that wave height peaks in the Gulf of Finland typically occur in autumn, potentially explaining the stormy conditions in September that the model could not replicate. Future studies should consider these salinity variations and autumnal wave height peaks to enhance model performance in these areas.

Fig. 14 presents the residual analysis at different points and horizons based on quantile and histogram analysis of errors. The residual analysis showed that the GRU model consistently captures the median sea level values with minimal bias across all points and horizons, demonstrating its reliability in modeling central tendencies. However, predictive uncertainty grows with longer horizons, particularly at P1, due to more complex dynamics, highlighting challenges in maintaining accuracy over time. Future work could improve uncertainty quantification by integrating probabilistic Bayesian forecasting, enabling interval-based predictions instead of relying solely on point forecasts.

Furthermore, to validate the GRU model's forecasted results, we compared them with independent satellite altimetry observations from Sentinel-3A and Sentinel-3B during the test period. The input data, derived from bias-corrected sea level data referenced to a high-resolution geoid model (NKG 2015), was aligned with satellite SSH measurements through data transformation and atmospheric corrections. Temporal matching was also performed for the 3 h ahead forecasts. Validation results against independent satellite observations in Fig. 15 showed a discrepancy of better than 5 cm for most tracks within the Gulf of Finland. This proves the compatibility between the employed HDM DT source and the reliable DT forecasts by GRU when compared with ground truth sea level measurements. Despite differences in resolution, with satellite data offering finer details than the model predictions, we ensured comparability by selecting tracks in regions with minimal geoid inaccuracies. Previous studies have shown good correlation between the satellite altimetry and model data and the final results also confirm that these datasets are indeed comparable for the study region.

Finally, the two applied models generally had similar performance with the subtle differences not always obvious. However, given the slightly superior results achieved by the GRU model and its faster training, reduced complexity, simpler structure with fewer gates and parameters than LSTM, then GRU is more advantageous for short-term sea level time series forecasting (up to 24 h) within the study area.

In comparison with prior research utilizing similar sea level data source in the Baltic Sea, the GRU model achieved an average 24 h ahead DT forecast with RMSE of 4.96 cm. This performance is comparable to the 4.7 cm accuracy achieved by a CNN model in the previous study (Rajabi-Kiasari et al., 2023). It is worth noting that the earlier study adopted a longer time window (7 days), a resource allocation not feasible in our current research due to computational limitations. Additionally, our focus here was on accurate high spatio-temporal resolution, providing hourly predictions across various future horizons with a restricted lag window of only 12 h. Despite these constraints, the results are promising.

Several studies have reported the effectiveness of traditional ML models like ANNs in sea-level forecasting (Kim et al., 2019; Sztobryn, 2003), while others have utilized ML models like random forests (Bellinghausen et al., 2023). However, for spatiotemporal frameworks,

recent research has increasingly employed hybrid techniques, such as combining convolutional neural networks with recurrent models, to capture both spatial and temporal features (Li et al., 2024; Wang et al., 2021). The study tested an early, non-optimized version of a hybrid CNN-LSTM model, as suggested in the literature, but excluded them due to their time-consuming training and the need for a computationally efficient high-resolution spatiotemporal model. This model did not appear to outperform deep recurrent models, which have been prioritized for sea level forecasting in the region (Tiggeloven et al., 2021).

Future studies can consider these recurrent models (LSTM and GRU) in hybrid structures along with hyper-parameter tuning through various optimization techniques. Moreover, here the methodology was examined for the Gulf of Finland basin, given its higher DT variation rates, within the Baltic Sea region, known for its highly nonlinear patterns, as the sea level controlling factors (wind speed and pressure) are noisy in time (Kulikov et al., 2015). Further research in diverse regions may yield even more improved results. Additionally, as possible improvements, the signal decomposition techniques e.g., wavelet transforms can be applied to the data to convert them to frequency domain which has been shown a powerful tool in sea level forecasting problems (Fu et al., 2019; Jin et al., 2023; Raj et al., 2022; Song et al., 2022; Zhao et al., 2021).

7. Conclusion and future studies

This study explores the potential of two popular recurrent-based neural networks (LSTM and GRU models), to spatio-temporally forecast sea level/dynamic topography (DT) at different time horizons (3 h, 6 h, 9 h, 12 h and 24 h time steps) in the Gulf of Finland, Baltic Sea. One of the novelties of the method examined is the usage of sea level obtained from a hydrodynamic model (HDM) that has been corrected for vertical biases using geoid-referred tide gauges. The benefit of this is in using a more accurate sea level data source, allowing better estimates of the connections made with other drivers affecting the sea levels. As a result, careful exploration of various input components revealed that historical DT, SST, air pressure, and uwind (zonal wind speed) were the dominant features with mostly influenced future DT in the eastern Baltic Sea. Input data selection was based on their established importance in sea level dynamics, as supported by both exploratory data analysis and prior studies. While we acknowledge the limitations of excluding variables such as salinity, river discharge, and meridional winds, this study highlights the foundational role of the selected factors. It also lays the groundwork for future research to investigate the contributions of additional variables, which could further enhance model accuracy and deepen our understanding of sea level dynamics. Since we have also taken the direction of forecasting DT, which is referred to the geoid, this advantageously enables an independent external validation possible with sea level data obtained from satellite altimetry.

The results showed a slightly greater spatio-temporal forecasting superiority of the GRU model at all time steps when compared with the LSTM model. For instance, the GRU model showed good spatial and temporal capabilities with an averaged R^2 of 0.93 and RMSE of 4.96 cm within the test period, hence this model can be regarded as the most suitable method for short-term high-resolution sea-level forecasting applications. Even though both models performed reasonably well, some problematic areas were identified in the eastern Gulf of Finland which may be due to river discharge, salinity or vwind (meridional) being excluded from the input components. Also, other complex hydrodynamics such as strong local currents, resonance-induced seiches and seasonal ice cover may have influenced these problematic areas. The forecasting of sea level maxima events was noticeably problematic, and this was apparent by larger discrepancies. Our method and results suggest that whilst various DL models are sufficiently capable of forecasting sea level data, it is also necessary to meticulously choose input variables that may also account for sea level maxima, as they impact result accuracy.

As a result, our examination of recurrent-based neural network approaches revealed that they are more or less suitable for normal sea level conditions with a tendency to underestimate extreme sea levels, particularly in the eastern Gulf of Finland. Future research that forecasts sea levels would require a longer historical data set, additional input variables (e.g. wind stress and wave properties (Harter et al., 2024) along with water volume index, river runoff, evaporation, and precipitation) and even different ML approaches or ensemble of different models.

Also, given the complexity of the study area and in order to overcome resource constraints, future studies could focus on specific geographical locations with distinct characteristics. Such consideration allows the utilization of more advanced model types, such as hybrid models to have the benefits of both convolutional and recurrent models or attention-based architectures to help the model focus on more relevant parts of features and previous timesteps. Implementing advanced optimization techniques could also enhance model performance and extend its applicability to a wider range of conditions. Future studies can also enhance model interpretability by incorporating explainable AI techniques and integrating data-driven methods with physics-based models and Bayesian inference.

It should be noted though that whilst other future approaches can be considered, the proposed GRU-based spatiotemporal sea level forecasting framework offers significant practical advantages, making it highly suitable for real-world applications. Its ability to simultaneously predict sea levels across multiple grid points and time steps improves both efficiency and accuracy by capturing spatial correlations and extending forecast horizons. This is crucial for timely decision-making in applications such as flood warnings, maritime navigation, and coastal infrastructure planning. Unlike traditional computationally intensive numerical models, the GRU framework provides accurate predictions within seconds, making it an excellent alternative for operational forecasting. Its relatively simple architecture enables faster training, fewer parameters, and easier deployment. These features make the model particularly advantageous for real-time sea level tracking, vessel scheduling, under-keel clearance management, port operations, and supporting ecosystem management strategies to mitigate the impacts of sea level rise and extreme weather.

The paper also advances sea level forecasting by applying DL to regions with complex and nonlinear dynamics such as the Baltic Sea. The model leverages fine-scale hydrodynamic data, enabling it to generalize across varying weather systems and accurately predict sea levels. Such a high-resolution framework, both in space and time, can be more effective in capturing the rapid and localized sea level changes in the Gulf of Finland, driven by short-term atmospheric factors like wind and air pressure (Medvedev et al., 2021). For example, hourly time steps may enable the model to track these quick fluctuations in wind patterns, which occur over a few hours to days, more accurately (Kulikov and Medvedev, 2013), which could be important for applications like maritime safety, storm surge forecasting, and coastal management. This level of detail can be necessary to capture variations occurring within hours to a few days, offering potentially more precise and timely predictions compared to daily or weekly time steps. Focusing on the Gulf of Finland, the study addresses rapidly changing small-scale features like eddies and fronts, which require a resolution of 0.5–1 nautical mile (Soomere et al., 2008; Kowalewski and Kowalewska-Kalkowska, 2017). The GRU model provides centimeter-level accuracy, validated by Sentinel-3A/3B data. It is more efficient than traditional models, offering near-instant multistep forecasts for real-time applications. It also acknowledges and addresses limitations, such as underestimation in the eastern Gulf of Finland, while highlighting the significant benefits of integrating DL models with high-resolution hydrodynamic data. For example, DL models like LSTM and GRU efficiently capture complex, nonlinear relationships in data with lower computational costs compared to traditional hydrodynamic models, making them ideal for operational forecasting. Our forecasting framework, initially designed for the micro-tidal Baltic Sea, specifically the Gulf of Finland, focuses on

meteorological factors such as wind and atmospheric pressure. Adapting this framework to other coastal areas with different climatic and hydrodynamic conditions is feasible with architectural modifications and transfer learning techniques, but success depends on the availability of high-quality input data and the appropriate level of model complexity and resolution. The GRU model is adaptable, but it may face challenges in generalization if new conditions differ significantly from its training data. Therefore, further testing would be needed to assess its scalability across diverse environments.

CRediT authorship contribution statement

Saeed Rajabi-Kiasari: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Artu Ellmann:** Writing – review & editing, Validation, Supervision, Conceptualization. **Nicole Delpeche-Ellmann:** Writing – review & editing, Validation, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research was supported by the Estonian Research Council grants PRG1129 and PRG1785 “Development of continuous DYNAmic vertical REFerence for maritime and offshore engineering by applying machine learning strategies /DYNAREF/". The authors wish to convey their appreciation to Dr. V. Jahanmard for facilitating access to the corrected HDM model data. We would also like to express our gratitude to the anonymous reviewers for their valuable and constructive feedback.

Data availability

Data will be made available on request.

References

Accarino, G., Chiarelli, M., Fiore, S., Federico, I., Causio, S., Coppini, G., Aloisio, G., 2021. A multi-model architecture based on long short-term memory neural networks for multi-step sea level forecasting. Fut. Gen. Comput. Syst. 124, 1–9. https://doi.org/10.1016/j.future.2021.05.008.

Ågren, J., Strykowski, G., Bilker-Koivula, M., Omang, O., Märdla, S., Forsberg, R., Ellmann, A., Oja, T., Liepiņš, I., Paršeliūnas, E., Kaminskis, J., Sjöberg, L.E., Valsson, G., 2016. The NKG2015 gravimetric geoid model for the Nordic-Baltic region. In: 1st Joint Commission 2 and IGFS Meeting International Symposium on Gravity, Geoid and Height Systems, pp. 8–9.

Ahmed, N., Assadi, M., Zhang, Q., 2023. Investigating the impact of borehole field data's input parameters on the forecasting accuracy of multivariate hybrid deep learning models for heating and cooling. Energy Build. 301, 113706. https://doi.org/10.1016/j.enbuild.2023.113706.

Alenius, P., Myrberg, K., Nekrasov, A., 1998. The physical oceanography of the Gulf of Finland: a review. Boreal Environ. Res. 3, 97–125.

Altunkaynak, A., Kartal, E., 2021. Transfer sea level learning in the Bosphorus Strait by wavelet based machine learning methods. Ocean Eng. 233, 109116. https://doi.org/10.1016/j.oceaneng.2021.109116.

Ayinde, A.S., Yu, H., Wu, K., 2023. Sea level variability and modeling in the Gulf of Guinea using supervised machine learning. Sci. Rep. 13 (1), 21318.

Bai, T., Tahmasebi, P., 2023. Graph neural network for groundwater level forecasting. J. Hydrol. 616, 128792. https://doi.org/10.1016/j.jhydrol.2022.128792.

Balogun, A.L., Adebisi, N., 2021. Sea level prediction using ARIMA, SVR and LSTM neural network: assessing the impact of ensemble ocean-atmospheric processes on models' accuracy. Geomat, Nat. Hazard. Risk 12, 653–674. https://doi.org/10.1080/19475705.2021.1887372.

Baran, K., Neumann, T., 2023. A comparative analysis of seaports in terms of the development of maritime tourism in the area of the Baltic sea. Water 15, 3721. https://doi.org/10.3390/w15213721.

Bellinghausen, K., Hünicke, B., Zorita, E., 2023. Short-term prediction of extreme sea-level at the Baltic Sea coast by Random Forests. Nat. Hazards Earth Syst. Sci. Discuss. 2023, 1–48.

Bengio, Y., Frasconi, P., Simard, P., 1993. The problem of learning long-term dependencies in recurrent networks. In: IEEE international conference on neural networks. IEEE, pp. 1183–1188. https://doi.org/10.1109/ICNN.1993.298725.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5, 157–166. https://doi.org/10.1109/72.279181.

Braakmann-Folgmann, A., Roscher, R., Wenzel, S., Uebbing, B., Kusche, J., 2017. Sea level anomaly prediction using recurrent neural networks. In: In: proceedings of the 2017 Conference on Big Data from Space.

Chen, Y., Wang, Y., Dong, Z., Su, J., Han, Z., Zhou, D., Zhao, Y., Bao, Y., 2021. 2-D regional short-term wind speed forecast based on CNN-LSTM deep learning model. Energy Convers. Manag. 244, 114451. https://doi.org/10.1016/j.enconman.2021.114451.

Chia, M.Y., Huang, Y.F., Koo, C.H., Ng, J.L., Ahmed, A.N., El-Shafie, A., 2022. Long-term forecasting of monthly mean reference evapotranspiration using deep neural network: a comparison of training strategies and approaches. Appl. Soft Comput. 126, 109221. https://doi.org/10.1016/j.asoc.2022.109221.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

De Biasio, F., Baldin, G., Vignudelli, S., 2020. Revisiting vertical land motion and sea level trends in the northeastern Adriatic sea using satellite altimetry and tide gauge data. J. Mar. Sci. Eng. 8, 949. https://doi.org/10.3390/jmse8110949.

Delpeche-Ellmann, N., Giudici, A., Rätsep, M., Soomere, T., 2021. Observations of surface drift and effects induced by wind and surface waves in the Baltic Sea for the period 2011–2018. Estuar. Coast Shelf Sci. 249, 107071. https://doi.org/10.1016/j.ecss.2020.107071.

Delpeche-Ellmann, N., Mingelaitė, T., Soomere, T., 2017. Examining lagrangian surface transport during a coastal upwelling in the Gulf of Finland, Baltic sea. J. Mar. Syst. 171, 21–30. https://doi.org/10.1016/j.jmarsys.2016.10.007.

Delpeche-Ellmann, N., Torsvik, T., Soomere, T., 2016. A comparison of the motions of surface drifters with offshore wind properties in the Gulf of Finland, the Baltic Sea. Estuar. Coast Shelf Sci. 172, 154–164. https://doi.org/10.1016/j.ecss.2016.02.009.

Elman, J., 1990. Finding structure in time. Cogn. Sci. 14, 179–211. https://doi.org/10.1016/0364-0213(90)90002-E.

Espinosa, R., Palma, J., Jiménez, F., Kamińska, J., Sciavicco, G., Lucena-Sánchez, E., 2021. A time series forecasting based multi-criteria methodology for air quality prediction. Appl. Soft Comput. 113, 107850. https://doi.org/10.1016/j.asoc.2021.107850.

Fu, Y., Zhou, X., Sun, W., Tang, Q., 2019. Hybrid model combining empirical mode decomposition, singular spectrum analysis, and least squares for satellite-derived sea-level anomaly prediction. Int. J. Remote Sens. 40, 7817–7829. https://doi.org/10.1080/01431161.2019.1606959.

Gao, Z., Liu, X., Yv, F., Wang, J., Xing, C., 2023. Learning wave fields evolution in North West Pacific with deep neural networks. Appl. Ocean Res. 130, 103393. https://doi.org/10.1016/j.apor.2022.103393.

Harter, L., Pineau-Guillou, L., Chapron, B., 2024. Underestimation of extremes in sea level surge reconstruction. Sci. Rep. 14 (1), 14875.

Hawkins, D.M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44 (1), 1–12.

Hazrin, N.A., Chong, K.L., Huang, Y.F., Ahmed, A.N., Ng, J.L., Koo, C.H., Tan, K.W., Sherif, M., El-shafie, A., 2023. Predicting sea levels using ML algorithms in selected locations along coastal Malaysia. Heliyon 9, e19426. https://doi.org/10.1016/j.heliyon.2023.e19426.

He, X., Shi, S., Geng, X., Yu, J., Xu, L., 2023. Multi-step forecasting of multivariate time series using multi-attention collaborative network. Expert Syst. Appl. 211, 118516. https://doi.org/10.1016/j.eswa.2022.118516.

Helcom, 2021. Input of nutrients by the seven biggest rivers in the Baltic Sea region 1995–2017. In: Baltic Sea Environment Proceedings, 178.

Hordoir, R., Axell, L., Höglund, A., Dieterich, C., Fransner, F., Gröger, M., Liu, Y., Pemberton, P., Schimanke, S., Andersson, H., Ljungemyr, P., Nygren, P., Falahat, S., Nord, A., Jönsson, A., Lake, I., Döös, K., Hieronymus, M., Dietze, H., Löptien, U., Kuznetsov, I., Westerlund, A., Tuomi, L., Haapala, J., 2019. Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas – research and operational applications. Geosci. Model. Dev. 12, 363–386. https://doi.org/10.5194/gmd-12-363-2019.

Imani, M., Kao, H.-C., Lan, W.-H., Kuo, C.-Y., 2018. Daily sea level prediction at Chiayi coast, Taiwan using extreme learning machine and relevance vector machine. Glob. Planet Change 161, 211–221. https://doi.org/10.1016/j.gloplacha.2017.12.018.

Imani, M., You, R.-J., Kuo, C.-Y., 2014. Forecasting Caspian Sea level changes using satellite altimetry data (June 1992–December 2013) based on evolutionary support vector regression algorithms and gene expression programming. Glob. Planet Change 121, 53–63. https://doi.org/10.1016/j.gloplacha.2014.07.002.

Intergovernmental Panel on Climate Change (IPCC), 2023. Climate Change 2022 - Mitigation of Climate Change. Cambridge University Press. https://doi.org/10.1017/9781009157926.

Ishida, K., Tsujimoto, G., Ercan, A., Tu, T., Kiyama, M., Amagasaki, M., 2020. Hourly-scale coastal sea level modeling in a changing climate using long short-term memory neural network. Sci. Total Environ. 720, 137613. https://doi.org/10.1016/j.scitotenv.2020.137613.

Jahanmard, V., Delpeche-Ellmann, N., Ellmann, A., 2022. Towards realistic dynamic topography from coast to offshore by incorporating hydrodynamic and geoid models. Ocean Model. 180, 102124. https://doi.org/10.1016/j.ocemod.2022.102124.

Jahanmard, V., Delpeche-Ellmann, N., Ellmann, A., 2021. Realistic dynamic topography through coupling geoid and hydrodynamic models of the Baltic Sea. Cont. Shelf Res. 222, 104421. https://doi.org/10.1016/j.csr.2021.104421.
Jahanmard, V., Delpeche-Ellmann, N., Ellmann, A., 2023b. Absolute Dynamic Topography: Corrected Nemo-Nordic Model for the Baltic Sea. SEANOE. https://doi.org/10.17882/96784.
Jahanmard, V., Hordoir, R., Delpeche-Ellmann, N., Ellmann, A., 2023a. Quantification of hydrodynamic model sea level bias utilizing deep learning and synergistic integration of data sources. Ocean Model. 186, 102286. https://doi.org/10.1016/j.ocemod.2023.102286.
Jiang, W., Liu, B., Liang, Y., Gao, H., Lin, P., Zhang, D., Hu, G., 2024. Applicability analysis of transformer to wind speed forecasting by a novel deep learning framework with multiple atmospheric variables. Appl. Energy 353, 122155. https://doi.org/10.1016/j.apenergy.2023.122155.
Jin, H., Zhong, R., Liu, M., Ye, C., Chen, X., 2023. Using EEMD mode decomposition in combination with machine learning models to improve the accuracy of monthly sea level predictions in the coastal area of China. Dyn. Atmosp. Oceans 102, 101370. https://doi.org/10.1016/j.dynatmoce.2023.101370.
Jung, A., 2022. Machine Learning, Machine Learning: Foundations, Methodologies, and Applications. Springer Nature Singapore, Singapore. https://doi.org/10.1007/978-981-16-8193-6.
Karimi, S., Kisi, O., Shiri, J., Makarynskyy, O., 2013. Neuro-fuzzy and neural network techniques for forecasting sea level in Darwin Harbor, Australia. Comput. Geosci. 52, 50–59. https://doi.org/10.1016/j.cageo.2012.09.015.
Keevallik, S., Soomere, T., 2003. Directional and extreme wind properties in the Gulf of Finland. Proc. Estonian Acad. Sci. Eng. 9, 73. https://doi.org/10.3176/eng.2003.2.01.
Kim, S., Pan, S., Mase, H., 2019. Artificial neural network-based storm surge forecast model: practical application to Sakai Minato, Japan. Appl. Ocean Res. 91, 101871. https://doi.org/10.1016/j.apor.2019.101871.
Kowalewski, M., Kowalewska-Kalkowska, H., 2017. Sensitivity of the Baltic Sea level prediction to spatial model resolution. J. Mar. Syst. 173, 101–113. https://doi.org/10.1016/j.jmarsys.2017.05.001.
Kulikov, E.A., Medvedev, I.P., 2013. Variability of the Baltic Sea level and floods in the Gulf of Finland. Oceanology 53, 145–151.
Kulikov, E.A., Medvedev, I.P., Koltermann, K.P., 2015. Baltic sea level low-frequency variability. Tellus A: Dyn. Meteorol. Oceanogr. 67, 25642. https://doi.org/10.3402/tellusa.v67.25642.
Leppäranta, M., Myrberg, K., 2009. Physical Oceanography of the Baltic Sea. Springer.
Li, X., Cao, J., Guo, J., Liu, C., Wang, W., Jia, Z., Su, T., 2022. Multi-step forecasting of ocean wave height using gate recurrent unit networks with multivariate time series. Ocean Eng. 248, 110689. https://doi.org/10.1016/j.oceaneng.2022.110689.
Li, X., Zhou, S., Wang, F., Fu, L., 2024. An improved sparrow search algorithm and CNN-BiLSTM neural network for predicting sea level height. Sci. Rep. 14, 4560. https://doi.org/10.1038/s41598-024-55266-4.
Liu, J., Jin, B., Wang, L., Xu, L., 2022. Sea surface height prediction with deep learning based on attention mechanism. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/LGRS.2020.3039062.
Man, Y., Yang, Q., Shao, J., Wang, G., Bai, L., Xue, Y., 2023. Enhanced LSTM model for daily runoff prediction in the Upper Huai River Basin, China. Engineering 24, 229–238. https://doi.org/10.1016/j.eng.2021.12.022.
Mao, W., Zhu, H., Wu, H., Lu, Y., Wang, H., 2023. Forecasting and trading credit default swap indices using a deep learning model integrating Merton and LSTMs. Expert Syst. Appl. 213, 119012. https://doi.org/10.1016/j.eswa.2022.119012.
Medvedev, I.P., Kulikov, E.A., 2021. Extreme storm surges in the Gulf of Finland: frequency-spectral properties and the influence of low-frequency sea level oscillations. Oceanology 61, 459–468.
Morrow, R., Fu, L.-L., Rio, M.-H., Ray, R., Prandi, P., Le Traon, P.-Y., Benveniste, J., 2023. Ocean circulation from space. Surv. Geophys. 44, 1243–1286. https://doi.org/10.1007/s10712-023-09778-9.
Mostafavi, M., Delpeche-Ellmann, N., Ellmann, A., 2021. Accurate Sea surface heights from Sentinel-3A and Jason-3 retrackers by incorporating high-resolution marine geoid and hydrodynamic models. J. Geodet. Sci. 11, 58–74. https://doi.org/10.1515/jogs-2020-0120.
Mostafavi, M., Delpeche-Ellmann, N., Ellmann, A., Jahanmard, V., 2023. Determination of accurate dynamic topography for the Baltic Sea using satellite altimetry and a marine geoid model. Remote Sens. 15, 2189. https://doi.org/10.3390/rs15082189.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pellikka, H., Leijala, U., Johansson, M.M., Leinonen, K., Kahma, K.K., 2018. Future probabilities of coastal floods in Finland. Cont. Shelf Res. 157, 32–42. https://doi.org/10.1016/j.csr.2018.02.006.
Pindsoo, K., Soomere, T., 2020. Basin-wide variations in trends in water level maxima in the Baltic Sea. Cont. Shelf Res. 193, 104029. https://doi.org/10.1016/j.csr.2019.104029.
Primo de Siqueira, B.V., Paiva, A., de, M., 2021. Using neural network to improve sea level prediction along the southeastern Brazilian coast. Ocean Model. 168, 101898. https://doi.org/10.1016/j.ocemod.2021.101898.
Qin, Y., Su, C., Chu, D., Zhang, J., Song, J., 2023. A review of application of machine learning in storm surge problems. J. Mar. Sci. Eng. 11, 1729. https://doi.org/10.3390/jmse11091729.
Raj, N., Gharineiat, Z., Ahmed, A.A.M., Stepanyants, Y., 2022. Assessment and prediction of Sea level trend in the South Pacific region. Remote Sens. 14, 986. https://doi.org/10.3390/rs14040986.
Rajabi-Kiasari, S., Delpeche-Ellmann, N., Ellmann, A., 2023. Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea. Comput. Geosci. 178, 105406. https://doi.org/10.1016/j.cageo.2023.105406.
Rajabi-Kiasari, S., Hasanlou, M., 2020. An efficient model for the prediction of SMAP sea surface salinity using machine learning approaches in the Persian Gulf. Int. J. Remote Sens. 41, 3221–3242. https://doi.org/10.1080/01431161.2019.1701212.
Ramos-Valle, A.N., Curchitser, E.N., Bruyère, C.L., McOwen, S., 2021. Implementation of an artificial neural network for storm surge forecasting. J. Geophys. Res. Atmos. 126 (13), e2020JD033266. https://doi.org/10.1029/2020JD033266.
Rosentau, A., Klemann, V., Bennike, O., Steffen, H., Wehr, J., Latinović, M., Bagge, M., Ojala, A., Berglund, M., Becher, G.P., Schoning, K., Hansson, A., Nielsen, L., Clemmensen, L.B., Hede, M.U., Kroon, A., Pejrup, M., Sander, L., Stattegger, K., Schwarzer, K., Lampe, R., Lampe, M., Uścinowicz, S., Bitinas, A., Grudzinska, I., Vassiljev, J., Nirgi, T., Kublitskiy, Y., Subetto, D., 2021. A Holocene relative sea-level database for the Baltic Sea. Quat. Sci. Rev. 266, 107071. https://doi.org/10.1016/j.quascirev.2021.107071.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536. https://doi.org/10.1038/323533a0.
Sabililah, R.N., Adytia, D., 2023. Time series forecasting of sea level by using transformer approach, with a case study in Pangandaran, Indonesia. In: In: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE, pp. 1–6. https://doi.org/10.1109/I2CT57861.2023.10126216.
Sarveswararao, V., Ravi, V., Vivek, Y., 2023. ATM cash demand forecasting in an Indian bank with chaos and hybrid deep learning networks. Expert Syst. Appl. 211, 118645. https://doi.org/10.1016/j.eswa.2022.118645.
Skriptunov, N.A., Gorelits, O.V., 2001. Wind-induced variations in water level in river mouths. Water Resourc. 28 (2), 174–179.
Slobbe, D.C., Klees, R., Gunter, B.C., 2014. Realization of a consistent set of vertical reference surfaces in coastal areas. J. Geodesy 88, 601–615. https://doi.org/10.1007/s00190-014-0709-9.
Song, C., Chen, X., Xia, W., Ding, X., Xu, C., 2022. Application of a novel signal decomposition prediction model in minute sea level prediction. Ocean Eng. 260, 111961. https://doi.org/10.1016/j.oceaneng.2022.111961.
Song, T., Han, N., Zhu, Y., Li, Z., Li, Y., Li, S., Peng, S., 2021. Application of deep learning technique to the sea surface height prediction in the South China Sea. Acta Oceanol. Sin. 40, 68–76. https://doi.org/10.1007/s13131-021-1735-0.
Soomere, T., 2003. Anisotropy of wind and wave regimes in the Baltic proper. J. Sea Res. 49, 305–316. https://doi.org/10.1016/S1385-1101(03)00034-0.
Soomere, T., Leppäranta, M., Myrberg, K., 2009. Highlights of the physical oceanography of the Gulf of Finland reflecting potential climate change. Boreal Environ. Res. 14, 152–165.
Soomere, T., Myrberg, K., Leppäranta, M., Nekrasov, A., 2008. The progress in knowledge of physical oceanography of the Gulf of Finland: a review for 1997-2007. Oceanologia 50, 287–362.
Sukhachev, V.N., Zakharchuk, E.A., Tikhonova, N.A., 2014. On the mechanisms of dangerous sea level rise in the eastern part gulf of Finland and possible reasons for the increase in their frequency in the second half of XX and the beginning of the XXI century. In: In: 2014 IEEE/OES Baltic International Symposium (BALTIC). IEEE, pp. 1–12. https://doi.org/10.1109/BALTIC.2014.6887843.
Sun, C., He, Z., Lin, H., Cai, L., Cai, H., Gao, M., 2023. Anomaly detection of power battery pack using gated recurrent units based variational autoencoder. Appl. Soft Comput. 132, 109903. https://doi.org/10.1016/j.asoc.2022.109903.
Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S.M., Wang, Jinbo, Lin, C., Cristea, N., Tong, D., Carande, W.H., Ma, X., Rao, Y., Bednar, J.A., Tan, A., Wang, Jianwu, Purushotham, S., Gill, T.E., Chastang, J., Howard, D., Holt, B., Gangodagamage, C., Zhao, P., Rivas, P., Chester, Z., Orduz, J., John, A., 2022. A review of earth artificial intelligence. Comput. Geosci. 159, 105034. https://doi.org/10.1016/j.cageo.2022.105034.
Suursaar, Ü., Sooäär, J., 2007. Decadal variations in mean and extreme sea level values along the Estonian coast of the Baltic Sea. Tellus A: Dyn. Meteorol. Oceanogr. 59, 249. https://doi.org/10.1111/j.1600-0870.2006.00220.x.
Sztobryn, M., 2003. Forecast of storm surge by means of artificial neural network. J. Sea Res. 49 (4), 317–322.
Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., Ward, P.J., 2021. Exploring deep learning capabilities for surge predictions in coastal areas. Sci. Rep. 11, 17224. https://doi.org/10.1038/s41598-021-96674-0.
Tsimplis, M.N., Woodworth, P.L., 1994. The global distribution of the seasonal sea level cycle calculated from coastal tide gauge data. J. Geophys. Res. Oceans 99 (C8), 16031–16039.
Varbla, S., Ågren, J., Ellmann, A., Poutanen, M., 2022. Treatment of tide gauge time series and marine GNSS measurements for vertical land motion with relevance to the implementation of the Baltic Sea Chart datum 2000. Remote Sens. 14, 920. https://doi.org/10.3390/rs14040920.
Vuolio, T., Visuri, V.-V., Sorsa, A., Ollila, S., Fabritius, T., 2020. Application of a genetic algorithm based model selection algorithm for identification of carbide-based hot metal desulfurization. Appl. Soft Comput. 92, 106330. https://doi.org/10.1016/j.asoc.2020.106330.
Wang, Bao, Wang, Bin, Wu, W., Xi, C., Wang, J., 2020. Sea-water-level prediction via combined wavelet decomposition, neuro-fuzzy and neural networks using SLA and wind information. Acta Oceanol. Sin. 39, 157–167. https://doi.org/10.1007/s13131-020-1569-1.
Wang, Bao, Liu, S., Wang, Bin, Wu, W., Wang, J., Shen, D., 2021. Multi-step ahead short-term predictions of storm surge level using CNN and LSTM network. Acta Oceanologica Sinica 40, 104–118. https://doi.org/10.1007/s13131-021-1763-9.

Wang, G., Wang, X., Wu, X., Liu, K., Qi, Y., Sun, C., Fu, H., 2022. A hybrid multivariate deep learning network for Multistep ahead sea level anomaly forecasting. J. Atmos. Ocean Technol. 39, 285–301. https://doi.org/10.1175/JTECH-D-21-0043.1.

Weisse, R., Dailidienė, I., Hünicke, B., Kahma, K., Madsen, K., Omstedt, A., Parnell, K., Schöne, T., Soomere, T., Zhang, W., Zorita, E., 2021. Sea level dynamics and coastal erosion in the Baltic Sea region. Earth Syst. Dyn. 12, 871–898. https://doi.org/10.5194/esd-12-871-2021.

Werbos, P.J., 1988. Generalization of backpropagation with application to a recurrent gas market model. Neural Netw. 1, 339–356. https://doi.org/10.1016/0893-6080(88)90007-X.

Xiao, C., Chen, N., Hu, C., Wang, K., Xu, Z., Cai, Y., Xu, L., Chen, Z., Gong, J., 2019. A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data. Environ. Model. Softw. 120, 104502. https://doi.org/10.1016/j.envsoft.2019.104502.

Yu, H., Yang, L., Feng, Q., Barzegar, R., Adamowski, J.F., Wen, X., 2024. Ensemble learning of decomposition-based machine learning models for multistep-ahead daily streamflow forecasting in northwest China. Hydrol. Sci. J. 69 (11), 1501–1522.

Zhang, Z., Stanev, E.V., Grayek, S., 2020. Reconstruction of the basin-wide sea-level variability in the north sea using coastal data and generative adversarial networks. J. Geophys. Res. Oceans 125 (12), e2020JC016402.

Zhao, J., Cai, R., Sun, W., 2021. Regional sea level changes prediction integrated with singular spectrum analysis and long-short-term memory network. Adv. Space Res. 68, 4534–4543. https://doi.org/10.1016/j.asr.2021.08.017.

Zhou, Y., Lu, C., Chen, K., Li, X., 2022. Multilayer fusion recurrent neural network for sea surface height anomaly field prediction. IEEE Trans. Geosci. Remote Sens. 60, 1–11. https://doi.org/10.1109/TGRS.2021.3126460.

Zilong, T., Yubing, S., Xiaowei, D., 2022. Spatial-temporal wave height forecast using deep learning and public reanalysis dataset. Appl. Energy 326, 120027. https://doi.org/10.1016/j.apenergy.2022.120027.

Zou, Y., Wang, J., Lei, P., Li, Y., 2023. A novel multi-step ahead forecasting model for flood based on time residual LSTM. J. Hydrol. 620, 129521. https://doi.org/10.1016/j.jhydrol.2023.129521.

Appendix 3

Publication III

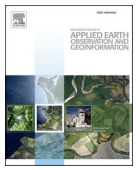
Rajabi-Kiasari, S., Delpeche-Ellmann, N., Ellmann, A., & Soomere, T. (2026). Forecasting sea level maxima using ML with explainability and extreme value analysis. International Journal of Applied Earth Observation and Geoinformation, 146, 105064, doi: doi.org/10.1016/j.jag.2025.105064.

International Journal of Applied Earth Observation and Geoinformation 146 (2026) 105064

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Forecasting sea level maxima using Machine learning with explainability and extreme value analysis☆

Saeed Rajabi-Kiasari [a,*], Nicole Delpeche-Ellmann [b], Artu Ellmann [a], Tarmo Soomere [b]

[a] Department of Civil Engineering and Architecture, School of Engineering, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
[b] Department of Cybernetics, School of Science, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia

ARTICLE INFO

Keywords:
Extreme sea-level prediction
Sea level forecasting
Baltic Sea
Extreme value analysis
Machine Learning
Explainable machine learning

ABSTRACT

Accurately forecasting sea level maxima (SLM) and extremes, is vital for maritime management, engineering, and navigation. Most Machine learning (ML) models focus on moderate surges and often underestimate extremes. We use a two-fold forecasting framework: ML/DL for short-term daily SLM forecasting and extreme value theory (EVT) for long-term extremes (<100 years). ML models include Random Forest, Extreme gradient boosting, and Multilayer perceptron, CNN-LSTM and CNN-GRU. Long-term extremes are analysed via EVT using a block maximum method. The Baltic Sea, a semi-enclosed micro-tidal basin prone to extremes, serves as case study. The analysis used six tide gauge stations (Narva, Ristna, Oulu, Kungsholmsfort, Władysławowo, and Greifswald). Key features—wind speed, surface air pressure, Baltic Sea Index (BSI), and significant wave height (SWH), were selected using a mutual information and models' hyperparameters tuned using Bayesian optimization.

Neural networks models, specifically the CNN-GRU and MLP, performed best (RMSE 7–15 cm) with strong generalization. Most models captured storm events, but underestimated extreme peaks (>150 cm), due to the rarity in the training, incomplete meteorological representation, and missing local physical processes. CNN-GRU excelled in RMSE, recall, and F1, while MLP led in R^2 and precision. EVT analysis showed winter extremes have ~ 5–7 years in the north-east (Narva and Oulu). Explainability analysis of CNN-GRU showed prefilling dominates SLM at all stations; BSI, pressure, and winds drive west, south, and north, while local pressure, wind, and SWH dominate in the east. The framework supports early warning and long-term risk assessment, though forecasting rare extremes remains challenging.

1. Introduction

Monitoring and forecasting sea level maxima (SLM) and extremes, both the short and long term, is crucial, as high-water events threats coastal communities through flooding, erosion, and loss of human life (Haigh et al., 2014). SLM are often sudden (minutes to days), site-specific, and primarily driven by strong storms but can be amplified by factors like pressure gradients, winds, waves, tides, coastal topography (Grossman et al., 2023), or local prefilling (Weisse et al., 2021).

Forecasting these events is challenging due to their rarity and the non-linear interplay of several multiple factors. Long-term datasets (multi-decadal to centennial) are needed to capture statistical behaviour and linking SLM to global warming adds complexity (Eelsalu et al., 2025). Historical records from tide gauges, satellite altimetry and hydrodynamic model hindcasts often have limited spatial coverage, temporal resolution and contain gaps. High-resolution numerical models provide detailed insights but are computationally expensive, sensitive to forcing and the parametrization (Lorenz and Gräwe, 2023). In contrast, Statistical and machine learning (ML) approaches offer lower computational costs (Harter et al., 2024) but are often limited by data timespan (Kirezci et al., 2020).

In semi-enclosed seas, accurate SLM forecasting is critical, as impacts can exceed those in open ocean. The Baltic Sea, bordered by nine countries and home to 15 million inhabitants within 10 km of the coast, is highly sensitive (HELCOM, 2021) due to intense maritime and coastal activities. Historically, sea levels in the Baltic have exceeded 3 m above the national datum (Averkiev and Klevannyy, 2010), with the highest surge (4.21 m above the so-called Kronstadt zero) on November 19, 1824 in the Gulf of Finland (Wolski and Wiśniewski, 2021). SLM in Baltic result from local storms, basin-wide water volume changes

☆ This article is part of a special issue entitled: 'ESL & Coastal Hazards' published in International Journal of Applied Earth Observation and Geoinformation.

* Corresponding author.
E-mail address: saeed.rajabi@taltech.ee (S. Rajabi-Kiasari).

https://doi.org/10.1016/j.jag.2025.105064
Received 14 February 2025; Received in revised form 25 December 2025; Accepted 25 December 2025
Available online 6 January 2026

(prefilling), seiches, wave-driven local effects and large-scale atmospheric conditions like winds, cyclones and North Atlantic Oscillation (Weisse et al., 2021).

Traditional statistical approaches such as Generalized Extreme Value (GEV), Gumbel and Weibull distributions for extreme water levels (Eelsalu et al., 2014) have been used to predict return periods of extreme sea levels; they are suitable for long-term planning but insufficient for short-term forecasting. Data assimilation and ML methods are often applied, though data assimilation is computationally expensive and sensitive to model forcing; many hydrodynamic models also omit surface waves, which can influence extremes (Harter et al., 2024).

Data-driven Machine and Deep Learning (ML/DL) models have been developed for extreme sea level predictions. First group uses storm characteristics (location, central pressure, the radius of maximum winds, wind direction near the storm centre) with learning methods like Artificial Neural Networks (ANN) (Ramos-Valle et al., 2021), Convolutional Neural Networks (CNN) (Park et al., 2022; Xie et al., 2023), or Generative Adversarial Network (GAN) (Mulia et al., 2023). Ramos-Valle et al. (2021) showed that ANNs can predict moderate surge levels accurately but tend to underestimate peak values. CNN-based approaches applied to historical (Lee et al., 2021) or simulated typhoon data (Xie et al., 2023), have achieved robust regional surge forecasts, capturing non-linear storm surge processes (Hashemi et al., 2016), though they require large datasets and are sensitive to storm parameter changes (Lee et al., 2021).

The second group uses multivariate time series forecasting with meteorological features. Di Nunno et al. (2021) applied Nonlinear Autoregressive Exogenous (NARX) neural networks to predict extreme surges in Venice. Regression and ANN models reconstructed surges in the Danish Straits and globally (Dubois et al., 2024; Bruneau et al., 2020) often struggling with peaks. Other studies also employed tree boosting systems (XGBoost or XGB) (Pachev et al., 2023) and Ensemble ML models (Sun and Pan, 2023). LSTM models effectively capture temporal dependencies, demonstrated regional and global capabilities (Ishida et al., 2020; Nagaraj et al., 2025; Sreeraj et al., 2025) and outperform ANN, CNN, and ConvLSTM architectures (Tiggeloven et al., 2021).

Recent advancements in developing hybrid models like CNN-LSTM model enhances forecasting accuracy by combining CNNs' ability to capture short-term patterns with recurrent-based models' strength in learning long-term dependencies (Yang et al., 2021). Applications in China and U.S. showed 4–10 % higher accuracy than standalone CNN/LSTM models or SVR/MLP models (Wang et al., 2021; (Wei and Nguyen, 2022)). Remaining challenges include underestimating peak SLM, limited training data, uncertainty in key drivers, and sensitivity to model choice and parameters.

In the Baltic Sea, despite numerous studies on SLM (Andrée et al., 2023; Pindsoo and Soomere, 2020; Rutgersson and Kjellström, 2022), prediction capabilities remain limited and region specific studies are scarce. Moreover, most models are regional or global (Bruneau et al., 2020; Tiggeloven et al., 2021) and may not capture the complex interactions (Gräwe and Burchard, 2012; Lorenz and Gräwe, 2023). Furthermore, its shallow depth and irregular shape (Leppäranta and Myrberg, 2009), the preconditioning effects (Andrée et al., 2023), and bi-directional strong winds (Soomere et al., 2024) adds complexity to its dynamics, requiring specialized models.

While extreme event classification has improved (Bellinghausen et al., 2025), forecasting the time series and peak intensities remains challenging (Qin et al., 2023). ANN models at Polish coast of the Baltic Sea showed underestimation of peaks (Sztobryn, 2003). Rajabi-Kiasari et al. (2023) developed a CNN model for day-ahead sea level forecasting of the Baltic Sea, achieving ~ 73 % accuracy in detecting maxima (>70 cm) events but underestimated higher levels (> 100 cm) in the Gulf of Riga and Gulf of Finland. LSTM and GRU models applied with high-resolution (hourly; spatially – one nautical mile) sea-level forecasts in the Gulf of Finland (Rajabi-Kiasari et al., 2025). While their model performed well in capturing moderate sea levels, it still underestimated extremes, particularly in the eastern Gulf of Finland.

Moreover, previous studies often used a single model type across tide gauges or grids, limiting performance evaluation, however, the performance of various models needs to be evaluated. Additionally, many time series models prioritize normal sea levels, underestimating peaks, due to lack of adequate extreme events in the training data (Qin et al., 2023; Ramos-Valle et al., 2021; Watson, 2022). Balanced training sets and incorporating diverse SLM drivers, such as storm surges and high winds, can improve extreme forecasts (Bruneau et al., 2020).

Hyperparameter selection and model optimization are also crucial for improving generalization and capturing complex peak patterns (Li et al., 2024). Various techniques are available to determine the best hyperparameter values for ML models. In this study, we use Bayesian optimization (BO) for hyperparameter tuning, which effectively tunes multiple hyperparameters (Dai et al., 2023). ML/DL models are often black boxes (Karpatne et al., 2019; Lamberti, 2023). The explainability can be enhanced using, e.g., the Shapley Additive Explanations (SHAP) method (Lundberg and Lee, 2017), as utilized in this study. The key benefit of SHAP values is their capability to illustrate the impact of features on each sample (Fan et al., 2024).

Despite advances, ML-based sea level models face limitations, particularly in complex regions like the Baltic Sea, often underestimating surges above 100 cm. Improving accuracy requires diverse input features, multiple model evaluation, and explainability analyses to address the “black box” nature of models. This study proposes hybrid CNN-LSTM and CNN-GRU models with Bayesian optimization for hyperparameter tuning to capture the Baltic's nonlinear sea level dynamics. Comparative analyses include baseline models (XGB, RF, and MLP). We perform comprehensive exploratory data analysis, feature selection, optimizations, and evaluation through multiple statistical metrics, including storm event detection. SHAP is applied to quantify the variable contribution, while Extreme Value Theory (EVT) corrects systematic underestimation of extreme peaks, supporting risk assessment. To our knowledge, this is the first study integrating Bayesian Optimization with hybrid CNN-LSTM and CNN-GRU models for sea level forecasting, combined with a transparent, explainable framework and comparative framework for daily SLM forecasting. The analysis uses 1971–2022 data, incorporating meteorological factors such as pressure, wind speed and direction, wind gust, significant wave height, river runoff, precipitation, evaporation, and BSI. The main objectives are to:

(i) Characterize SLM (extremes) by magnitude, frequency, and duration;
(ii) Identify the most effective ML/DL models for short-term one-day-ahead forecasting;
(iii) Compare models for peaks/storm event detection;
(iv) Determine the key contributors via Explainable ML;
(v) Estimate return levels and periods of extreme sea levels for long-term prediction (<100 years).

The paper is structured as follows. Section 2 describes the proposed methodology and theoretical explanation of the applied models. Section 3 explains the data sets used and the study area specifications. Section 4 discusses the preprocessing and models' configurations. The results of this study are explained in Section 5. The discussion is presented in Section 6, while the conclusions and future research directions are provided in Section 7.

2. Background and methodology

2.1. Problem Definition

SLM forecasting encompasses multiple stages, accounting for dynamically changing variables across space and time. Accurate forecasts require linking historical sea level observations with climatic and

geographic drivers. Given a multivariate time series dataset $X_t = \left[X_{t,1}^{(s)}, X_{t,2}^{(s)}, \cdots, X_{t,D}^{(s)}\right]$ for the station s, the forecast for time $t+1$ is:

$$\widehat{X}_{t+1}^{(s)} = f\left(X_{t-L+1}^{(s,1)}, X_{t-L+2}^{(s,1)}, \cdots, X_t^{(s,1)}, X_{t-L+1}^{(s,2)}, X_{t-L+2}^{(s,2)}, \cdots, X_t^{(s,2)}, \cdots, X_{t-L+1}^{(s,D)}, X_{t-L+2}^{(s,D)}, \cdots, X_t^{(s,D)}\right) \quad (1)$$

where L is the lag window, D the number of features, and f the forecasting model.

The workflow (displayed in Fig. 1) includes:

1. Selecting relevant features (surface atmospheric pressure, wind speed, SWH, river runoff, precipitation, evaporation, BSI); filling gaps using a linear interpolation technique; performing feature selection using nonlinear mutual information (MI).
2. Normalizing and structuring the data for time series forecasting.
3. Developing three ML and two DL models optimized through Bayesian optimization.
4. Evaluating models for time series forecasting, peaks/storm detection, computing underestimation levels by return periods, and explainability to determine feature contributions.

2.2. ML and DL models

ML is a subfield of artificial intelligence, uses algorithms to identifying patterns in data for tasks like regression and classification. Core components include input features, a model (hypothesis), a loss function to measure errors and optimization of hyperparameters to minimize loss, with performance evaluated iteratively. DL, a specialized branch of ML, uses multi-layered neural networks to automatically learn features directly from complex nonlinear data, making it powerful for images, text, and time series analysis. Unlike ML, DL requires higher computational demands. In this study, three ML —MLP, RF, and XGB— and two hybrid DL models —CNN-LSTM and CNN-GRU— are employed to forecast SLM. Detailed descriptions of the models, along with their advantages and limitations, and algorithms are provided in the following subsections.

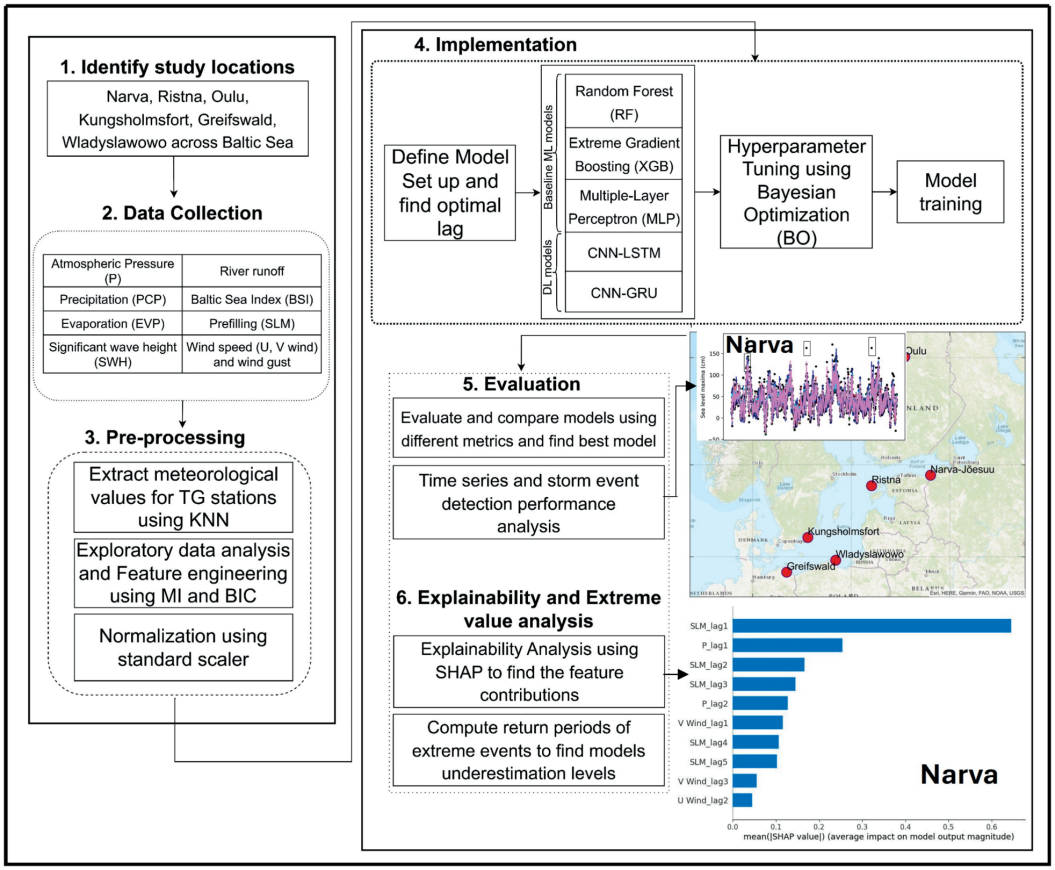


Fig. 1. Overview of the proposed strategy for sea level maxima forecasting using ML/DL models.

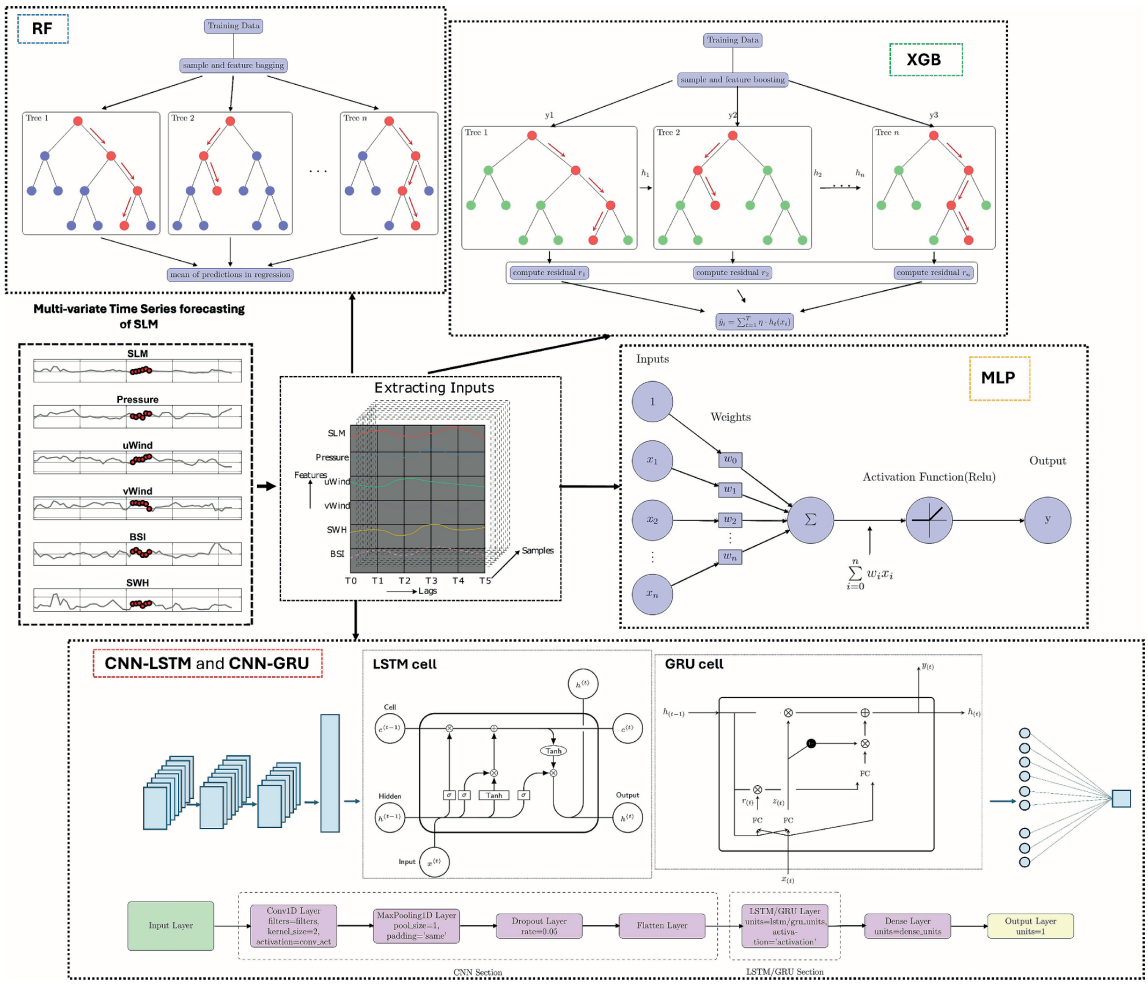


Fig. 2. Workflow of the XGB, RF, MLP, CNN-LSTM and CNN-GRU models' architecture.

2.2.1. Baseline ML methods

Random Forest (RF)

RF (Breiman, 2001) is an ensemble of decision trees built using bagging. It has been applied for binary classification of extreme sea levels (Bellinghausen et al., 2025) and regression (Pachev et al., 2023) tasks. Trees are trained on random subsets of data and features ('mtry') (Fig. 2) to reduce correlation, mitigate overfitting, and improve generalization (Rajabi-Kiasari and Hasanlou, 2020). Each tree is trained on a slightly different dataset, enhancing robustness. Predictions are then aggregated by averaging for regression or majority voting for classification. Key hyperparameters are the number of decision trees and 'mtry'. A detailed description of the RF method can be found in Breiman (2001).

Extreme gradient boosting (XGB)

Extreme gradient boosting (XGB) (Chen and Guestrin, 2016) is a boosting-based ensemble ML method, widely used in applications such as urban floods (Xu et al., 2023), maximum water level increase (Sun and Pan, 2023) and tsunami early warning systems (Rabbani et al., 2023). Boosting—sequentially combines weak learners to form a stronger model—unlike bagging (e.g., RF), where trees operate independently.

XGB adds new decision trees iteratively to minimize a loss function with a regularization term Ω:

$$objective = \sum_{i=1}^{n} L\left[y_i, \hat{y}_i^{t-1} + h_t(x_i) + \Omega(h_t)\right], \tag{2}$$

where, *objective* is the loss function L at iteration t, measuring the difference between actual value y_i and predicted $\hat{y}_i^{t-1} + h_t(x_i)$, Ω is the regularization term, $\hat{y}_i^{t-1}$ is the prediction from the ensemble at iteration $t-1$, h_t is the weak learner (decision tree) added at iteration t, and x_i is the feature vector for the i th sample.

The final prediction is the sum of predictions from all weak learners (Fig. 2) weighted by a learning rate η:

$$\hat{y}_i = \sum_{t=1}^{T} \eta h_t(x_i). \tag{3}$$

where T is the total number of decision trees in the ensemble.

Key hyperparameters mainly include the learning rate, the number of trees, and maximum depth.

Multi-layer perceptron (MLP) neural network

MLP neural networks are a well-known ML model frequently applied for extreme sea level prediction (Bruneau et al., 2020). They consist of an input layer, one or more hidden layers, and an output layer, with neurons connected through weighted links. Each neuron computes a weighted sum of its inputs, adds a bias, and applies an activation function to capture non-linear relationships (Fig. 2). Mathematically, the output of a neuron in a hidden layer can be represented as:

$$z = f\left(\sum_{i=1}^{n} \omega_i x_i + b\right), \tag{4}$$

where x_i represents the input features, n is the number of features, ω_i denotes the weight associated with each input feature, b is the bias term, and f is the activation function, introducing non-linearity into the model. MLPs are trained using algorithms such as backpropagation and gradient descent to adjust the weights and biases to minimize a defined loss function, enabling them to learn complex patterns.

2.2.2. CNN-LSTM model

A hybrid CNN-LSTM model combines CNN's spatial feature extractions and LSTM's temporal sequences modelling, making it effective for multivariate time series forecasting. The CNN layer, which acts as a feature extractor from the input X_t:

$$Y_{l,t} = \sigma\left(\sum_{k=1}^{K_l} W_{l,k} * X_t + b_l\right), \tag{5}$$

The output $Y_{l,t}$ of the l-th convolutional layer at time step t is a feature map that captures local patterns in the input sequence X_t, σ is the activation function applied element-wise (e.g., ReLU, tanh), $W_{l,k}$ represents the k-th filter (kernel) of the l-th convolutional layer, * denotes the convolution operation, X_t is the input vector at time step t, b_l is the bias vector for the l-th convolutional layer, and K_l is the number of filters in the l-th convolutional layer.

The extracted features are passed to the LSTM layer (see Fig. 2), which models temporal dependencies using different gating mechanisms—input (i_t), forget (f_t), and output (o_t) gates—to regulate information flow and capture long-range dependencies. This capability helps mitigate issues like the gradient vanishing problem encountered in simple recurrent neural networks (RNNs), allowing LSTMs to effectively model long-range dependencies. The LSTM equations can be expressed as:

$$f_t = \sigma\left(W_f[H_{t-1}, x_t] + b_f\right) \tag{6}$$

$$i_t = \sigma(W_i[H_{t-1}, x_t] + b_i), \widehat{C}_t = \tanh(W_c[H_{t-1}, x_t] + b_c), \tag{7}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widehat{C}_t, \tag{8}$$

$$o_t = \sigma(W_o[H_{t-1}, x_t] + b_o), H_t = o_t \cdot \tanh(C_t) \tag{9}$$

where x_t is the input data at time t which includes multiple features and their lags, i_t, f_t, and o_t are the input, forget, and output gates, respectively, $\widehat{C}_t$ is the candidate cell state, C_t is the cell state, H_t is the hidden state output at time t, W_i, W_f, W_o, W_c are weight matrices, b_i, b_f, b_o, b_c are bias vectors, and $\cdot$ is element-wise multiplication product.

2.2.3. CNN-GRU model

The hybrid CNN-GRU model combines CNNs for spatial feature extraction with GRUs for temporal modeling, making it suitable for multivariate time series forecasting. It share the same structure as the CNN-LSTM model but replaces the LSTM network with the GRU layer (Fig. 2). GRUs (Cho et al., 2014), simplifies the LSTM by merging the input and forget gates into a single update gate, reducing the model complexity and overfitting risk, while maintaining performance. The GRU operations are described as:

$$z_t = \sigma(W_u[x_t, h_{t-1}] + b_t), \tag{10}$$

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r), \tag{11}$$

$$\widetilde{h}_t = \tanh(W_h[r_t * x_t, h_{t-1}] + b_h), h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t \tag{12}$$

where the terms z_t and r_t represent the outputs of the update and reset gates, respectively, W_u and W_r are the weight matrices associated with these gates at time step t, x_t corresponds to the input at the current time step t, while h_{t-1} denotes the previous hidden state at time step $t-1$. The notation b_* signifies the bias term, and σ refers to the sigmoid activation function. Finally, h_t stands for the updated hidden state at time step t, $\widetilde{h}_t$ represents the candidate memory content at t and * is the convolution operator. The CNN-GRU model architecture (layers and dimensions) is displayed in Table S4 in supplementary material S1.

2.2.4. Extreme value theory

This study employs extreme value theory (EVT) to analyse and predict rare events, such as extreme sea levels. Two common approaches for EVT are the block maxima and peak over threshold (POT). We adopt the block maxima method, which fits a Generalized Extreme Value (GEV) distribution to maxima extracted from long time intervals (e.g., seasonal or yearly blocks), following EVT theory (Coles, 2001; Männikus et al., 2020). The POT (models the exceedances over a threshold using the Generalized Pareto Distribution) is not used in our analysis due to low temporal resolution of the available observations and measurements.

The GEV distribution generalizes the Gumbel, Fréchet, and Weibull families, with the shape parameter ξ controlling tail behaviour ($\xi > 0$ Fréchet, $\xi < 0$ Weibull, and $\xi = 0$ Gumbel distribution). This flexibility makes GEV highly effective for modelling sea level extremes (Okacha et al., 2024), including the Baltic Sea, where high water levels often contain outliers (Soomere et al., 2018). The GEV cumulative distribution function (CDF) is given by (Coles, 2001):

$$F(x; \mu, \sigma, \xi) = \exp\left(-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right), \tag{13}$$

where x is the value for which the CDF is evaluated, μ (location parameter) indicates the mode of the distribution if $\xi = 0$, $\sigma > 0$ (scale parameter) reflects the spread, and ξ (shape parameter) determines the tail behaviour. Seasonal maxima from six stations were fitted using maximum-likelihood estimation and return periods of 2, 5, 10, 15, 20, 50, and 100 years period were derived, providing a reliable basis for long-term risk assessment.

2.2.5. Models' explainability using SHapley Additive exPlanations (SHAP)

SHAP (Lundberg and Lee, 2017) explains ML model predictions by attributing feature importance based on cooperative game theory, accounting for feature's interactions. SHAP values computing the contribution of each feature contribution to the deviation from the mean prediction and are expressed as (Zhang et al., 2023):

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \bigcup \{i\}) - f(S)], \tag{14}$$

where ϕ_i presents the SHAP value for the feature i in model f, N is the total number of features, S represents the subset of features, $f(S \bigcup \{i\})$ is the model's output when the feature i is included in S, $f(S)$ is the model's output when the feature i is excluded from S, and the summation is taken over all possible subsets S of features excluding i. The SHAP modelling was implemented using the 'shap' package of Python 3.11 with 1000 samples, as recommended for robust estimation.

2.2.6. Models' evaluation

Models' performances were assessed using R-squared (R^2) index,

Root Mean Square Error (RMSE), and the correlation coefficient (R), measuring explained variance, prediction error, and linear association between observed and predicted sea levels, respectively.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}\left(\widehat{y}_{(t)} - y_{(t)}\right)^2}{\sum_{t=1}^{n}\left(\widehat{y}_{(t)} - \overline{y}\right)^2}, \overline{y} = \frac{1}{n}\sum_{t=1}^{n} y_{(t)}, \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\widehat{y}_{(t)} - y_{(t)}\right)^2} \tag{16}$$

$$R = \frac{\sum_{t=1}^{n}\left(y_{(t)} - \overline{y}\right)\left(\widehat{y}_{(t)} - \overline{\widehat{y}}\right)}{\sqrt{\sum_{t=1}^{n}\left(y_{(t)} - \overline{y}\right)^2 \sum_{t=1}^{n}\left(\widehat{y}_{(t)} - \overline{\widehat{y}}\right)^2}} \tag{17}$$

Here $y_{(t)}$ and $\widehat{y}_{(t)}$ are the actual and forecasted sea level, and $\overline{y}$ and $\overline{\widehat{y}}$ are the mean of actual and forecasted values at time step t, respectively.

Error dispersion was further evaluated using the Interquartile Range (IQR). The Bayesian Information Criterion (BIC) was used to select the optimal lag length, favoring models with the lowest BIC, defined as:

$$BIC = kln(n) - 2ln(\widehat{L}) \tag{18}$$

where k is the number of variables, n is the number of observations, and $\widehat{L}$ is the maximized value of the likelihood function of the model.

Moreover, to evaluate performance during the extreme events, Precision, Recall, and F1-score, which are well-suited for imbalanced dataset. These metrics quantify correct peak detection (TP), false alarms (FP), and missed events (FN), calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{21}$$

The computed values for the metrics in equations (19–21) across all models, thresholds, and stations are reported in Fig. A1 in Appendix.

3. Study area and data collection

3.1. Study area

The Baltic Sea is a semi-enclosed, micro-tidal estuarine sea in

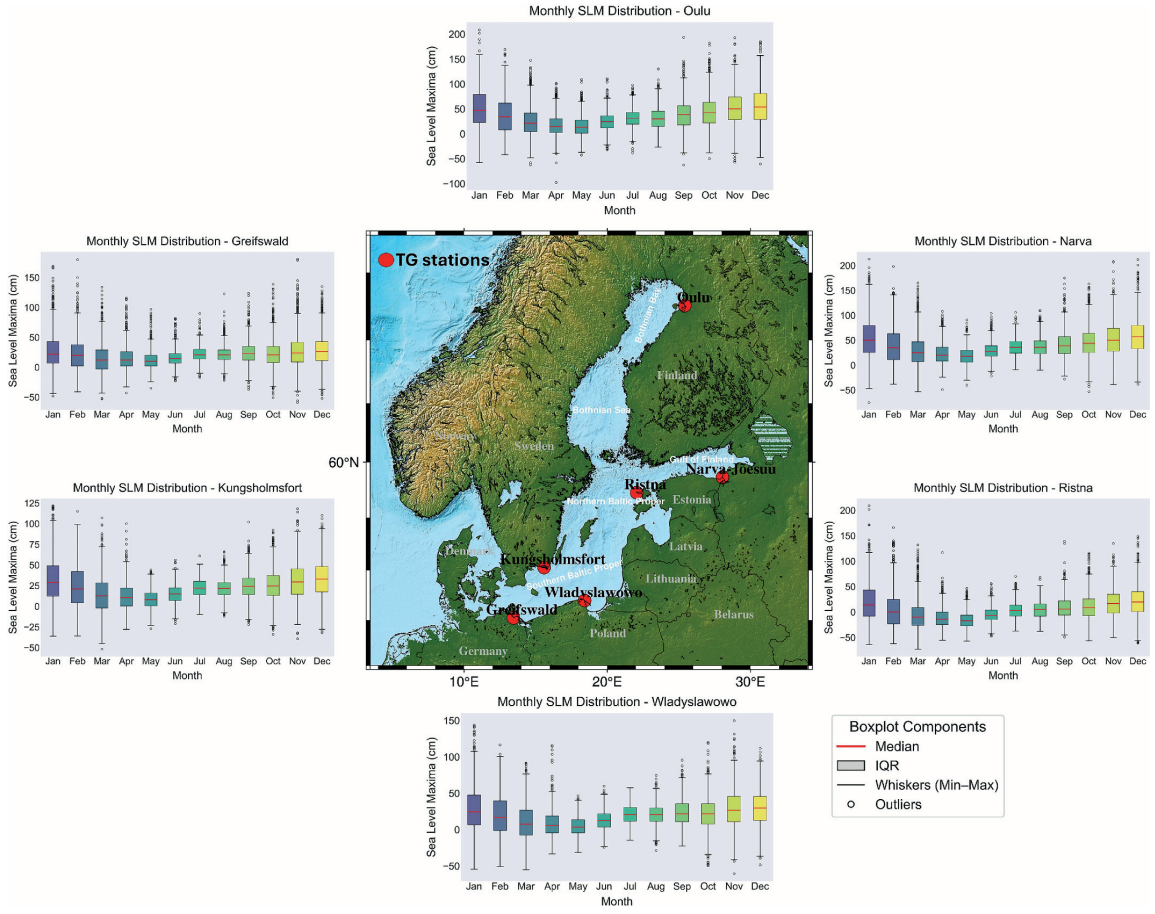


Fig. 3. Locations of selected tide gauge stations in the Baltic Sea: Narva- (Gulf of Finland), Ristna (north-eastern Baltic proper), Oulu (Bothnian Bay), Greifswald (south-western Baltic), Władysławowo (southern Baltic), and Kungsholmsfort (southern Baltic proper) along with the monthly boxplots of SLM variations 1971–2022.

northern Europe, divided into several sub-basins (Leppäranta and Myrberg, 2009). Freshwater input from rivers and limited saltwater exchange through Danish straits results in variable salinity, while tidal effects are negligible (Rutgersson and Kjellström, 2022). The Baltic region is one of the world's most active maritime regions (HELCOM, 2021). Long-term observations has shown a rising mean sea level, from 1.3-1.6 mm/year in the 20th century to 3.3 mm/yr for recent decades (Madsen et al., 2019; Mostafavi et al., 2024).

The Baltic Sea's extensive tide gauge records (Weisse et al., 2021), makes it an ideal location to evaluate ML-based SLM forecasting. Extremes are driven by winds, atmospheric pressure, prefilling (preconditioning), waves, precipitation and evaporation rates, the flow of rivers into the sea, seiches and meteotsunamis, and the basin geometry, with the highest levels occurring in the southwestern, eastern, and northern coasts (Weisse et al., 2021; Elken et al., 2024).

Although, typical SLM peak at 0.8 m above the mean sea level, severe events occur, especially in winter (Weisse et al., 2021). Preconditioning can amplify the extremes even on moderate winds (Andrée et al., 2023). Wave set-up, can cause sea level increases by up to 70–80 cm and SWHs have reached 8.2 m in Dec 2004 (Soomere et al., 2013, 2020).

Six TG stations— Narva-Jõesuu (hereafter Narva, for brevity; located in the eastern end of the Gulf of Finland), Ristna (north-eastern Baltic proper), Oulu (Bothnian Bay), Greifswald (south-western Baltic Sea), Władysławowo (southern Baltic Sea), and Kungsholmsfort (southern Baltic proper)—were selected based on data quality, regional coverage, and ability to capture extremes. Monthly boxplots of SLM (Fig. 3) highlight that sea level peaks predominantly between October and March, aligning with the stormy season. Extreme peaks are most intense in January. Narva, Ristna, and Oulu have recorded highest extremes (> 2 m) during the study period.

3.2. Data collection

3.2.1. In-situ tide gauge data

Hourly relative sea level data from six tide gauge stations were collected from different organizations for 1971–2022 (Table 1). Relative sea level observations were converted to Baltic Sea Chart Datum (BSCD2000) (Liebsch et al., 2023) to ensure consistency, following the conversions outlined in Jahanmard et al. (2022). Missing data were minimal (0.1–0.5 %) and were filled using a linear interpolation technique, which had negligible impact on models' performance.

3.2.2. Meteorological features

Meteorological variables, including wind components, surface pressure, evaporation, precipitation, and river runoff were obtained from the ERA5 re-analysis dataset (Hersbach and Bell, 2020). The Baltic Sea Index (BSI) representing large-scale atmospheric forcing and Baltic–North Sea water exchange, was also included. BSI is defined by the difference in normalized pressure anomalies between the coordinates 53°30′N, 14°30′E (Szczecin, Poland) and 59°30′N, 10°30′E (Oslo, Norway) (Lehmann et al., 2002). ERA5 data were extracted from nearest 0.25°×0.25° grid point to each tide gauge station.

SWH data were sourced from the SWAN wave model 1971–2005 (Booij et al., 1999; Björkqvist et al., 2018) and the WAM wave model 2006–2022 (WAMDI, 1988). Differences between the wave models are negligible, for bulk parameters such as SWH (Umesh et al., 2018), with demonstrated agreement even in complex coastal settings (Björkqvist et al., 2020). SWH values were extracted as the maximum within a 0.1°×0.1° box around each station, to account for spatial variability and avoid data gaps.

Although hourly data were available, daily SLM was used by taking the highest hourly value per day. Same aggregation approach was also applied to all meteorological and oceanographic features, including daily minimum atmospheric pressure (Table 2).

4. Data preprocessing

Ensuring that the time series is stationary is crucial in forecasting tasks (Raj and Prakash, 2024). Stationarity of the sea level time series was assessed using the Augmented Dickey–Fuller (ADF) test with up to 15 lags (Lopez, 1997). All stations showed strong rejection of the null hypothesis of non-stationarity, with ADF statistics below critical values at the 1 %, 5 %, and 10 % significance levels. Prior to ML implementation, optimal lag selection and data normalization were performed. The lag selection was based on Bayesian Information Criterion (BIC) criterion (Aertsen et al., 2010), due to its ability to evaluate model performance, which favors parsimonious models. The optimal lag was 5 days for Narva, Kungsholmsfort, Władysławowo, Oulu and Greifswald, and 3 days for Ristna. Model performance deteriorated for both smaller and larger lags, confirming these choices. All input features were also feature-wise normalized using the standard scaler function in Python. The results for feature selections, and lag selection are provided in supplementary material S1.

The Bayesian Optimization (BO) process ran for 50 iterations. Neural network models used the Adam optimizer to minimize Mean Squared Error, with a batch size of 32, 200 epochs, and a kernel size of 2 for hybrid DL models. Early stopping was applied with a patience of 10. BO explored different activation functions (ReLU, Leaky ReLU, tanh) to add non-linearity term. Hyperparameter ranges were chosen based on prior studies and practical considerations. Table S5 (supplementary material S1) summarizes the tuning ranges and final values determined by BO. After performing the preprocessing, we prepared the training, validation, and test sets with 70–15-15 % split ratio: the first 70 % of the data was used for model training, the next 15 % for model evaluation, and the remaining 15 % to assess the model's performance on unseen data.

5. Results and analysis

5.1. Performance of ML/DL models

This Section compares the performance of the five developed models across selected TG stations. Fig. 4 summarizes the training and test results and shows that the proposed CNN-GRU model consistently outperforms the other approaches in terms of test RMSE and R^2. Across all station, test RMSE for CNN-GRU varied from 7 cm at Kungsholmsfort to

Table 1
Geographical details on selected tide gauges in the Baltic Sea.

Station	Latitude (°N)	Longi-tude (°E)	Country	Datum	Missing data rate	Conversion of TG records to BSCD2000 (cm)	Source
Narva-Jõesuu	59.4691	28.0421	Estonia	EH2000	0.109 %	–500	envir.ee
Ristna	58.9212	22.0552	Estonia	EH2000	0.285 %	–500	envir.ee
Oulu	65.0403	25.4182	Finland	N2000	0	–	ilmatieteenlaitos.fi
Greifswald	54.0928	13.446	Germany	DHHN92	0	–496.9	wsv.bund.de
Wladyslawowo	54.7968	18.4187	Poland	PL-EVRF2007-NH	0.502 %	–494.4	imgw.pl
Kungsholmsfort	56.1053	15.5894	Sweden	RH2000	0	–	smhi.se

Table 2
Meteorological variables used in this study (computed across all stations).

Variable	Units	Source	Temporal resolution	Statistics		
				Min	Mean	Max
Zonal wind speed	m/s	ERA5	Hourly	–16.33	3.53	24
Meridional wind speed	m/s	ERA5	Hourly	–14.52	3.21	20.41
Wind gust	m/s	ERA5	Hourly	1.96	11.11	35.82
Surface atmospheric pressure	Mbar	ERA5	Hourly	944.98	1008.2	1052.2
Significant wave height	m	SWAN and WAM	Hourly	0	0.89	7.31
Evaporation minus precipitation	m	ERA5	Hourly	–0.00090	–0.000027	0.00019
Surface runoff	m	ERA5	Hourly	–4.34e-19	1.0196e-06	0.00082
Baltic Sea Index	–	ERA5	Hourly	–1.6192	0.2302	2.6213
Sea level	cm	TGs	Hourly	–97.5	23.2914	213

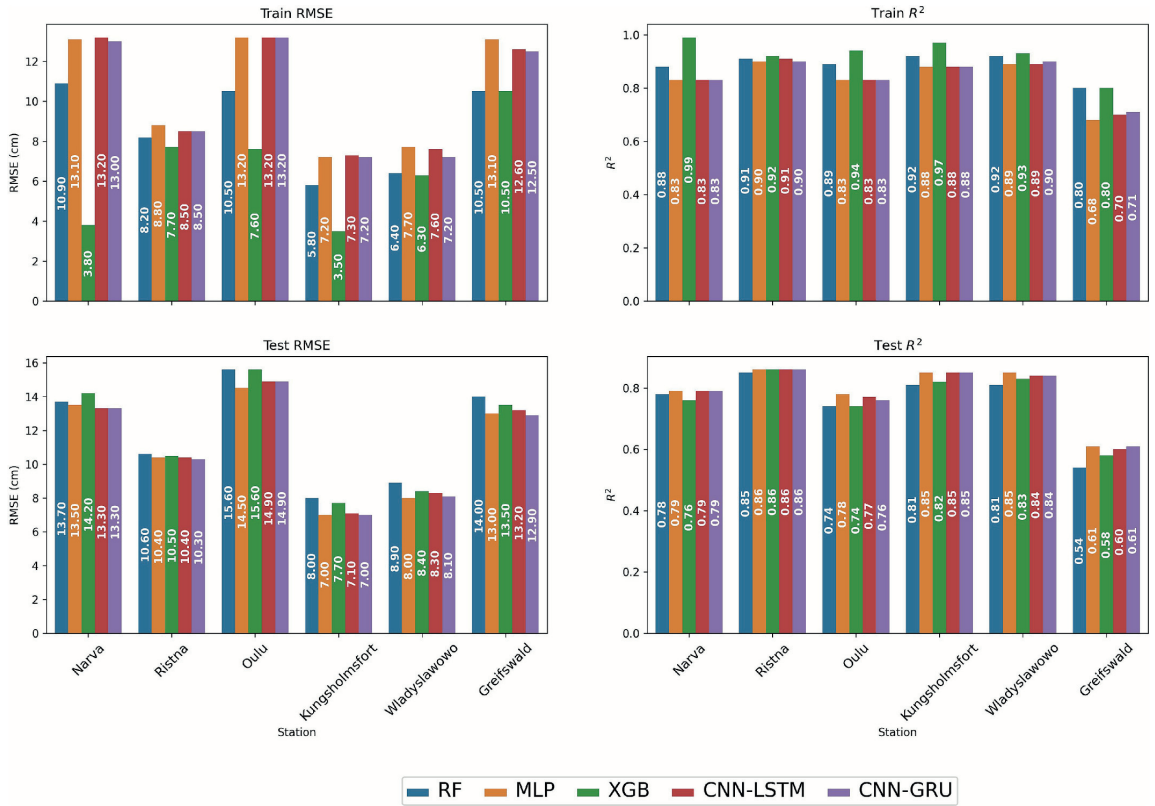


Fig. 4. Statistical evaluation of the models using R^2 and RMSE for training and test sets at different stations.

14.90 cm at Oulu, while R^2 values varied from 0.61 at Greifswald to 0.86 at Ristna, demonstrating adaptability and effectiveness for different local conditions. At Narva, both CNN-GRU and CNN-LSTM exhibited the best test performance (RMSE of 13.30 cm, $R^2 = 0.79$), indicating good generalization. The MLP ranked second, with similar training and test RMSE values. In contrast, XGB and RF models, showed pronounced overfitting, with test RMSE increasing sharply relative to training. At Ristna, all models performed similarly, with a small difference in RMSE and comparable R^2. However, CNN-GRU achieved the best test performance ($R^2 = 0.86$ and RMSE = 10.30 cm).

At Oulu, MLP achieved the best test results (with an RMSE = 14.50 cm, $R^2 = 0.78$). It showed a modest increase from the training RMSE of 13.20 cm. CNN-LSTM and CNN-GRU models performed comparably, whereas RF and XGB again showed significant overfitting. Similar results were obtained at Władysławowo, where MLP achieved the lowest test RMSE (8 cm) and highest R^2 (0.85), followed closely by CNN-GRU and CNN-LSTM. Tree-based models showed notable performance degradation from training to testing. At Kungsholmsfort, CNN-GRU and MLP shared the best performance (RMSE = 7 cm, $R^2 = 0.85$), while CNN-LSTM also generalized well. At Greifswald, the CNN-GRU again performed best (RMSE = 12.90 cm, $R^2 = 0.61$), with MLP close behind.

Overall, comparing training and test metrics suggests that RF and XGB frequently overfit, particularly at Narva, Oulu, and Kungsholmsfort, and Władysławowo. These decision-based models likely captured noise or localized extremes that did not generalize well. In contrast, neural-network-based models (MLP, CNN-GRU, and CNN-LSTM models) demonstrated better generalization capabilities across stations. Therefore, the CNN-GRU, MLP, and CNN-LSTM models emerge

as more reliable choices for forecasting SLM of the Baltic Sea. At Ristna, all ML/DL models performed similarly, suggesting more linear and predictable SLM patterns compared to other stations.

To evaluate whether the different SLM forecasting methods exhibit statistically significant differences, we applied the Friedman test (Demšar, 2006), a non-parametric statistical test suitable for comparing multiple models. The results revealed significance for both performance metrics: RMSE ($\chi^2 = 20.452$, p-value $= 0.0004$) and R^2 ($\chi^2 = 19.92$, p-value $= 0.0005$), indicating that the applied models differ significantly (95 % confidence interval).

5.2. Time series forecasting

Model time series forecasts were compared with observed SLM during the test period (Fig. 5). Overall, the MLP, CNN-GRU, and CNN-LSTM models performed reasonably well, with CNN-GRU providing the most consistent results across stations. CNN-based models accurately captured next-day SLM variability and major peaks at Narva and Ristna, including events exceeding 150 cm.

At Oulu and Kungsholmsfort, frequent moderate peaks were captured. However, both models underestimated the SLM peaks, specifically those exceeding 160 cm at Oulu and 120 cm at Kungsholmsfort. The MLP model closely followed observed SLM patterns and, in some cases (November 2016 and 2018), outperformed the CNN-based models in peaks representations. The XGB model showed occasional signs of overfitting and occasional overestimations. However, all models tended to underestimate sea levels greater than 150 cm.

Scatter density plots (Fig. 6) shows that CNN-GRU predictions align closely with observations, at most stations achieving correlation coefficients ~ 0.9 and IQR ranges of 7–8 cm. Performance was weaker at Oulu and Greifswald, where extremes were underestimated and error dispersion was higher. For Oulu, the performance of the CNN-GRU model was acceptable. Overall, CNN-GRU emerges as the most robust model for SLM forecasting, while underestimation of rare extremes remains a common challenge across all models.

5.3. Models' performance in scenario-based storm event detection

Storm events are a major contributor to SLM and pose the greatest challenges for all models. Models performance were evaluated for five major recent Baltic Sea storms (2017–2020). Figs. 7 and A.1b show the course of water level during the maxima of these major storms at six TG stations: Xavier (October 4–6, 2017, 118.6 cm at Władysławowo), Eleanor (January 2–4, 2018, 159 cm at Narva), Aapeli (January 1–2, 2019, 169 cm at Greifswald), Lorenzo (October 2–7, 2019, 107 cm at Narva), and Ciara (February 3–16, 2020, 161.30 cm at Oulu).

Overall, the models captured the timing and general evolution of storm-driven SLM, though peak magnitudes were often underestimated, especially during the most extreme events. The CNN-GRU and CNN-LSTM models consistently outperformed other approaches in reproducing storm-related SLM variability and peak timing across most stations. They showed particularly strong performance at Narva, Ristna, and Oulu, including sharp and rapidly evolving peaks during storms such as Lorenzo and Ciara. The MLP model also performed well and, in some cases, better captured peak magnitudes. In contrast, XGB frequently overestimated storm peaks, indicating limited robustness for extreme-event forecasting.

Underestimation was most pronounced during storm Aapeli, suggesting the influence of additional drivers not fully represented in the input data. Analysis of feature variability in supplementary materials (Fig. A1a) indicated that pressure and wind were the dominant contributors during most storms, while during storm Aapeli, pressure and wind variations were generally lower, suggesting that Aapeli's extreme impacts may have been underestimated. This suggests that wave prop-

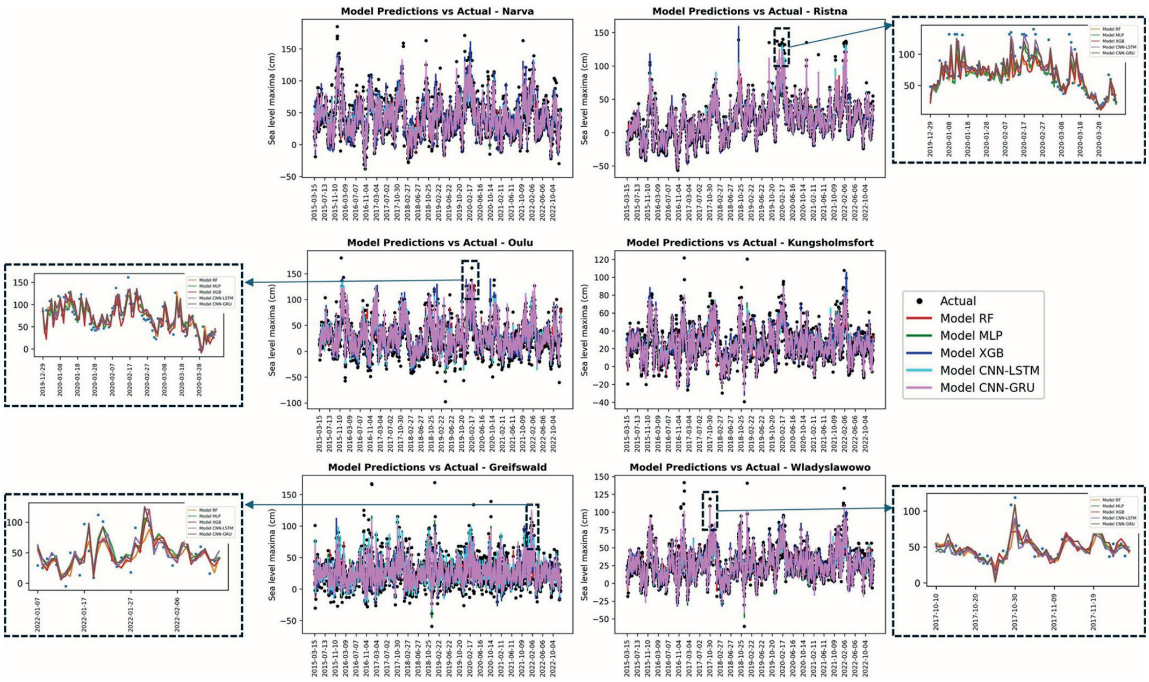


Fig. 5. Time series forecasting performance of different models at selected TG stations during the test period.

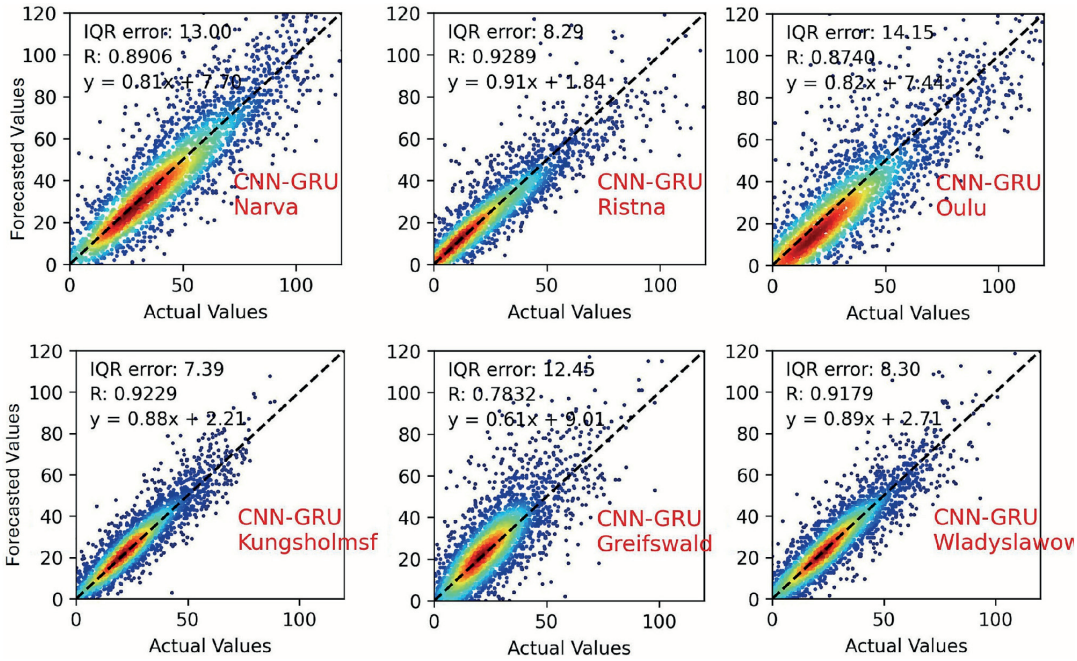


Fig. 6. Scatter density plot of CNN-GRU model's predictions vs. actual SLM for different stations.

erties like wave set-up (Gillibrand et al., 2011) were a key driver, a conclusion supported by Eelsalu et al. (2014), who found that such set-up is responsible for water level extremes that exceed standard return level projections.

Peak detection skill was further assessed using precision, recall, and F1 score across multiple alarm thresholds ranging from 60 to 100 cm. The results are presented in Fig. A2 in Appendix. To further explore these differences, we conducted post-hoc pairwise comparisons using the Wilcoxon Signed-Rank test (Demšar, 2006). Based on test RMSE, the models were categorized into two main groups: neural network–based approaches (CNN-GRU, CNN-LSTM, and MLP) and tree-based ensemble methods (RF and XGB). Regarding R^2 and precision, MLP achieved the highest performance, whereas CNN-GRU outperformed in terms of RMSE, recall, and F1 score. Detailed statistical test results and model comparisons are provided in supplementary material S1 and S2.

5.4. Model's explainability

Given its superior predictive capabilities (ranking as the top performer across more stations and metrics), the CNN-GRU was further analyzed using SHAP feature importance to understand the drivers behind its predictions. Fig. 8 shows the top ten contributing features for each TG station based on SHAP bar plot feature analysis. Globally, the most influential feature was the previous day's SLM (SLM_lag1), emphasizing the critical role of prefilling and current water volume in short-term SLM forecasting, a relationship addressed in previous studies (Wiśniewski and Wolski, 2011).

At eastern Baltic stations (Narva and Ristna) and Władysławowo, atmospheric pressure and wind components (U- and V-components) were the dominant meteorological drivers, with significant wave height playing a minor role only at Ristna. In contrast, at western and northern stations (Oulu, Kungsholmsfort, Greifswald), the Baltic Sea Index (BSI) lag 1 was the primary feature, followed by pressure and wind. This observation suggests that these locations are more affected by atmospheric forcing from the North Atlantic than the eastern, usually "downwind" stations where more localized effects are frequent and the impact of severe waves is often explicit (Eelsalu et al., 2014).

Overall, prefilling accounted for ~ 40 % of total feature importance, with pressure lags contributing 13 % and 9 %, and wind components adding further influence. These results highlight that while SLM_lag1 is the key predictor across all stations, the relative importance of meteorological drivers varies geographically, with eastern stations dominated by local wind and pressure effects, and western/northern stations more affected by large-scale atmospheric patterns captured by the BSI.

5.5. Extreme value analysis at different TG stations

Extreme value analysis using the GEV distribution with block maxima was applied to understand the intensity and frequency of SLM at six TG stations over 52 years. As displayed in Fig. 9, seasonal extremes were evaluated, revealing that winter consistently exhibits the highest water levels, while summer extremes rarely exceed 100 cm. The largest 50-year return levels were projected at Narva (212 cm) and Oulu (181 cm), with 2-year extremes also highest at these stations (Narva: 116 cm; Oulu: 112 cm). Other stations, such as Kungsholmsfort and Władysławowo, showed lower extreme values, with rare events exceeding 130 cm.

During the test period (2015–2022), some of the highest SLM values occurred, including 159 cm at Narva during storm Eleanor (January 2018) and 161.3 cm at Oulu during storm Ciara (February 2020). These events correspond to winter return periods of approximately 5 yr at Narva and 7 yr at Oulu (Fig. 9) and highlight the challenge for ML/DL models, which generally underestimated such extremes. These extremes pose significant forecasting challenges for ML/DL models and

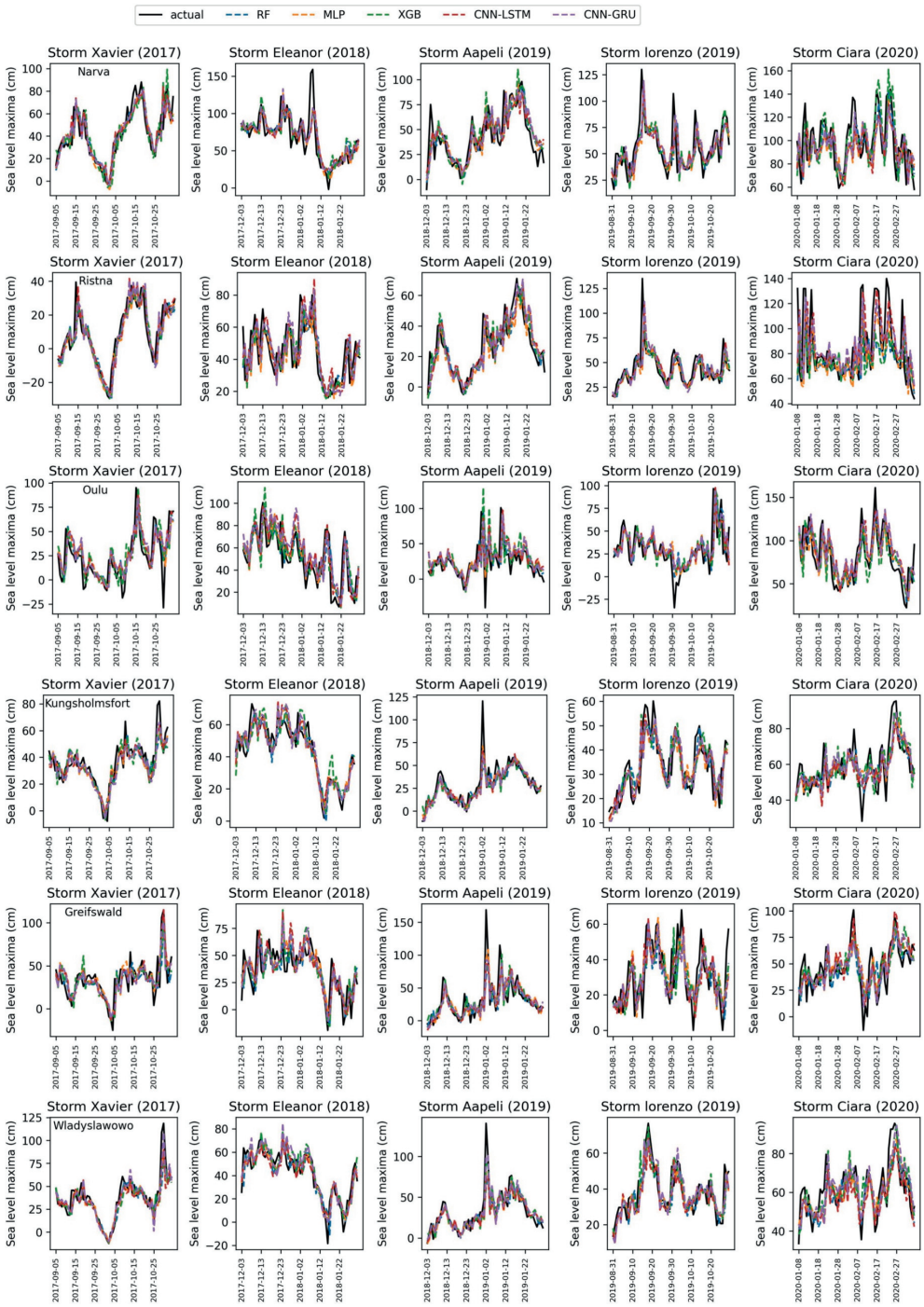


Fig. 7. Forecast by ML models and observed sea level (tide gauges) time series at six tide gauge stations capturing the SLM during storm events Xavier, Eleanor, Aapeli, Lorenzo, and Ciara.

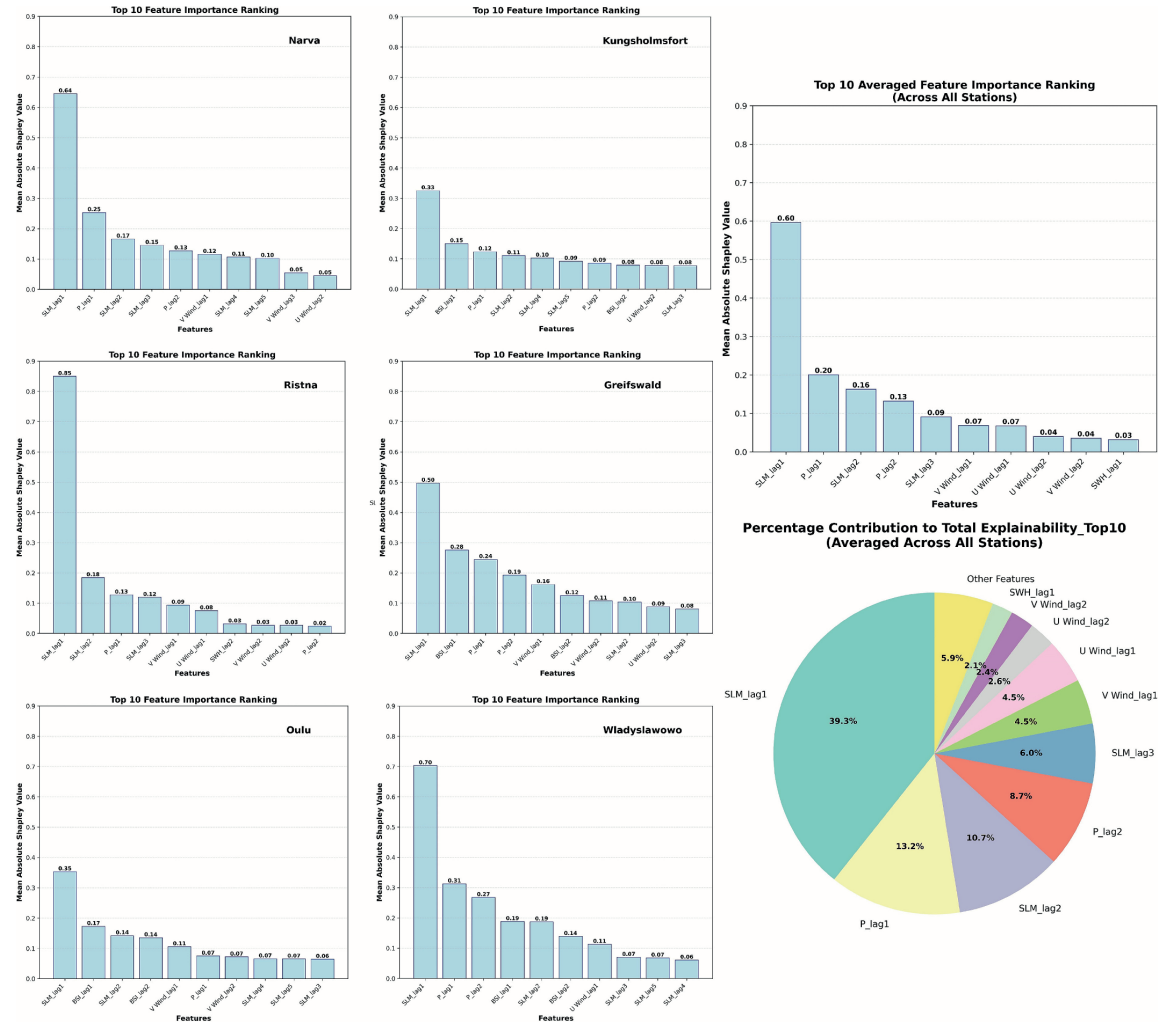


Fig. 8. Features' contributions (top ten) in the CNN-GRU model at different stations using the SHAP explainability analysis. SLM refers to sea level maxima, *p* stands for pressure, SWH denotes significant wave height, BSI is the Baltic Sea Index, *U* wind and *V* wind represent the zonal and meridional wind speed components, respectively. The lag number indicates how many previous time steps (days) the model uses to forecast future value (Section 4.1). For example, "P lag-2" means the pressure value shifted by two time steps in a time series analysis.

understanding their seasonality and return periods can help improve coastal management and preparedness strategies.

5.6. Study limitations and model applications

A key limitation of this study is the underestimation of sea level peaks above 150 cm. While predictions for 100–150 cm were significantly improved compared to previous regional studies, rare extremes are underrepresented in the training data, meteorological predictors may not fully capture extreme conditions, and local processes (e.g., surges, bathymetry, seiches) were not explicitly included. Non-stationarity of Baltic Sea extremes and systematic limitations in physics-based models further complicate forecasting of the most severe events. To mitigate these challenges, models were trained on daily maxima and weighted loss functions (Man et al., 2023) were tested but rarest peaks remained underestimated. Even physics-based circulation models show similar challenges; for instance, wind forcing adjustments reduced bias but introduced new errors (Lorenz and Gräwe, 2023). Future improvements could include longer observational records, additional coastal predictors, physics-informed loss functions, and ensemble approaches.

Station selection captured diverse extreme sea level regimes: Narva and Ristna represent wind-driven surges, Kungsholmsfort and Oulu capture open-sea and northern basin dynamics, while Władysławowo and Greifswald capture westerlies-dominated semi-enclosed conditions. This ensures models are validated across contrasting oceanographic

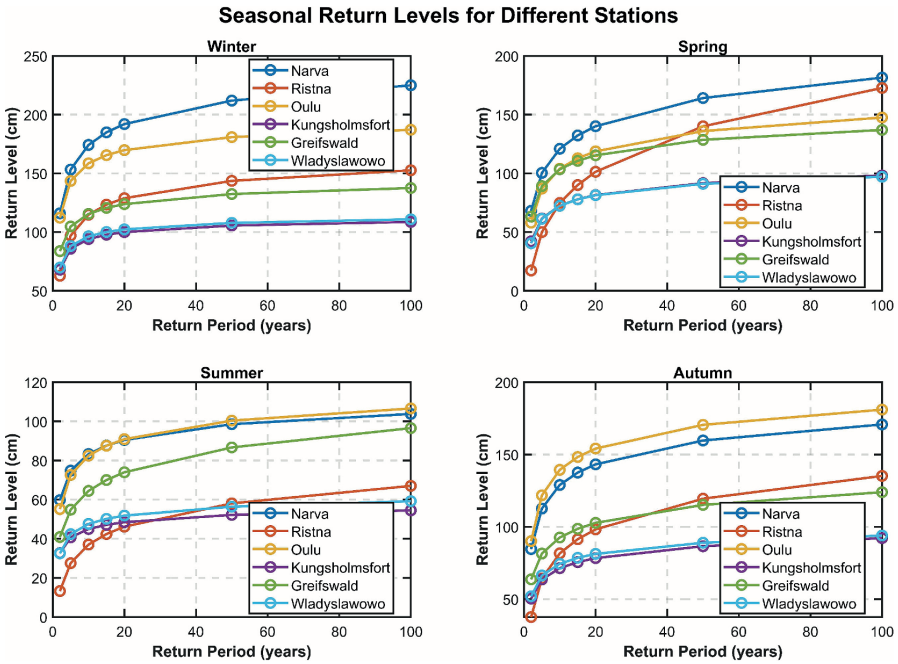


Fig. 9. Projected extreme water levels based on a seasonal block maxima method and a GEV distribution at different stations.

conditions. Operationally, extending forecasts to denser station networks, accounting for sea-level rise and storminess trends, and integrating threshold-based or multi-class approaches would improve extreme sea level predictions. Ultimately, these results have direct applications for coastal defences design, port infrastructure, ecosystem management, and climate adaptation planning.

6. Discussion

Motivated by persistent underprediction of extreme sea levels in previous ML studies, this work evaluates ML and DL methods for short-term SLM forecasting at six selected Baltic Sea tide gauges (Narva, Ristna, Oulu, Kungsholmsfort, Greifswald, and Władysławowo) over 1971–2022. Consistent with earlier findings, SLM extremes are often higher and more frequent in the eastern and northern Baltic, reflecting the spatial heterogeneity of dominant forcing mechanisms.

To complement the short-term ML/DL forecasts, Extreme Value Theory (EVT) was applied to quantify long-term risk through return periods and extreme magnitudes. Feature selection was performed using a nonlinear mutual information, complemented by correlation analysis, to reduce model complexity while retaining physically meaningful predictors. Predominant features are pressure, wind components and gust, BSI and SWH. Feature relevance varied spatially, with notable collinearity between SWH and wind gusts at Ristna, Kungsholmsfort, and Władysławowo. While prior knowledge of influential factors supports effective feature selection, rare winter extremes remain difficult to capture, suggesting that local drivers (e.g., seiches, wave set-up, etc.), not emphasized by statistical selection may be critical to improve forecasting of the most severe events.

SHAP explainability analysis for best-performing model (CNN-GRU) shows that prefilling is dominant at all stations and highlights clear regional differences: western stations (e.g., Greifswald) are mainly influenced by large-scale forcings, highlighting the role of Baltic–North Sea water exchange. Eastern stations (Narva and Ristna) are driven primarily by local pressure and wind speed with SWH additionally important at Ristna. BSI remains influential at Oulu and southern stations, consistent with known Baltic Sea dynamics. Wind-driven impacts varies spatially with zonal wind stronger at Władysławowo and Kungsholmsfort, whereas meridional winds dominating elsewhere. Variables such as evaporation, river runoff, and precipitation were not relevant for short-term SLM forecasting. Moreover, including these low-MI features generally degraded performance due to added noise.

Neural network-based (CNN-GRU, CNN-LSTM, and MLP) models consistently outperformed tree-based methods (XGB and RF) across most stations. The CNN-GRU showed best overall performance. Ranking highest in RMSE, recall, and F1-score at four stations (Narva, Ristna, Kungsholmsfort, and Greifswald; RMSE of 7–14.9 cm,R^2 = 0.61–0.86) and demonstrated superior generalization with test-to-training RMSE ratios close to 1. This may be attributed to the use of rigorous early stopping criteria or BO optimization, enhancing generalization performance (Zhou and Zhang, 2023). MLP performed best at Oulu and Władysławowo and achieved the highest R^2 and precision. CNN-LSTM showed stable but moderate performance. Tree-based models like XGB and RF exhibited overfitting, with strong training but weaker test results.

Time series evaluation (2015–2022) showed that neural-network models captured SLM variability and most peaks between 100–130 cm, representing a clear improvement over earlier Baltic studies that struggled > 100 cm (Rajabi-Kiasari et al., 2023, 2025). However, extreme peaks (>150 cm) at stations such as Narva, Oulu, and Greifswald were still underestimated. Statistical tests confirmed CNN-GRU and MLP as the strongest models, with CNN-GRU leading in balanced

extreme-event detection (recall and F1-score), MLP in precision and CNN-LSTM is steady but never dominant.

A scenario-based analysis of five major storms during the test period showed that models captured the timing and magnitude of moderate SLM events well, especially up to 130 cm, but consistently underestimated rarer extremes exceeding 150 cm during storm Eleanor (January 2018, Narva), Ciara (February 2020, Oulu), and Aapeli (January 2019, Greifswald). This limitation reflects both data scarcity—very few comparable extreme events exist in the training record—and the assumption of stationarity in extreme sea level formation. Recent studies indicate pronounced non-stationarity in Baltic Sea extremes, with spatially variable trends in frequency and magnitude (Viigand et al., 2025), further complicating short-term prediction.

Analysis of storm-related input feature showed that wind components and atmospheric pressure generally exhibit strong variability and were well captured by models. However, during storm Aapeli features variability was relatively weak at some stations despite high observed peaks, limiting model responsiveness. This suggests that additional physical drivers not included in the inputs—such as wave set-up, local surges, bathymetry, and seiches—may have played a critical role.

Despite these challenges, ML/DL approaches advanced Baltic Sea level forecasting by extending predictions from 100 cm to 150 cm (Rajabi-Kiasari et al., 2023, 2025). Extreme value analysis using seasonal GEV distribution provided complementary context, indicating that 150 cm events correspond to return periods of ~ 5 yr at Narva and ~ 7 yr at Oulu in winter. This estimate is consistent with recent studies in the Baltic Sea (Wolski et al., 2025). In contrast, at stations like Greifswald and Władysławowo, such peaks are considerably rarer. The scarcity of such events, combined with evolving climate conditions and missing local processes, explains the persistent underestimation of the most extreme SLM. This challenge is not unique to our approach, as even process-based Baltic Sea models struggle to fully reproduce extremes (Lorenz and Gräwe, 2023).

7. Concluding remarks

This study demonstrated that Bayesian optimized neural-network models performed well in forecasting sea level maxima across Baltic Sea. CNN-GRU achieved the best overall performance (lowest RMSE, highest recall, and F1-score), while MLP performed best in R^2 and precision. The framework reliably predicts most SLM events with 7–15 cm accuracy and extends the effective forecasting range from ~ 100 cm to ~ 150 cm. Model skills decreased for the rarest extremes (>150 cm), which mainly occur in winter and correspond to 5–7-year winter and 20–30-year autumn return periods in the eastern and northern Baltic. This limitation reflects data scarcity and missing local physical processes, emphasizing the need for improved observations via physics-informed strategies.

Explainability analyses (SHAP) revealed prefilling as the dominant predictor (~40 %) at all stations with strong regional contrasts: eastern stations are primarily driven by local pressure, wind and wave, whereas western and northern stations are more influenced by large-scale atmospheric forcing represented by BSI. Overall, integrating ML/DL-based short-term forecasting with EVT provides a robust and interpretable framework for extreme sea level prediction. Future studies should focus on longer observational records, ensemble and probabilistic approaches, and the inclusion of additional coastal processes to better capture rare extremes and improve the robustness of coastal forecasting, risk assessment, and climate adaptation.

Data & Scripting Statement.

Tide gauge sea level data were obtained from various national organizations: Estonia (envir.ee), Finland (ilmatieteenlaitos.fi), Sweden (smhi.se), Germany (wsv.bund.de), and Poland (imgw.pl). Meteorological data were sourced from ERA5, while significant wave height data were acquired from the WAM model (marine.copernicus.eu) and the SWAN model (wavelab.taltech.ee). Data analysis and visualization were performed using Python and MATLAB scripts. The written scripts for this study are available upon request.

Author contributions.

"SR; Conceptualization, Methodology and Analysis, Software, Visualization, Validation, Writing- Original Draft, ND; Conceptualization, Writing – review & editing, Validation, Supervision, AE; Conceptualization, Writing – review & editing, Validation, Supervision, Project administration, TS; Conceptualization, Writing – review & editing, Validation. All authors have reviewed and approved the published version of the work.".

CRediT authorship contribution statement

Saeed Rajabi-Kiasari: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Nicole Delpeche-Ellmann:** Writing – review & editing, Validation, Supervision, Conceptualization. **Artu Ellmann:** Writing – review & editing, Validation, Supervision, Project administration, Conceptualization. **Tarmo Soomere:** Writing – review & editing, Validation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research was supported by the Estonian Research Council grants PRG1129 and PRG1785 DYNAREF. We are grateful to anonymous reviewers for their insightful feedback, which significantly enhanced this manuscript.

Appendix A. – Complementary figures

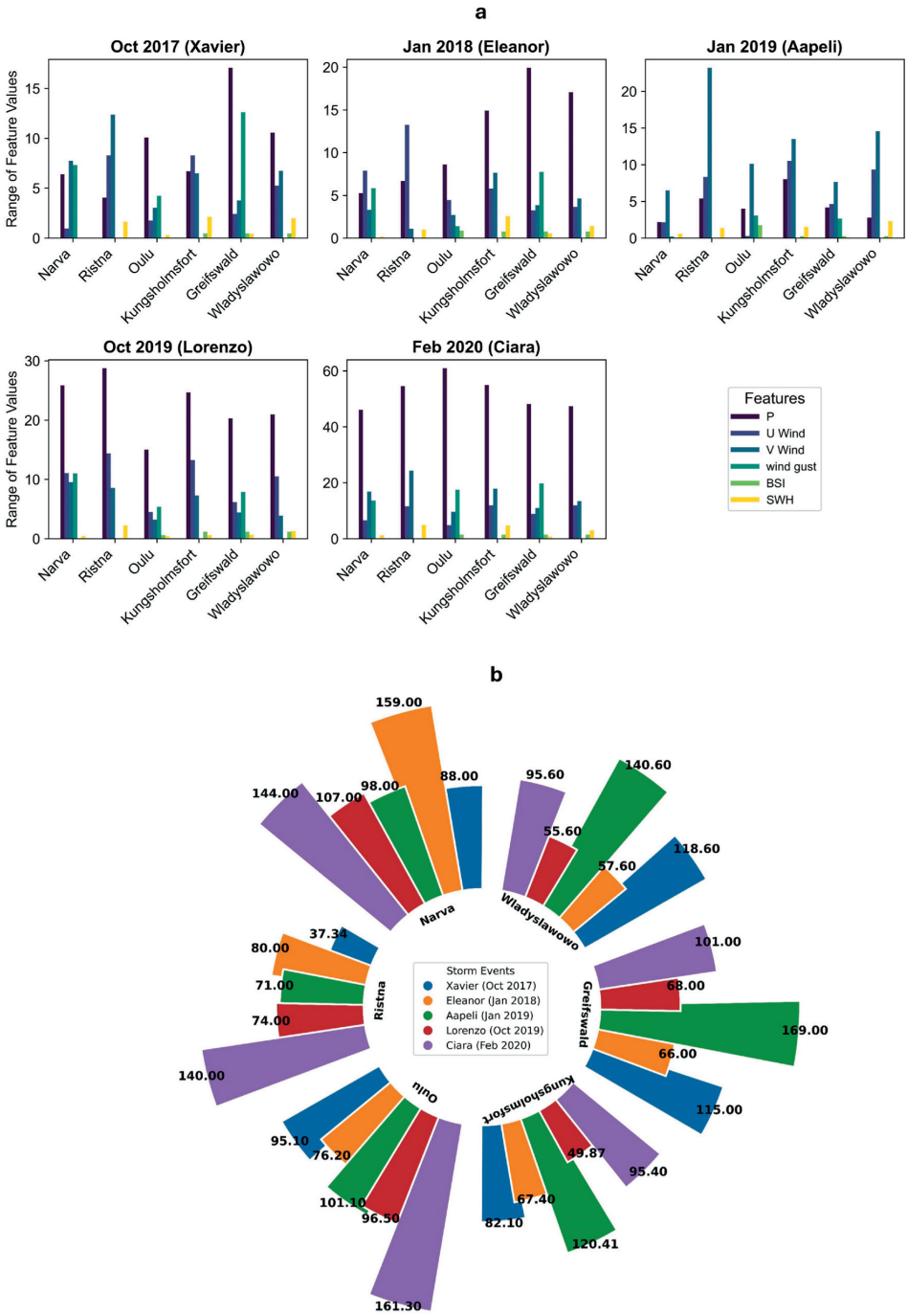


Fig. A1. a) Ranges of input feature variations during the five storm events (Xavier, Eleanor, Aapeli, Lorenzo and Ciara) across different stations; b) Circular bar chart showing the peaks of SLMs (in cm) during each storm event at different stations. Each station's data is represented by groups of bars, where each color corresponds to a different storm event. The bar heights indicate the reached sea level maximum.

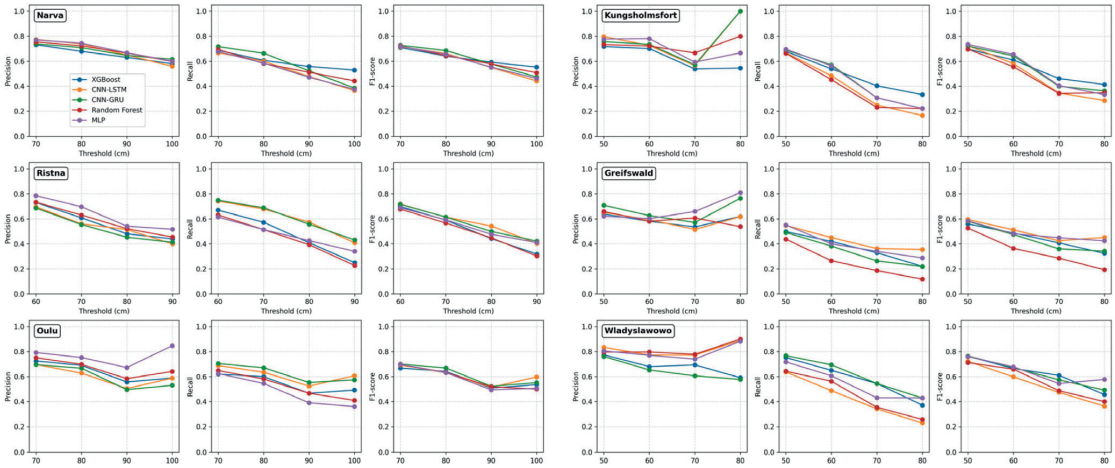


Fig. A2. Performance of different models in detecting peak events across stations, evaluated using precision, recall, and F1-score metrics.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2025.105064

Data availability

Data will be made available on request.

References

Aertsen, W., Kint, V., van Orshoven, J., Özkan, K., Muys, B., 2010. Comparison and ranking of different modeling techniques for prediction of site index in Mediterranean mountain forests. Ecol. Model. 221, 1119–1130. https://doi.org/10.1016/j.ecolmodel.2010.01.007.

Andrée, E., Su, J., Dahl Larsen, M.A., Drews, M., Stendel, M., Skovgaard Madsen, K., 2023. The role of preconditioning for extreme storm surges in the western Baltic Sea. Nat. Hazards Earth Syst. Sci. 23, 1817–1834. https://doi.org/10.5194/nhess-23-1817-2023.

Averkiev, A.S., Klevannyy, K.A., 2010. A case study of the impact of cyclonic trajectories on sea-level extremes in the Gulf of Finland. Cont. Shelf Res. 30, 707–714. https://doi.org/10.1016/j.csr.2009.10.010.

Bellinghausen, K., Hünicke, B., Zorita, E., 2025. Short-term prediction of extreme sea-level at the Baltic Sea coast by Random Forests. Nat. Hazards Earth Syst. Sci. 25, 1139–1162. https://doi.org/10.5194/nhess-25-1139-2025.

Björkqvist, J.-V., Lukas, I., Alari, V., van Vledder, G.P., Hulst, S., Pettersson, H., Behrens, A., Männik, A., 2018. Comparing a 41-year model hindcast with decades of wave measurements from the Baltic Sea. Ocean Eng. 152, 57–71. https://doi.org/10.1016/j.oceaneng.2018.01.048.

Björkqvist, J.V., Vähä-Piikkiö, O., Alari, V., Kuznetsova, A., Tuomi, L., 2020. WAM, SWAN and WAVEWATCH III in the finnish archipelago - the effect of spectral performance on bulk wave parameters. Journal of Operational Oceanography 13 (1), 55–70. https://doi.org/10.1080/1755876X.2019.1633236.

Booij, N., Ris, R., Holthuijsen, L., 1999. A third generation wave model for coastal regions, 1: model description and validation. J. Geophys. Res. Oceans 104 (C4), 7649–7666. https://doi.org/10.1029/98JC02622.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Bruneau, N., Polton, J., Williams, J., Holt, J., 2020. Estimation of global coastal sea level extremes using neural networks. Environ. Res. Lett. 15, 074030. https://doi.org/10.1088/1748-9326/ab89d6.

Chen, T., Guestrin, C., 2016. XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 785–794. Doi: 10.1145/2939672.2939785.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078.

Coles, S., 2001. An introduction to statistical modeling of extreme values. Springer, London.

Dai, Y., Yang, X., Leng, M., 2023. Optimized Seq2Seq model based on multiple methods for short-term power load forecasting. Appl. Soft Comput. 142, 110335. https://doi.org/10.1016/j.asoc.2023.110335.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

Di Nunno, F., Granata, F., Gargano, R., de Marinis, G., 2021. Forecasting of extreme storm tide events using NARX neural network-based models. Atmosphere (basel) 12, 512. https://doi.org/10.3390/atmos12040512.

Dubois, K., Andreas Dahl Larsen, M., Drews, M., Nilsson, E., Rutgersson, 2024. Extending sea level time series for the analysis of extremes with statistical methods and neighbouring station data. Ocean Science 20, 21–30. Doi: Doi: 10.5194/os-20-21-2024.

Eelsalu, M., Soomere, T., Pindsoo, K., Lagemaa, P., 2014. Ensemble approach for projections of return periods of extreme water levels in Estonian waters. Cont. Shelf Res. 91, 201–210. https://doi.org/10.1016/j.csr.2014.09.012.

Eelsalu, M., Soomere, T., Parnell, K., Viška, M., 2025. Attribution of alterations in coastal processes in the southern and eastern Baltic Sea to climate change driven modifications of coastal drivers. Oceanologia 67 (1), 67103. https://doi.org/10.5697/LXTZ5389.

Elken, J., Barzandeh, A., Maljutenko, I., Rikka, S., 2024. Reconstruction of Baltic Gridded Sea Levels from Tide Gauge and Altimetry Observations using Spatiotemporal Statistics from Reanalysis. Remote Sens. (Basel) 16, 2702. https://doi.org/10.3390/rs16152702.

Fan, D., Xue, K., Zhang, R., Zhu, W., Zhang, H., Qi, J., Zhu, Z., Wang, Y., Cui, P., 2024. Application of interpretable machine learning models to improve the prediction performance of ionic liquids toxicity. Sci. Total Environ. 908, 168168. https://doi.org/10.1016/j.scitotenv.2023.168168.

Gillibrand, P.A., Lane, E.M., Walters, R.A., Gorman, R.M., 2011. Forecasting extreme sea level events and coastal inundation from tides, surge and wave setup. Aust. J. Civ. Eng. 9, 99–112. https://doi.org/10.1080/14488353.2011.11463961.

Gräwe, U., Burchard, H., 2012. Storm surges in the Western Baltic Sea: the present and a possible future. Clim. Dyn. 39, 165–183. https://doi.org/10.1007/s00382-011-1185-z.

Grossman, E.E., Tehranirad, B., Nederhoff, C.M., Crosby, S.C., Stevens, A.W., Van Arendonk, N.R., Nowacki, D.J., Erikson, L.H., Barnard, P.L., 2023. Modeling extreme water levels in the Salish Sea: the importance of including remote sea level anomalies for application in hydrodynamic simulations. Water (basel) 15, 4167. https://doi.org/10.3390/w15234167.

Haigh, I.D., Wijeratne, E.M.S., MacPherson, L.R., Pattiaratchi, C.B., Mason, M.S., Crompton, R.P., George, S., 2014. Estimating present day extreme water level exceedance probabilities around the coastline of Australia: tides, extra-tropical storm surges and mean sea level. Clim. Dyn. 42, 121–138. https://doi.org/10.1007/s00382-012-1652-1.

Harter, L., Pineau-Guillou, L., Chapron, B., 2024. Underestimation of extremes in sea level surge reconstruction. Sci. Rep. 14, 14875. https://doi.org/10.1038/s41598-024-65718-6.

Hashemi, M.R., Spaulding, M.L., Shaw, A., Farhadi, H., Lewis, M., 2016. An efficient artificial intelligence model for prediction of tropical storm surge. Nat. Hazards 82, 471–491. https://doi.org/10.1007/s11069-016-2193-4.

HELCOM, 2021. Input of nutrients by the seven biggest rivers in the Baltic Sea region 1995–2017, in: Baltic Sea Environment Proceedings 178.

Hersbach, H., Bell, B., 2020. The ERA5 global reanalysis. Q. J. Roy. Meteor. Soc. 146 (730), 1999–2049. https://doi.org/10.1002/qj.3803.
Ishida, K., Tsujimoto, G., Ercan, A., Tu, T., Kiyama, M., Amagasaki, M., 2020. Hourly-scale coastal sea level modeling in a changing climate using long short-term memory neural network. Sci. Total Environ. 720, 137613. https://doi.org/10.1016/j.scitotenv.2020.137613.
Jahanmard, V., Delpeche-Ellmann, N., Ellmann, A., 2022. Towards realistic dynamic topography from coast to offshore by incorporating hydrodynamic and geoid models. Ocean Model. 102124. https://doi.org/10.1016/j.ocemod.2022.102124.
Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2019. Machine Learning for the Geosciences: challenges and Opportunities. IEEE Trans. Knowl. Data Eng. 31, 1544–1554. https://doi.org/10.1109/TKDE.2018.2861006.
Kirezci, E., Young, I.R., Ranasinghe, R., Muis, S., Nicholls, R.J., Lincke, D., Hinkel, J., 2020. Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st Century. Sci. Rep. 10, 11629. https://doi.org/10.1038/s41598-020-67736-6.
Lamberti, W.F., 2023. An overview of explainable and interpretable AI, in: AI Assurance. Elsevier, pp. 55–123. Doi: 10.1016/B978-0-32-391919-7.00015-9.
Lee, J.W., Irish, J.L., Bensi, M.T., Marcy, D.C., 2021. Rapid prediction of peak storm surge from tropical cyclone track time series using machine learning. Coastal Engineering 170, 104024.
Lehmann, A., Krauss, W., Hinrichsen, H.-H., 2002. Effects of remote and local atmospheric forcing on circulation and upwelling in the Baltic Sea. Tellus A 54, 299–316. https://doi.org/10.1034/j.1600-0870.2002.00289.x.
Leppäranta, M., Myrberg, K., 2009. Physical Oceanography of the Baltic Sea. Springer Science & Business Media, Praxis, Berlin, Heidelberg 2009. https://doi.org/10.1007/978-3-540-79703-6.
Li, X., Zhou, S., Wang, F., Fu, L., 2024. An improved sparrow search algorithm and CNN-BiLSTM neural network for predicting sea level height. Sci. Rep. 14, 4560. https://doi.org/10.1038/s41598-024-55266-4.
Lopez, J.H., 1997. The power of the ADF test. Econ. Lett. 57, 5–10. https://doi.org/10.1016/S0165-1765(97)81872-1.
Lorenz, M., Gräwe, U., 2023. Uncertainties and discrepancies in the representation of recent storm surges in a non-tidal semi-enclosed basin: a hind-cast ensemble for the Baltic Sea. Ocean Sci. 19 (6), 1753–1771. https://doi.org/10.5194/os-19-1753-2023.
Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model predictions. In: In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.
Madsen, K.S., Høyer, J.L., Suursaar, Ü., She, J., Knudsen, P., 2019. Sea level trends and variability of the Baltic Sea from 2D statistical reconstruction and altimetry. Front. Earth Sci. 7, 67. https://doi.org/10.3389/feart.2019.00243.
Man, Y., Yang, Q., Shao, J., Wang, G., Bai, L., Xue, Y., 2023. Enhanced LSTM model for daily runoff prediction in the upper Huai River basin, China. Engineering 24, 229–238. https://doi.org/10.1016/j.eng.2021.12.022.
Männikus, R., Soomere, T., Viška, M., 2020. Variations in the mean, seasonal and extreme water level on the Latvian coast, the eastern Baltic Sea, during 1961–2018. Estuar. Coast. Shelf Sci. 245, 106827. https://doi.org/10.1016/j.ecss.2020.106827.
Mostafavi, M., Ellmann, A., Delpeche-Ellmann, N., 2024. Long-term and decadal sea-level trends of the Baltic Sea using along-track satellite altimetry. Remote Sens. (Basel) 16, 760. https://doi.org/10.3390/rs16050760.
Mulia, I.E., Ueda, N., Miyoshi, T., Iwamoto, T., Heidarzadeh, M., 2023. A novel deep learning approach for typhoon-induced storm surge modeling through efficient emulation of wind and pressure fields. Sci. Rep. 13, 7918. https://doi.org/10.1038/s41598-023-35093-9.
Nagaraj, M., Rodriguez, A., Wahl, T., 2025. Regional modeling of storm surges using localized features and transfer learning. Journal of Geophysical Research: Machine Learning and Computation 2 (3), e2025JH000650. https://doi.org/10.1029/2025JH000650.
Okacha, A., Salhi, A., Bouchouou, M., Fattasse, H., 2024. Enhancing flood forecasting accuracy in data-scarce regions through advanced modeling approaches. J. Hydrol. 645 (b), 132283. https://doi.org/10.1016/j.jhydrol.2024.132283.
Pachev, B., Arora, P., Del-Castillo-Negrete, C., Valseth, E., Dawson, C., 2023. A framework for flexible peak storm surge prediction. Coast. Eng. 186, 104406. https://doi.org/10.1016/j.coastaleng.2023.104406.
Park, Y., Kim, E., Choi, Y., Seo, G., Kim, Y., Kim, H., 2022. Storm surge forecasting along Korea Strait using Artificial Neural Network. J Mar Sci Eng 10, 535. https://doi.org/10.3390/jmse10040535.
Pindsoo, K., Soomere, T., 2020. Basin-wide variations in trends in water level maxima in the Baltic Sea. Cont. Shelf Res. 193, 104029. https://doi.org/10.1016/j.csr.2019.104029.
Qin, Y., Su, C., Chu, D., Zhang, J., Song, J., 2023. A review of application of machine learning in storm surge problems. J Mar Sci Eng 11, 1729. https://doi.org/10.3390/jmse11091729.
Rabbani, E.S., Adytia, D., Husrin, S., 2023. Tsunami signal classification based on sea level data using Extreme Gradient Boosting method for tsunami early warning system modeling. In: In: 2023 International Conference on Data Science and Its Applications (ICoDSA). IEEE, pp. 373–378. https://doi.org/10.1109/ICoDSA58501.2023.10276491.
Raj, N., Prakash, R., 2024. Assessment and prediction of significant wave height using hybrid CNN-BiLSTM deep learning model for sustainable wave energy in Australia. Sustainable Horiz. 11, 100098. https://doi.org/10.1016/j.horiz.2024.100098.
Rajabi-Kiasari, S., Hasanlou, M., 2020. An efficient model for the prediction of SMAP sea surface salinity using machine learning approaches in the Persian Gulf. Int. J. Remote Sens. 41 (8), 3221–3242. https://doi.org/10.1080/01431161.2019.1701212.
Rajabi-Kiasari, S., Delpeche-Ellmann, N., Ellmann, A., 2023. Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea. Comput. Geosci. 178, 105406. https://doi.org/10.1016/j.cageo.2023.105406.
Rajabi-Kiasari, S., Ellmann, A., Delpeche-Ellmann, N., 2025. Sea level forecasting using deep recurrent neural networks with high-resolution hydrodynamic model. Appl. Ocean Res. 157, 104496. https://doi.org/10.1016/j.apor.2025.104496.
Ramos-Valle, A.N., Curchitser, E.N., Bruyère, C.L., McOwen, S., 2021. Implementation of an Artificial Neural Network for storm surge forecasting. J. Geophys. Res. Atmos. 126. https://doi.org/10.1029/2020JD033266.
Rutgersson, A., Kjellström, E., 2022. Natural hazards and extreme events in the Baltic Sea region. Earth Syst. Dyn. 13, 251–301. https://doi.org/10.5194/esd-13-251-2022.
Soomere, T., Pindsoo, K., Bishop, S.R., Käärd, A., Valdmann, A., 2013. Mapping wave set-up near a complex geometric urban coastline. Nat. Hazards Earth Syst. Sci. 13 (11), 3049–3061. https://doi.org/10.5194/nhess-13-3049-2013.
Soomere, T., Eelsalu, M., Pindsoo, K., 2018. Variations in parameters of extreme value distributions of water level along the eastern Baltic Sea coast. Estuar. Coast. Shelf Sci. 215, 59–68. https://doi.org/10.1016/j.ecss.2018.10.010.
Soomere, T., Pindsoo, K., Kudryavtseva, N., Eelsalu, M., 2020. Variability of distributions of wave set-up heights along a shoreline with complicated geometry. Ocean Sci. 16, 1047–1065. https://doi.org/10.5194/os-16-1047-2020.
Soomere, T., Eelsalu, M., Viigand, K., Giudici, A., 2024. Linking changes in the directional distribution of moderate and strong winds with changes in wave properties in the eastern Baltic proper. J. Coast. Res. Special Issue 113, 190–194. https://doi.org/10.2112/JCR-SI113-038.1.
Sreeraj, P., Swapna, P., Singh, M., Krishnan, R., 2025. Improved storm Surge Prediction and Extreme Sea Level Future Projections in the Indian Ocean using Deep Learning. Environ. Res. Lett. 20, 084058. https://doi.org/10.1088/1748-9326/ade9e0.
Sun, K., Pan, J., 2023. Model of storm surge maximum water level increase in a coastal area using ensemble machine learning and explicable algorithm. Earth Space Sci. 10 (12), e2023EA003243. https://doi.org/10.1029/2023EA003243.
Sztobryn, M., 2003. Forecast of storm surge by means of artificial neural network. J. Sea Res. 49, 317–322. https://doi.org/10.1016/S1385-1101(03)00024-8.
Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., Ward, P.J., 2021. Exploring deep learning capabilities for surge predictions in coastal areas. Sci. Rep. 11, 17224. https://doi.org/10.1038/s41598-021-96674-0.
Liebsch, G., Schwabe, J., Varbla, S., Ågren, J., Teitsson, H., Ellmann, A., Forsberg, R., Strykowski, G., Bilker-Koivula, M., Liepiņš, I., Paršeliūnas, E., Keller, K., Vestøl, O., Omang, O., Kaminskis, J.N., Wilde-Piórko, M., Pyrchla, K., Olsson, P.-A., Förste, C., Ince, E.S., Somla, J., Westfeld, P., Hammarklint, T., 2023. Release note for the BSCD2000 height transformation grid. The International Hydrographic Review 29 (2), 194–199. https://doi.org/10.58440/ihr-29-2-n11Liebsch.
The WAMDI Group, 1988. The WAM Model - A third generation ocean wave prediction model. Journal of Physical Oceanography, 18, 1775–1810. Doi: 10.1175/1520-0485(1988)018<1775:TWMTGO>2.0.CO;2.
Umesh, P.A., Swain, J., Balchand, A.N., 2018. Inter-comparison of WAM and WAVEWATCH-III in the North Indian Ocean using ERA-40 and QuikSCAT/NCEP blended winds. Ocean Eng. 164, 298–321. https://doi.org/10.1016/j.oceaneng.2018.06.053.
Viigand, K., Eelsalu, M., Soomere, T., 2025. Inherent non-stationarity in the GEV distribution for extreme sea levels: Implications for coastal vulnerability in the Baltic Sea. Ocean Eng. 341 (4), 122681. https://doi.org/10.1016/j.oceaneng.2025.122681.
Wang, B., Liu, S., Wang, B., Wu, W., Wang, J., Shen, D., 2021. Multi-step ahead short-term predictions of storm surge level using CNN and LSTM network. Acta Oceanol. Sin. 40, 104–118. https://doi.org/10.1007/s13131-021-1763-9.
Watson, P.A.G., 2022. Machine learning applications for weather and climate need greater focus on extremes. Environ. Res. Lett. 17, 111004. https://doi.org/10.1088/1748-9326/ac9d4e.
Wei, Z., Nguyen, H.C., 2022. Storm surge forecast using an encoder–decoder recurrent neural network model. J Mar Sci Eng 10, 1980. https://doi.org/10.3390/jmse10121980.
Weisse, R., Dailidienė, I., Hünicke, 2021. Sea level dynamics and coastal erosion in the Baltic Sea region. Earth Syst. Dyn. 12, 871–898. https://doi.org/10.5194/esd-12-871-2021.
Wiśniewski, B., Wolski, T., 2011. Physical aspects of extreme storm surges and falls on the polish coast. Oceanologia 53, 373–390.
Wolski, T., Wiśniewski, B., 2021. Characteristics and Long-Term Variability of Occurrences of storm Surges in the Baltic Sea. Atmosphere (basel) 12, 1679. https://doi.org/10.3390/atmos12121679.
Wolski, T., Giza, A., Wiśniewski, B., 2025. Application of the probability of extreme sea levels at selected Baltic Sea tide gauge stations. Water 17 (3), 291. https://doi.org/10.3390/w17030291.
Xie, W., Xu, G., Zhang, H., Dong, C., 2023. Developing a deep learning-based storm surge forecasting model. Ocean Model. 182, 102179. https://doi.org/10.1016/j.ocemod.2023.102179.
Xu, K., Han, Z., Xu, H., Bin, L., 2023. Rapid prediction model for urban floods based on a Light Gradient Boosting Machine approach and hydrological-hydraulic model.

International Journal of Disaster Risk Science 14, 79–97. https://doi.org/10.1007/s13753-023-00465-2.
Yang, Y., Xiong, Q., Wu, C., Zou, Q., Yu, Y., Yi, H., Gao, M., 2021. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. Environ. Sci. Pollut. Res. 28, 55129–55139. https://doi.org/10.1007/s11356-021-14687-8.
Zhang, Q., Li, P., Ren, X., Ning, J., Li, J., Liu, C., Wang, Y., Wang, G., 2023. A new real-time groundwater level forecasting strategy: Coupling hybrid data-driven models with remote sensing data. J Hydrol (amst) 625, 129962. https://doi.org/10.1016/j.jhydrol.2023.129962.
Zhou, R., Zhang, Y., 2023. Linear and nonlinear ensemble deep learning models for karst spring discharge forecasting. J Hydrol (amst) 627, 130394. https://doi.org/10.1016/j.jhydrol.2023.130394.

Articles in peer reviewed journals indexed in reputable databases (ETIS 1.1)

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A.; Soomere, T. (2026). Forecasting Sea Level Maxima using Machine Learning with Explainability and Extreme Value Analysis. International Journal of Applied Earth Observation and Geoinformation, 146, art. 105064. DOI: 10.1016/j.jag.2025.105064.

Rajabi-Kiasari, S.; Ellmann, A.; Delpeche-Ellmann, N. (2025). Sea level Forecasting using Deep Recurrent Neural Networks with High-Resolution Hydrodynamic Model. Applied Ocean Research, 157, #104496. DOI: 10.1016/j.apor.2025.104496.open access

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2023). Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea. Computers & Geosciences, 178, #105406. DOI: 10.1016/j.cageo.2023.105406.

Rajabi-Kiasari, S.; Hasanlou, M. (2020). An efficient model for the prediction of SMAP sea surface salinity using machine learning approaches in the Persian Gulf. International Journal of Remote Sensing, 41 (8), 3221−3242. DOI: 10.1080/01431161.2019.1701212.

Conference papers (ETIS 3.4)

Nikraftar, Z.; **Rajabi-Kiasari, S.**; Seydi, S. T. (2019). Genetic algorithm based feature selection for landslide susceptibility mapping in northern Iran. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, 42. ISPRS, 821−825. DOI: 10.5194/isprs-archives-XLII-4-W18-821-2019.open access

Rajabi-Kiasari, S.; Hasanlou, M.; Safari, A. R. (2017). Spatial and temporal analysis of sea surface salinity using satellite imagery in gulf of mexico. ISPRS Int Joint Conf of 2nd Conf on Geospatial Information Research GI Research, 4th Conf on Sensors and Models in Photogrammetry and Remote Sensing (SMPR), 6th Conf on Earth Observation of Environmental Changes (EOEC), OCT 07-10, 2017, Tehran, IRAN. Intl Soc Photogrammetry & Remote Sensing-Isprs, 219−223. (International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; 42-4). DOI: 10.5194/isprs-archives-XLII-4-W4-219-2017.open access

Conference abstracts (ETIS 5.2)

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2025). Conditional Diffusion Model for Sea level Prediction using SWOT Satellite Altimetry with Uncertainty Quantification. IAG Scientific Assembly 2025: Geodesy for a changing environment: Abstract Book: Rimini, Italy, September 1-5, 2025. Bologna: University of Bologna, 325−325.

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2025). Instantaneous Sea Surface Height Prediction Using Satellite Altimetry and Deep learning. lps25.esa.in: Living Planet Symposium (LPS25), Vienna, Austria, June 23-27, 2025. European Space Agency. (Session A.07.05: Monitoring and predicting surface water and flood dynamics).

Delpeche-Ellmann, N.; **Rajabi-Kiasari, S.**; Soomere, T.; Ellmann, A. (2025). Forecasting of Sea Level Extremes using Deep Learning and Extreme Value Analysis. Forecasting of Sea Level Extremes using Deep Learning and Extreme Value Analysis: EGU General Assembly 2025, Vienna, Austria & Online, 27 April–2 May 2025. Copernicus Meetings, #EGU25-13350. DOI: 10.5194/egusphere-egu25-13350.

Ellmann, A.; Delpeche-Ellmann, N.; Varbla, S.; **Rajabi Kiasari, S.**; Jahanmard, V.; Kupavõh, A. (2025). Development of continuous dynamic vertical reference for maritime and offshore engineering by applying geodetic and machine learning strategies. EGU General Assembly 2025: EGU General Assembly 2025, Vienna, Austria, April 27-May 2, 2025. Copernicus Meetings, #EGU25-20596. DOI: 10.5194/egusphere-egu25-20596.

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2025). Integrating Satellite Altimetry and Deep Learning for Enhanced Sea Level Forecasting. EGU General Assembly 2025: EGU General Assembly 2025, Vienna, Austria & Online, 27 April–2 May 2025. European Geosciences Union, #EGU25-9951. DOI: 10.5194/egusphere-egu25-9951.

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2025). Instantaneous Sea Surface Height Prediction Using Satellite Altimetry and Deep learning. lps25.esa.in: Living Planet Symposium (LPS25), Vienna, Austria, June 23-27, 2025. European Space Agency. (Session A.07.05: Monitoring and predicting surface water and flood dynamics).

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2025). Conditional Diffusion Model for Sea level Prediction using SWOT Satellite Altimetry with Uncertainty Quantification. IAG Scientific Assembly 2025: Geodesy for a changing environment: Abstract Book: Rimini, Italy, September 1-5, 2025. Bologna: University of Bologna, 325−325.

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2024). Forecasting Sea Level Maxima Using Machine Learning Models in the Baltic Sea. International Liège Colloquium 2024: Ocean Extremes: 55TH INTERNATIONAL LIÈGE COLLOQUIUM ON OCEAN DYNAMICS, 27 TO 31 MAY 2024. Liege, Belgium: University of Liège.

Delpeche-Ellmann, N.; **Rajabi Kiasari, S.**; Ellmann, A. (2023). Forecasting Of Absolute Dynamic Topography by Utilizing Machine Learning with Synergy of Satellite Altimetry Data. 13th Coastal Altimetry Workshop & Coastal Altimetry Training, 6-10 February 2023, Universidad de Cádiz, Spain: Abstract book: 13th Coastal Altimetry Workshop - Coastal Altimetry Training, 6-10 February 2023, Universidad de Cádiz, Spain. ESA, 42−42.

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2023). A Data-Fusion Technique for Forecasting of Absolute Sea Levels in the Baltic Sea. Machine learning and data analysis in oceanography: 54TH INTERNATIONAL LIÈGE COLLOQUIUM ON OCEAN DYNAMICS MACHINE LEARNING AND DATA ANALYSIS IN OCEANOGRAPHY - 8 TO 12 MAY 2023. Liege, Belgium: Université de Liège.

Rajabi Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2023). A multivariate-multistep-ahead forecasting of dynamic topography using convolutional encoder-decoder network in the Baltic Sea: XXVIII General Assembly of the International Union of Geodesy and Geophysics (IUGG), Berlin, Germany, 11-17 July 2023. Potsdam: GFZ German Research Centre for Geosciences. DOI: 10.57757/IUGG23-1664.

Rajabi-Kiasari, S.; Ellmann, A.; Delpeche-Ellmann, N.; Ghorbani Afzal, F. (2022). Evaluation of SMAP and Sentinel-2 Sea Surface Salinity measurements in the Baltic Sea. The electronical abstract book, session E3.04 Baltic Sea Regional Applications and Science: Living Planet Symposium (LPS22), Bonn, Germany, May 23-27, 2022. European Space Agency.

Rajabi-Kiasari, S.; Delpeche-Ellmann, N.; Ellmann, A. (2022). Machine learning Forecasting of Absolute Dynamic Topography in the Baltic Sea. Nordic Geodetic Commission General Assembly: Planet Ocean and Geodesy, Copenhagen, Denmark, September 5-8, 2022. Nordic Geodetic Commission.

Mostafavi, M.; **Rajabi Kiasari, S.**; Jahanmard, V.; Delpeche-Ellmann, N.; Ellmann, A. (2022). Reconstruction of Dynamic Topography Using Cyclostationary Empirical Orthogonal Functions in the Baltic Sea. Session: Geodynamics and Earth Observation: Nordic Geodetic Commission General Assembly: Planet Ocean and Geodesy, Copenhagen, Denmark, September 5-8, 2022. Nordic Geodetic Commission, 25−25.

Mostafavi, M.; Jahanmard, V.; **Rajabi Kiasari, S.**; Delpeche-Ellmann, N.; Ellmann, A. (2022). Absolute Sea Level Trend Forecasting using an Ensemble Empirical Mode Decomposition Method for Satellite Altimetry data. Nordic Geodetic Commission General Assembly: Planet Ocean and Geodesy, Copenhagen, Denmark, September 5-8, 2022. Session: 1. Planet Ocean and Geodesy: Nordic Geodetic Commission, 26−26.

Curriculum vitae

Personal data

Name:	Saeed Rajabi Kiasari
Date of birth:	26.04.1990
Place of birth:	Sari, Iran
Citizenship:	Iranian

Contact data

E-mail:	saeed.rajabi@taltech.ee

Education

2021–2026	Tallinn University of Technology, PhD
2013–2016	University of Tehran, IRN, MSc
2008–2013	University of Tehran, IRN, BSc

Language competence

English	Fluent
Persian	Native
Estonian	Working Proficiency
German	Elementary Proficiency
Turkish	Elementary Proficiency

Elulookirjeldus

Isikuandmed

Nimi:	Saeed Rajabi Kiasari
Sünniaeg:	26.04.1990
Sünnikoht:	Sari, Iraan
Kodakondsus:	Iraani

Kontaktandmed

E-post:	saeed.rajabi@taltech.ee

Hariduskäik

2021–2026	Tallinna Tehnikaülikool, PhD
2013–2016	University of Tehran, MSc
2008–2013	University of Tehran, BSc

Keelteoskus

Inglise keel	kõrgtase
Pärsia keel	emakeel
Eesti keel	algtase
Saksa keel	esmatasand
Türgi keel	esmatasand

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-478-0 (PDF)