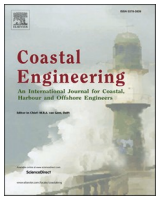
Coastal Engineering 200 (2025) 104766

Contents lists available at ScienceDirect

Coastal Engineering

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



Assessing the role of a probabilistic model for guiding storm surge barrier maintenance

Sunke Trace-Kleeberg [a,*], Krijn Saman [b], Robert Vos [b], Elja Huibregtse [b], Ivan D. Haigh [a], Marc Walraven [b], Annette Zijderveld [b], Susan Gourvenec [c]

[a] School of Ocean and Earth Science, University of Southampton, European Way, Southampton, SO14 3ZH, UK
[b] Ministry of Infrastructure and Water Management, Rijkswaterstaat, Rotterdam, Netherlands
[c] School of Engineering, University of Southampton, Southampton, SO16 7BQ, UK

ARTICLE INFO

Keywords:
Storm surge barriers
Management
Maintenance and operation
Probabilistic model
Decision support system
Ensemble forecasting

ABSTRACT

Storm surge barriers provide flood protection to many major coastal cities in estuaries around the world. Maintenance of these assets is critical to ensure they remain reliable and continue to comply with national legal protection standards. There are often critical thresholds of environmental conditions, beyond which maintenance work is unsafe to be carried out. However, as storm surge barriers age and with climate change effects such as sea-level rise and possible changes in storminess, periods when environmental conditions exceed set thresholds will occur more frequently, so carrying out the required work in available maintenance windows will become increasingly challenging. Probabilistic models enable the use of ensemble forecasts of upcoming water levels to determine the likelihood of conditions exceeding the threshold and so can inform on decision making regarding maintenance. This paper evaluates a probabilistic model currently in operational use by Rijkswaterstaat, the Dutch Ministry of Infrastructure and Water Management, to guide maintenance decisions at the Maeslant barrier in the Netherlands. Sixteen years of historic highwater level forecasts from a combination of European Centre for Medium-Range Weather Forecasts and Dutch Continental Shelf Model v5 are used with observations from the Hoek van Holland tide gauge to evaluate and sensitivity test the probabilistic model. Binary classification is used to assess the performance of the probabilistic model. Findings show that the model is conservative with 33.1 % of outcomes resulting in a False Alarm. Changing the baseline parameters of critical probability and water level threshold impacts the balance between False Alarm and Miss outcomes. Increasing the critical probability reduces the number of False Alarms but increases the Miss situations, emphasising the trade-off between acceptable risk and time available to carry out maintenance work. This study highlights the delicate balance between model parameter selection and the associated risk with respect to the maintenance of storm surge barriers.

1. Introduction

Many coastal cities and towns are located in estuaries which are particularly vulnerable to natural disasters such as flooding, due to their location at the interface of the sea and rivers. The impacts of climate change, such as accelerating sea-level rise and increases in storminess, along with changes in rainfall and river discharge, are increasing the risk of damage to infrastructure and loss of life in low-lying coastal regions (Brown et al., 2013; Del-Rosal-Salido et al., 2021; Hinkel et al., 2014).

In estuaries with long exposed coastlines, where space is limited, storm surge barriers can provide a technical and economic solution for new and/or improved flood protection measures (Aerts et al., 2014; Jonkman et al., 2013; Kirshen et al., 2020). Storm surge barriers are hard engineering structures consisting of partly or fully movable gates (Mooyaart and Jonkman, 2017), located in an estuary, river (Mooyaart et al., 2014) or lake (Deltares, 2018). These flood defences can be closed temporarily to prevent extreme water levels from propagating inland and thus protect the hinterland from flooding (Zhong et al., 2012). Presently, there are over 50 storm surge barriers in operation worldwide (Smaling, 2024), including: the Hollandsche IJssel (Hamerslag and Bakker, 2023), Eastern Scheldt (Knoester et al., 1984) and Maeslant (Bol, 2005) barriers in the Netherlands; the Thames (Wilkes and Lavery, 2005) and Boston (ICE, 2020) barriers in the UK; the MOSE barrier in Italy (Munaretto et al., 2012); New Bedford Hurricane protection barrier

* Corresponding author.
E-mail address: s.trace-kleeberg@soton.ac.uk (S. Trace-Kleeberg).

https://doi.org/10.1016/j.coastaleng.2025.104766
Received 9 July 2024; Received in revised form 23 April 2025; Accepted 23 April 2025
Available online 25 April 2025

in Massachusetts (US Army Corps of Engineers, 2015) and the Lake Borgne Surge Barrier which is part of the Hurricane Storm Damage Risk Reduction System in New Orleans, USA (Flood Protection Authority East, 2014). In addition to the operational barriers, many new schemes are being planned or constructed such as in Galveston (Merrell et al., 2011), New York and New Jersey (Kluijver et al., 2019) in the US; Bridgwater in the UK (Somerset Council, 2024) and Nieuwpoort in Belgium (Jan De Nul, 2024).

Like other engineering structures, storm surge barriers require specialist management, maintenance and operation due to their complexity (Walraven et al., 2022). Regular maintenance is required to ensure storm surge barriers are reliable, remain functional and comply with national legal protection requirements. Maintenance is highly varied involving test closures, inspections, repair and replacement of key parts, along with major upgrades of systems. These maintenance tasks can require substantial time to complete, and the more often a barrier closes, the more inspection and maintenance work needs to be carried out to ensure the barrier remains reliable. In the UK, for example, maintenance is carried out year-round. Whereas, in other counties, like the Netherlands, maintenance is currently completed from the middle of April to the end of September during the so-called maintenance season. However, this puts pressure on the teams carrying out the work, as a large number and wide range of tasks need to be completed in a limited time window. Furthermore, maintenance can only be carried out when conditions are safe to ensure the safety of the workers at the barrier. Therefore, thresholds of environmental conditions such as water level are defined beyond which maintenance work cannot be carried out. These health and safety thresholds vary among barriers. At the Maeslant barrier in the Netherlands, which is located 6 km from the coast, maintenance work is required to stop when the water level reaches a certain height, as elaborated in Section 2.1. At the Eastern Scheldt barrier in the Netherlands, which is located on an exposed coastline, maintenance work is stopped when water level or the combination of water level and significant wave height exceed defined thresholds. In London at the Thames barrier maintenance work is impacted by river discharge as well as water level (Haigh et al., 2024).

Completing the required maintenance for storm surge barriers is getting increasingly challenging (Walraven et al., 2022), due to a number of key reasons. Firstly, existing barriers are ageing. Some, barriers were constructed over 40 years ago, meaning they require additional maintenance to continue being reliable. Secondly, over time, sea-level rise will result in barriers having to close more often (Haasnoot et al., 2018), at some locations possibly exacerbated by increased storminess and changes in river discharge (Chen et al., 2020). Closure of the Thames Barrier in May 2020, where previously closures had not occurred later than March, is an illustration that climate change may already have an influence on the traditional closure season at this barrier. As time progresses, storm surge barriers will need to close more often and increasingly in summer months (Haigh et al., 2024), impacting planned maintenance projects. Thirdly, climate change will not only continue to influence the number of closures but will also increase the number of times maintenance thresholds are reached, interrupting work; especially as maintenance thresholds will typically be lower than closure thresholds resulting in these being exceeded more often. Such that the influence of sea-level rise will be noticed more for maintenance thresholds than closure levels. A detailed assessment of past and likely future maintenance threshold exceedances (under different climate scenarios) was undertaken for the Maeslant barrier (Trace-Kleeberg et al., 2023). Findings showed that of the past maintenance threshold exceedances, 13 % occurred during the maintenance season which could have interrupted the planned maintenance work (Trace-Kleeberg et al., 2023). As sea-level rise increases mean sea level, a point will be reached when the highest astronomic tides alone exceed the maintenance threshold, without a meteorological contribution, decreasing the availability of safe working windows (Trace-Kleeberg et al., 2023).

As it is getting harder to carry out the required maintenance, investigations are beginning into new ways to maximise the available time to complete maintenance work. For example, Dutch governmental departments are exploring means of carrying out maintenance work year-round by considering more extensive use of operational forecasts of environmental conditions. To do this, a thorough risk analysis of the consequences of wrong predictions is required besides mitigation measures. For instance, adequate precautions need to be in place to ensure the safety of workers and to guarantee that the barrier will be ready for an operational closure, if called upon (Trace-Kleeberg et al., 2023). This requires accurate forecasting of upcoming water levels to manage, maintain and operate the storm surge barriers.

At many storm surge barriers, water level forecasts are used to determine when they are required to close. These forecasts could also be used to aid maintenance decision making of when upcoming conditions are safe. At the Maeslant storm surge barrier in the Netherlands, staff have recently developed a probabilistic model based on ensemble water level forecasts, to provide guidance in the short term (next 7 days) to aid decision making whether or not it is safe to do maintenance. The developed model has - up to this point - not been evaluated in depth to test, and identify possible adjustments, to maximise its performance. Thus, this paper explores the use of forecasts to guide when maintenance can be safely carried out. A detailed evaluation of the probabilistic model using historic forecasts over the 16-year period from 2008 to 2023 is conducted. Three specific objectives are defined: (1) evaluate the performance of the probabilistic model using the existing model parameters; (2) conduct sensitivity tests to explore how adjusted parameters affect model performance; and (3) analyse instances when the probabilistic model outcome is incorrect.

This paper is structured as follows. The case study barrier and probabilistic model are described in Section 2, followed by the method used for evaluating the model baseline, sensitivity testing and analysing incorrect model outcomes which are outlined in Section 3. Section 4 presents the results of the three paper objectives. These are discussed in Section 5 and the paper finishes with conclusions in Section 6.

2. Background to case study barrier and model

This paper uses the Maeslant storm surge barrier in the Netherlands as a case study. The following section provides an overview of the Maeslant barrier and maintenance thresholds (Section 2.1) and describes the probabilistic model that has been developed to aid decision making regarding maintenance (Section 2.2).

2.1. Overview of the Maeslant barrier and maintenance threshold

The Maeslant barrier is located approximately 30 km west of Rotterdam city in the Netherlands (Fig. 1a). It consists of two horizontal sector gates (Fig. 1b). Three water level thresholds are important for barrier management, maintenance and operation. The gates close when the forecast water level in central Rotterdam exceeds 300 cm above NAP or the forecast water level at Dordrecht exceeds 290 cm above NAP (Dutch: Normaal Amsterdams Peil, NAP; Amsterdam's Ordnance Datum). This happened for the first time in December 2023 (Zijderveld et al., 2024). At a forecast water level of 260 cm above NAP, the operational team is called onsite. The Maeslant barrier is operated by a computer system which switches state from "at rest" to "operational" when predicted water levels exceed 230 cm above NAP.

In addition to the thresholds mentioned above, a health and safety threshold exists. When water levels at the barrier reach 170 cm above NAP, maintenance work is stopped. This is due to the risk posed by waves and wakes from passing vessels, which reach the lowered terrain between the ball joint and barrier gate impeding access to the structure, and water overtopping the dock doors making work in the docks dangerous (Fig. 1c).

The maintenance season at the Maeslant Barrier is currently between the 15th of April and the 30th of September, with the storm season in the

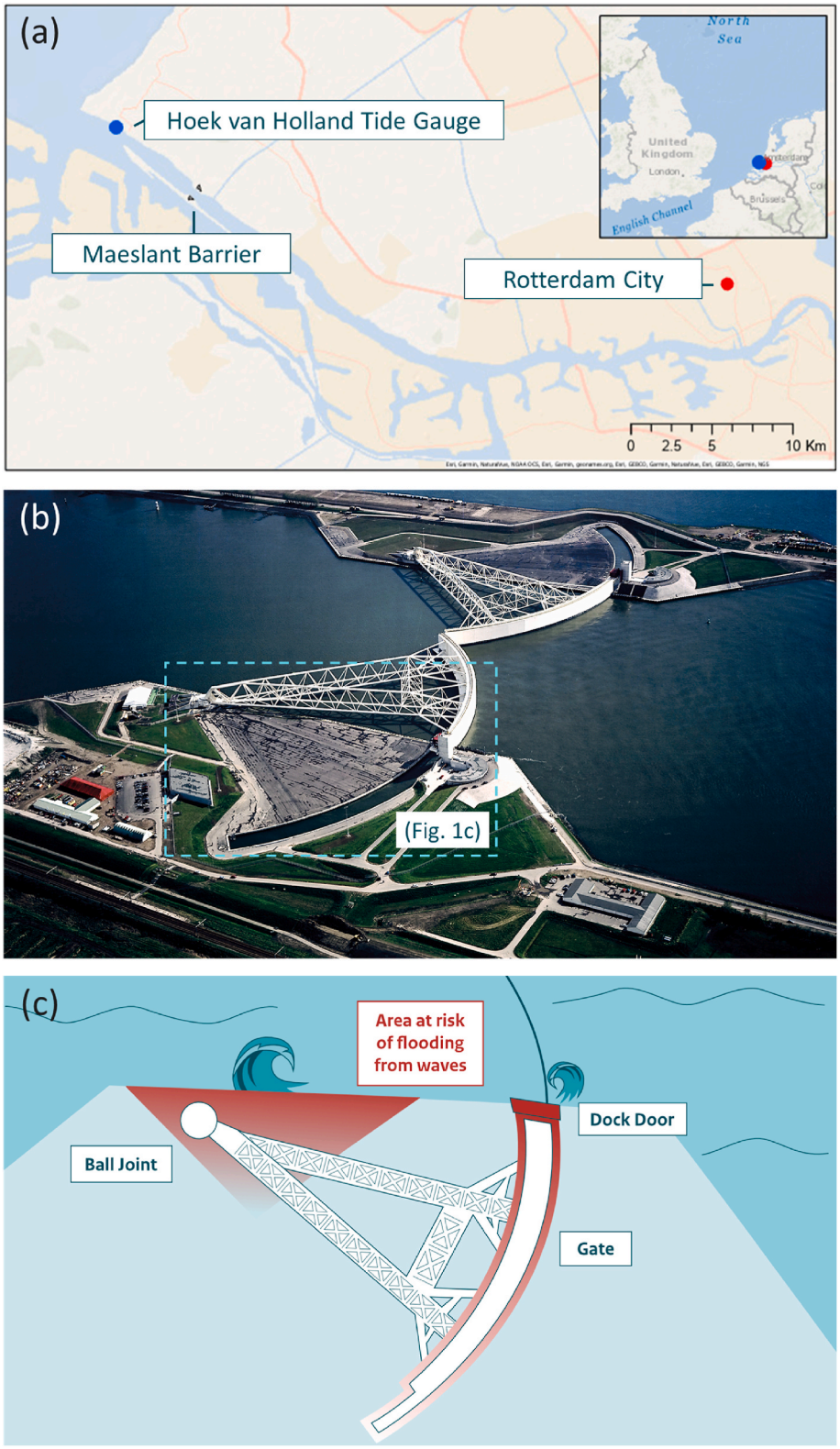


Fig. 1. (a) Overview of case study area indicating location of tide gauge, Maeslant barrier and Rotterdam city (b) Arial view of Maeslant barrier. Dotted box indicates region represented in panel c. (Image credits: Rijkswaterstaat, 2022) (c) Schematic illustration of Maeslant barrier and regions at risk of flooding at maintenance threshold of 170 cm.

remaining months. Inspections and the annual test closure before the start of storm season effectively shorten the actual duration available for maintenance by a month. Currently, maintenance works are not undertaken at the Maeslant barrier during the storm season. However, analysis has shown that water levels in past storm seasons have been below the maintenance threshold for extended periods, meaning work could have been carried out (Trace-Kleeberg et al., 2023). Maintenance jobs vary in frequency and duration from short daily tasks to infrequent

replacement that can take up to four successive seasons to complete. Maintenance planning and delivery is a challenge due to the quantity and complexity of the work and short time-period when work is carried out. This will be exacerbated by sea-level rise, which reduces the safe weather windows as the number of maintenance threshold exceedances per year increase (Trace-Kleeberg et al., 2023). Therefore, it is important to utilise all potentially available maintenance windows including possibly undertaking maintenance work all year round. However, to achieve this, adequate precautions need to be in place to ensure the safety of workers and to guarantee that the barrier will be ready if needed for an operational closure. One way of doing this is to use forecasts to predict when upcoming water levels are likely to be lower than the maintenance threshold, accounting for uncertainty. Such a system can aid decision making of when maintenance work can safely be carried out in relation to the criteria considered.

2.2. Probabilistic model

Conventional forecasting systems produce a single, deterministic forecast, which is subject to two main sources of error: (1) uncertainty in initial conditions and (2) approximation of processes in the models (Buizza, 2006). These sources of uncertainties limit the skill of deterministic forecasts in an unpredictable way. An ensemble prediction technique addresses these issues by producing several forecasts over the same period, allowing quantitative estimates of uncertainty (Flowerdew et al., 2010). The combination of models used in this study determine the probability that upcoming water levels exceed certain thresholds. This output is used to guide decision making for whether it is safe or not to carry out short term maintenance at a storm surge barrier.

To reach a decision multiple steps are needed, these are outline below and shown in Fig. 2.

1. A 50-member ensemble of meteorological conditions is computed by the medium range Ensemble Prediction System (EPS) at the European Centre for Medium-range Weather Forecasts (ECMWF), with a ten-day lead time (Buizza, 2006). This model has undergone multiple updates which are described in the annual technical memorandums (e.g., ECMWF Tech. Memos. 880, 884 and 902).
2. The 10-m wind and mean sea level pressure ensembles from the ECMWF EPS are used to drive the depth-averaged hydrodynamic Dutch Continental Shelf Model (DCSMv5) (Gerritsen et al., 1995; Zijl et al., 2013). This model produces fifty forecasts of surge height for the coming seven days (Irazoqui Apecechea, 2018). To obtain forecasts of total water level, calculated astronomical tide is needed to add to the forecast surge height. The astronomic tides are calculated by running the DCSMv5 model without wind forcing.
3. Total high waters taken from the forecasted time series, are used as input to the newly developed probabilistic model named 4SVK. This model calculates the probability the different threshold levels explained in section 2.1 are exceeded.
4. The model outcome is used by staff at the barrier to guide their decision making for whether the upcoming water levels are safe for maintenance work to be carried out or not.

Some parts of the operational system use different time references, this has been accounted for to ensure correct comparison. Note, new versions of the meteorological and hydrodynamic models are now available, that have replaced the ECMWF EPS/DCSMv5 forecasting

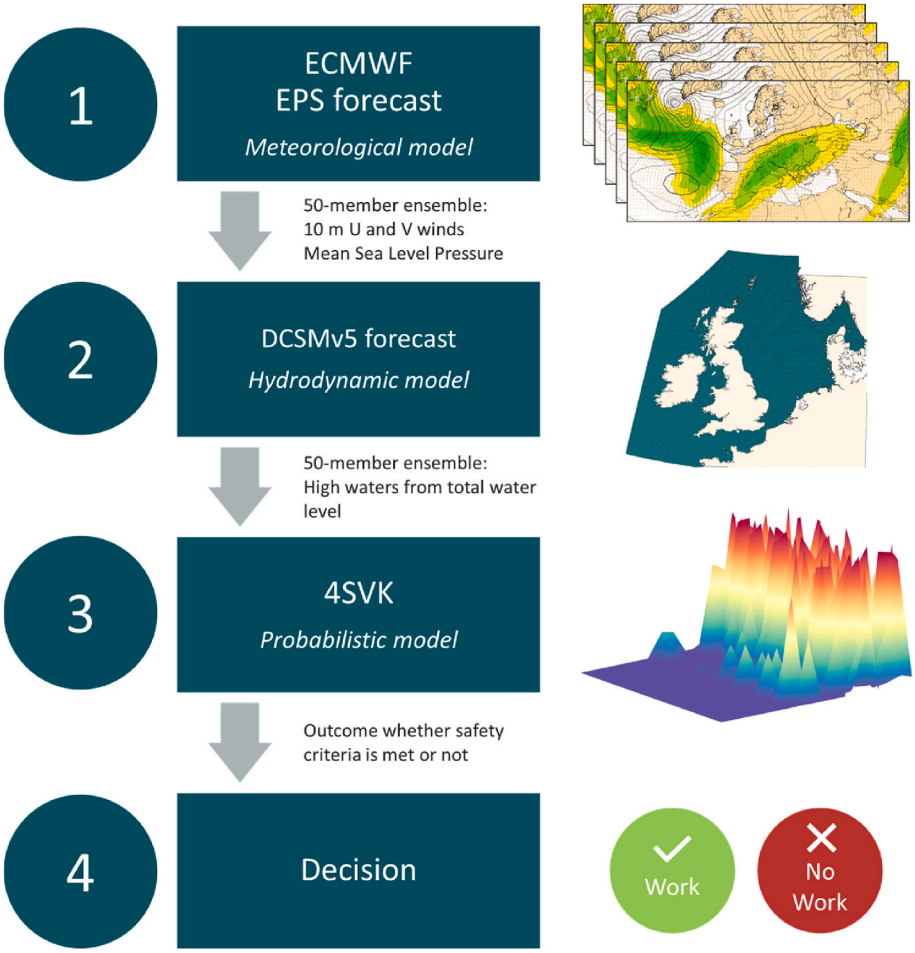


Fig. 2. Overview of forecast models and the process steps to reach a decision on upcoming short-term maintenance at storm surge barrier.

combination. However, the older version is used in this study as 16-years of past data are available, which is not yet the case for the new model combination.

Step three, outlined above, is a novel probabilistic model. To produce an outcome on the risk associated with upcoming water levels, four calculation steps are needed. These are as follows.

1. The 50-member ensemble forecast of upcoming highwater levels from the combination of ECMWF and DCSMv5 models (Fig. 3a) are used as input to the probabilistic model 4SVK. For the following equations this is generalised to N members where each member has NT high waters. Therefore, each high water is denoted as $H_{n,it}$ where $n = 1 \ldots N$ and $it = 1 \ldots NT$. It is assumed that every member, at each timestep has equal probability of occurring (Buizza et al., 1999; Leutbecher and Palmer, 2008; Stephenson et al., 2005). The assumption that each ensemble member has an equal probability of occurrence is a common simplification used in ensemble-based analyses, this is a conservative assumption. As such, each high water level value ($H_{n,it}$) is defined as a stochastic variable with a Gaussian probability density function and standard deviation, σ. The operational and maintenance water level thresholds introduced earlier are denoted by H_{op}.
2. As a function of both forecast surge height and lead time (Fig. 3b), the standard deviation, $\sigma_{n,it}$, is calculated for each member ($n = 1 \ldots N$) at every timestep ($it = 1 \ldots NT$). If the forecast surge is less than 50 cm then $\sigma_{n,it}$ is set to 10 cm. If the surge (H_{surge}) is greater than 50 cm, then the standard deviation is calculated as $\sigma_{n,it} = 10 + 0.1*(H_{surge}\text{-}50.0)$. The contribution of lead time is that for every successive high water 0.5 cm is added to $\sigma_{n,it}$ such that $\sigma_{n,it} = \sigma_{n,it\text{-}1} + 0.5$. These values are based on verification of the combination of ECMWF and DSCMv5 models (de Vries, 2008, 2009; Wagenaar, 2018; Zijl et al., 2013).
3. The cumulative density function ($P_{n,it}$) is calculated for each ensemble member (n) at each time step (it) from the forecast high water level ($H_{n,it}$) (Fig. 3c) and is given by:

$$P_{n,it} = \varphi(-\beta_{n,it}) \quad \text{(Eq. 1)}$$

Where φ is the notation for the probability of exceeding the standard normal distribution with associated standardised β-factor:

$$\beta_{n,it} = \frac{(H_{n,it} - H_{op})}{s_{n,it}} \quad \text{(Eq. 2)}$$

The cumulative density function is calculated for every value of H_{op}.

4. The resulting probabilities ($P_{n,it}$) of the N members at each timestep (it) are then averaged (Fig. 3d):

$$\overline{P}_{it} = \frac{1}{N}\sum_{n=1}^{N} P_{n,it} \quad \text{(Eq. 3)}$$

To calculate the cumulative exceedance probability for the high waters at each timestep (it) as follows:

$$Pcum_{it} = 1 - \prod_{j=1}^{j=it} (1 - \overline{P}_{it}) \quad \text{(Eq. 4)}$$

Resulting probabilities are compared to a critical probability value (P_{crit}) (Fig. 3d). The safety criterion is met when the cumulative exceedance probability ($Pcum_{it}$) is less than or equal to the critical probability ($Pcum_{it} \leq P_{crit}$). Conversely, the safety criterion is not met when the cumulative exceedance probability is greater than the critical probability ($Pcum_{it} > P_{crit}$).

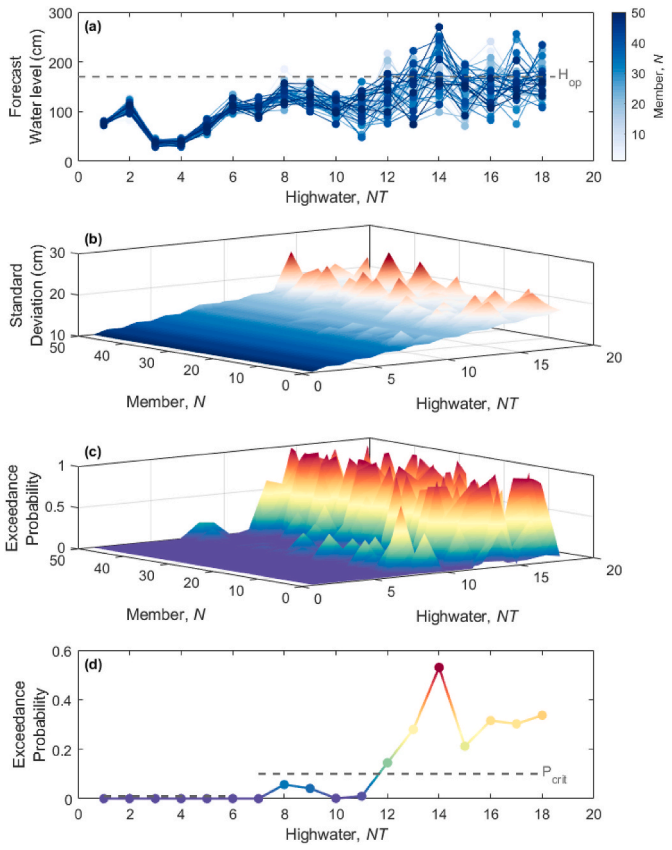


Fig. 3. Overview of probabilistic model (a) Ensemble of future highwater levels with the horizontal line indicating a value of H_{op}. (b) Standard deviation $\sigma_{n,it}$, for each member ($n = 1 \ldots N$) at every timestep ($it = 1 \ldots NT$). (c) Exceedance probability of H_{op} for each member at every timestep (d) Average ensemble cumulative exceedance probability of H_{op} with horizontal lines indicating critical probability values (P_{crit}). Instances where $P_{cum} \leq P_{crit}$ safety criterion is met, while when $P_{cum} > P_{crit}$ the safety criterion is not met.

Depending on the value of H_{op}, the forecasted high waters are split into different forecast horizons to compare calculated probability against a critical probability value (Fig. 4). The model criteria depicted in Fig. 4 correspond to the baseline.

This probabilistic model has been used at the Maeslant barrier to guide decision making in real time since late 2021. The critical probability values illustrated in Fig. 4, were selected by expert judgement, however, to date, no detailed evaluation of the model performance has been undertaken, especially not considering past surge height forecasts. Thus, in this paper, the model is evaluated in detail, as described in the next section.

3. Methodology - model evaluation

The following sections outline the framework developed to evaluate the performance of the probabilistic model (Section 3.1), conduct sensitivity tests (Section 3.2) and analyse any incorrect model outcomes (Section 3.3).

3.1. Baseline evaluation

The first objective is to evaluate the performance of the probabilistic model using historic forecasts between 2008 and 2023. To illustrate the evaluation framework, the results from the model outcomes at H_{op} 170 cm and in the time frame of days 1–3 are presented in this paper.

Results from the combined ECMWF/DCSMv5 ensemble forecasting approach are used to run the probabilistic model in hindcast. Water level

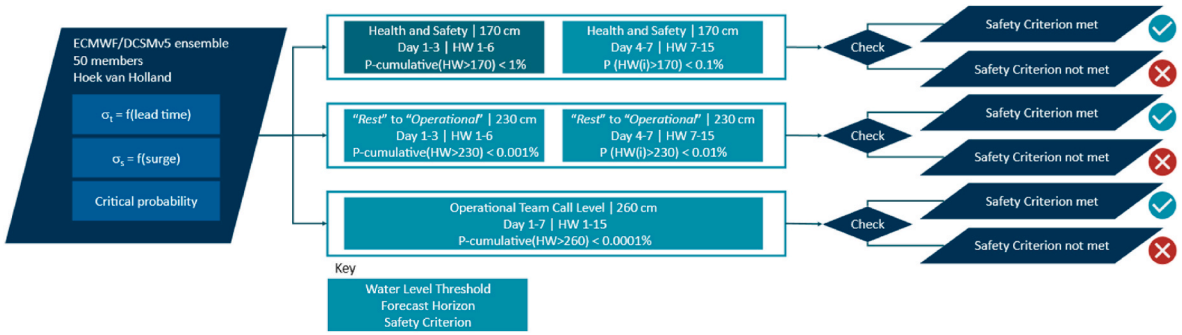


Fig. 4. Schematic overview of input, forecast horizons and output of the probabilistic decision support model (4SVK).

observations from the Hoek van Holland tide gauge (51.9775°N, 4.12°E) are used to assess the outcome of the probabilistic model. The observational data were provided directly by Rijkswaterstaat, at 10-min recording interval.

The total number of model outcomes analysed is 5802. This is less than the number of days over the analysed period because on 42 occasions there were fewer than 50 ensemble members in the forecast or there were no forecasts available on that day. For consistency, these days are omitted from the analysis. Operationally, in instances where a forecast update is missing, the previous forecast remains accessible, ensuring that the system continues to function. The probabilistic model outcome displays the date and time of the last update, allowing users to identify the most recent model run and assess its reliability. The ECMWF forecast models are run twice daily, this study uses the forecast from the midnight run as this run is available at the beginning of the working day and so is used at the barrier for decision making of upcoming maintenance work.

To evaluate the performance of the probabilistic model, binary classification is used, i.e., a framework where models predict one of two mutually exclusive classes, the positive and negative class (Fahmy, 2022; Starovoitov and Golub, 2020). For this probabilistic model, the positive class is water levels are at or above threshold, while the negative class is water levels are below threshold. The outcome of the probabilistic model is compared against observed water levels and the results are used to populate a 2x2 confusion matrix with four possible results: (1) *Correct - True Negative* (TN), (2) *Hit - True Positive* (TP), (3) *False Alarm - False Positive* (FP), and (4) *Miss - False Negative* (FN) (Figure). In this case, C*orrect - True Negatives* correspond to correct model outcomes where the safety criterion is met, and observed water levels were below the threshold (top left quadrant in Fig. 5). While *Hits - True Positives* are also correct model outcomes of safety criterion not being met and observed water levels above the threshold value (bottom right quadrant in Fig. 5). In this way, *False Alarms - False Positives* are defined as incorrect model outcomes where safety criterion was not met but the observed water levels were below the operational threshold (top right quadrant in Fig. 5). These instances mean that a potential maintenance window has been missed as water levels were safe for maintenance work, but the model outcome forecasted otherwise. On the other hand, *Misses - False Negatives* outcomes are when the probabilistic model output that the safety criterion is met, but observations show that water levels were above the threshold (bottom left quadrant in Fig. 5). Therefore, these occasions are "increased risk" situations. To increase the useable intervals when maintenance work is carried out at the barrier, it is important to minimise the number of False Alarms as these restrict maintenance work. On the other hand, it is also important to mitigate the number of Misses as these can pose a risk to health and safety.

Binary classification defines metrics to quantify model performance (Fawcett, 2006). Outlined below are the metrics used in this paper and the equations to calculate them. All metrics applied are expressed as percentages.

The metric *Recall* shows how sensitive the model is towards identifying the true positives in the positive class (Eq. (5)). This is calculated as the number of true positives divided by the sum of true positives and false negatives, as follows:

$$Recall = \frac{TP}{(TP + FN)}\ x\ 100 \qquad \text{(Eq. 5)}$$

This makes it a suitable metric for identifying models with low numbers of misses or "increased risk" situations as high model Recall indicates low miss rates.

Specificity is the ratio of the number of correctly classified negative outcomes (true negatives) to the number of actual negative events (Eq. (6)) which is the sum of true negative and false positives.

$$Specificity = \frac{TN}{(TN + FP)}\ x\ 100 \qquad \text{(Eq. 6)}$$

This metric is useful when trying to limit the number of false alarm or "waste of time" situations, as high model Specificity indicates a low false alarm rate.

Prevalence shows how many of the total outcomes are in the positive class (Eq. (7))), this is calculated as follows:

$$Prevalence = \frac{FN + TP}{(FN + TP + TN + FP)}\ x\ 100 \qquad \text{(Eq. 7)}$$

This indicates the number of instances when observed water levels were above the maintenance threshold.

In classification problems when the number of instances within each class differ by an order of magnitude or more, they are unbalanced (Starovoitov and Golub, 2020). In such cases, *balanced accuracy* shows

		Outcome of model	
		Safety Criterion met	Safety Criterion not met
Reality	Water level < H_{op}	Correct True Negative	False Alarm False Positive (*Waste of time*)
	Water level ≥ H_{op}	Miss False Negative (*Increased risk*)	Hit True Positive

Fig. 5. Binary classification confusion matrix indicating Correct-True Negative, False alarm-False positive, Miss-False Negative and Hit-True Positive outcomes.

the number of correct model outcomes (Eq. (8))). It is calculated by dividing the sum of Recall and Specificity by two:

$$Balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{(TP+FN)} + \frac{TN}{(TN+FP)}\right) x\ 100 \qquad \text{(Eq. 8)}$$

3.2. Sensitivity testing

The second objective is to explore how adjusting parameters effect the model performance. This is done by altering two model parameters: critical probability (P_{crit}) and water level threshold (H_{op}). In the baseline model critical probability (P_{crit}) is set at 1 %. The values tested are increased at 10 % intervals from 10 % to 50 %, resulting in five critical probability sensitivity tests. In the baseline probabilistic model, the values for water level threshold (H_{op}) are 170, 230 and 260 cm. Five alternative water level thresholds are tested ranging from 110 cm to 210 cm at 20 cm intervals. In these sensitivity tests the baseline critical probability value of 1 % is used. Water level threshold values lower than 170 cm provide a proxy for sea-level rise, where a lower criterion applied now represents a smaller water level margin that may be available in the future. The thresholds of 110 cm, 130 cm and 150 cm correspond to a proxy of 60 cm, 40 cm and 20 cm sea-level rise. This could be experienced at Hoek van Holland by 2101, 2072 and 2043 respectively under a sea-level rise scenario of SSP2-4.5 (50th percentile intermediate scenario) or 2083, 2065 and 2042 respectively under a sea-level rise scenario of SSP5-8.5 (50th percentile high scenario) (Fox-Kemper et al., 2021; Van Dorland et al., 2024). Water level thresholds greater than 170 cm allow the impact of increasing the operational water level threshold on the model performance to be assessed. In practice, the operational threshold can either be altered procedurally or through physical alterations at the barrier, this is discussed in more detail in section 5.2.

3.3. Analysis and interpretation of incorrect model outcomes

The third objective is to analyse the occasions when the outcome of the probabilistic model is incorrect. There are two types of incorrect outcomes: (1) False Alarms and (2) Misses. Results of incorrect model outcomes are presented as annual and monthly percentage occurrence. For the False Alarms, percentage occurrence is calculated by determining the proportion of False Alarms from the positive class, while the Misses are out of the total of negative class events.

In addition, the extent to which the False Alarms in the end actually were below the H_{op} threshold (referred to hereon as freeboard) and the amount by which Misses exceeded the threshold (referred to hereon as threshold exceedance) are also shown. Freeboard is calculated by determining the minimum difference between the observed highwaters and threshold H_{op} in the forecast days 1–3. Threshold exceedance is calculated by determining the maximum difference between the day 1–3 observed highwaters and the threshold H_{op}.

4. Results

The following sections present results from the probabilistic model baseline evaluation (Section 4.1), sensitivity testing (Section 4.2) and analysis of incorrect model outcomes (Section 4.3).

4.1. Baseline evaluation

The distribution of outcomes for the baseline model considered is depicted in Fig. 6 (and Supplementary Table 1). Results show that in the baseline configuration (i.e., critical probability of 1 % and water level threshold of 170 cm), 4175 (71.9 %) outcomes matched the status according to the observed water levels. Of which 3284 (78.7 %) accounted for correct outcomes with water levels below threshold and the other 891 (21.3 %) were Hit outcomes when water levels are above threshold. The remaining 1627 (28.0 %) outcomes were incorrect with 1626 False Alarms and 1 Miss.

4.2. Sensitivity testing

Results for the critical probability and water level threshold sensitivity tests are presented in sub-sections 4.2.1 and 4.2.2, respectively.

4.2.1. Critical probability (P_{crit})

The first set of sensitivity tests presented here alter the critical probability (P_{crit}). Results from the model outcome classification and metrics are illustrated in Fig. 6 (and Supplementary Table 2).

Classification of the model outcome shown in Fig. 6a illustrate the shift in distribution of model outcomes depending on the critical probability. Results show that as the critical probability is increased from the baseline of 1 % – 50 % the number of Correct outcomes increase from

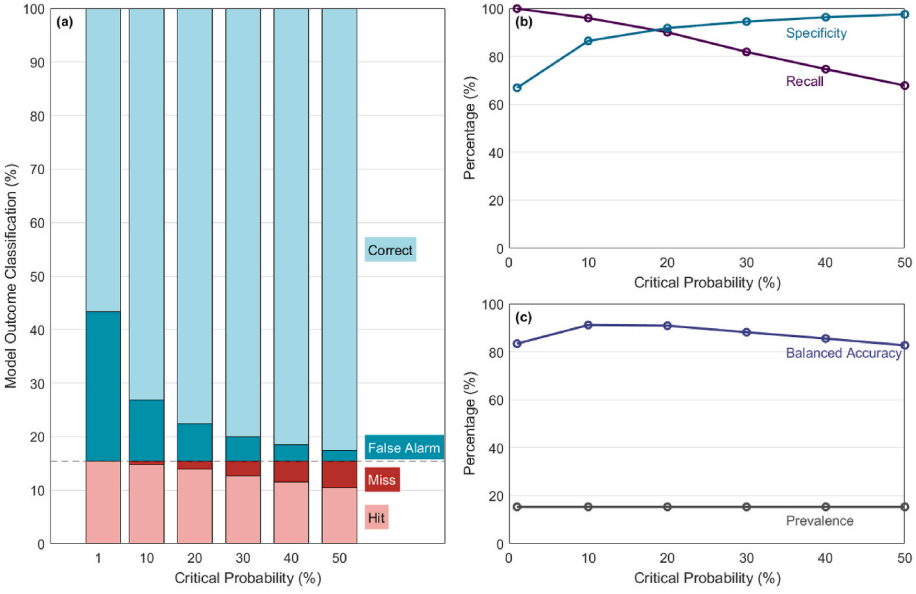


Fig. 6. (a) model outcome classification for critical probability sensitivity tests at H_{op} 170 cm. Dotted horizonal line indicates prevalence. Model metrics for critical probability sensitivity tests (b) Recall and Specificity (c) Balanced Accuracy and Prevalence.

3284 (56.6 %) to 4789 (82.5 %), while the number of Hits decrease from 891 (15.3 %) to 605 (10.4 %). The number of False Alarms decrease from 1626 (28.0 %) to 121 (2.1 %), while the number of Misses increase from 1 (0.017 %) to 287 (4.9 %). The incorrect model outcomes from the critical probability sensitivity tests are analysed in further detail in section 4.3.1. When the critical probability is set to 40 % there are a greater number of Misses than False Alarms. The ratio of Correct and False Alarms (light and dark blue) to Misses and Hits (light and dark red) is constant at 15.3 % in each sensitivity test as shown in Fig. 6a. This is because the positive and negative class totals are dependent on the observed water levels which remain the same in each of the sensitivity tests as shown by the constant Prevalence value of 15.4 % in Fig. 6c.

The model performance metrics are visualised in Fig. 6b and c. As critical probability increases the model Recall decreases from 99.9 % to 67.8 % shown in Fig. 6b, as the number of Misses increase. While the Specificity increases from 66.8 % to 97.5 % due to a decrease in the number of False Alarms. Balanced accuracy increases from 83.3 % in the baseline model to 91.1 % in the 10 % critical probability test. At higher critical probability values, the balanced accuracy decreases to 82.6 % at 50 % P_{crit}.

4.2.2. Water level threshold (H_{op})

Results from the model outcome classification and metrics are illustrated in Fig. 7 (and Supplementary Table 3).

Classification of the model outcomes shown in Fig. 7a illustrate the shift in distribution of classification depending on the water level threshold. Results show that as the water level threshold is increased from 110 cm to 210 cm the correct outcomes increase significantly from 38 to 5,303, while the number of Hits decrease from 5213 to 119. The number of False Alarms increase from 551 in the 110 cm test to 2250 in the 150 cm test after which the number of False Alarms decrease again to 380 at 210 cm threshold value. While the number of Hit outcomes remain low with only three tests returning Hits, these are 150 cm with 2 instances, 170 cm with 1 instance and 190 cm with 8 Hits. The incorrect model outcomes from water level threshold sensitivity tests are analysed in further detail in section 4.3.2.

The model performance metrics are visualised in Fig. 7b and c. The model Recall is fairly stable for the different water level thresholds tested. The drop in Recall to 97.5 % occurs at 190 cm when 8 Misses occur. The model Specificity increases greatly from 6.4 % at 110 cm to 93.3 % at 210 cm which is related to a decrease in the number of False Alarms. As the number of outcomes in the positive and negative classes changes depending on the water level the Prevalence changes in these sensitivity tests. It decreases from 89.8 % at 110 cm to 2.1 % at 210 cm. Conversely, the Balanced Accuracy increases from 53.2 % at the lowest water level tested to 96.6 % in the test of highest water level.

4.3. Analysis and interpretation of incorrect model outcomes

Incorrect model outcomes for the critical probability and water level threshold sensitivity tests are presented in sub-sections 4.3.1 and 4.3.2, respectively.

4.3.1. Critical probability (P_{crit})

The distribution of False Alarms are shown in Fig. 8 and the Misses are illustrated in Fig. 9, for the critical probability tests. In the baseline configuration, the model resulted in 1626 False Alarm outcomes which correspond to 33.1 % of the negative class. There was only one Miss in the baseline hindcast (Supplementary Table). This occurred on the January 16, 2016 when the measured water level was 180 cm, so 10 cm above the operational water level threshold of 170 cm.

In the critical probability sensitivity tests, the percentage of incorrect outcomes decreased as the critical probability increased, from 12 % to 7 % under the 10 % and 50 % critical probability tests respectively. Within the incorrect outcomes, the number of False Alarms fell while the number of Misses rose (Supplementary Table 2). The total number of False Alarms decreased from 667 in the 10 % to 121 in the 50 % critical probability tests. The annual distribution of False Alarms is shown in Fig. 8a–f. This shows a slight decrease in the annual number of False Alarms. The trend is most noticeable in the 1 % critical probability test shown in Fig. 8a, as this has the highest number of False Alarms. The monthly distribution of model False Alarms is shown in Fig. 8g–l. This reveals a seasonal pattern in the occurrence of False Alarms, with fewer occurring in the months of April to August, while the model returns more False Alarms between September and March. The number of False Alarms in August decreases noticeably as critical probability value is increased. However, this seems to be dependent on the given P_{crit} value and not on the month.

Fig. 9 illustrates the number of Misses where the left column shows annual distribution (panels a–d) and the right column monthly distribution (panels e–h). The number of Miss outcomes increase from 36 to 287 under the 10 % and 50 % tests respectively. When a Miss occurs in a

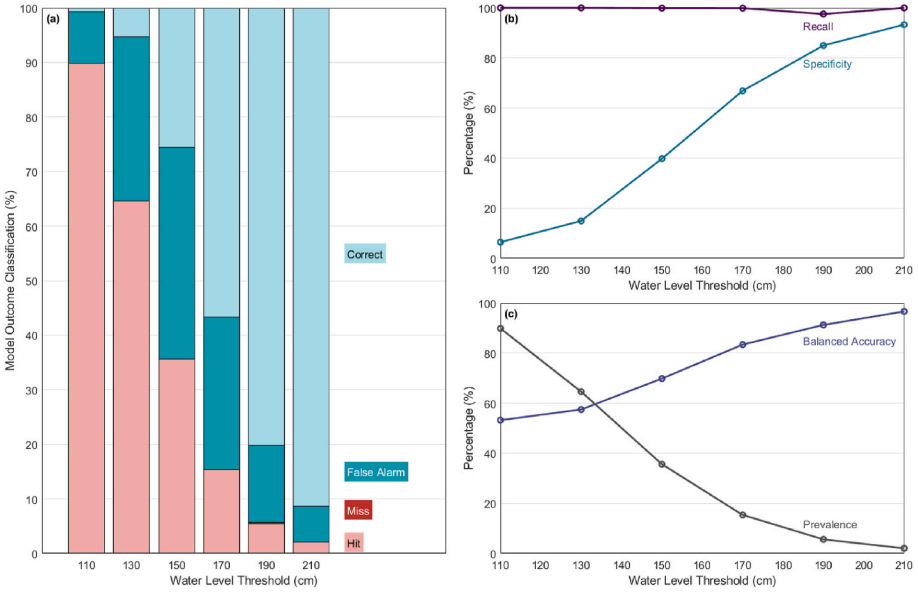


Fig. 7. (a) model output classification for water level threshold sensitivity tests, where the critical probability value is set at 1 %. Model metrics for water level threshold sensitivity tests (b) Recall and Specificity (c) balanced accuracy and prevalence.

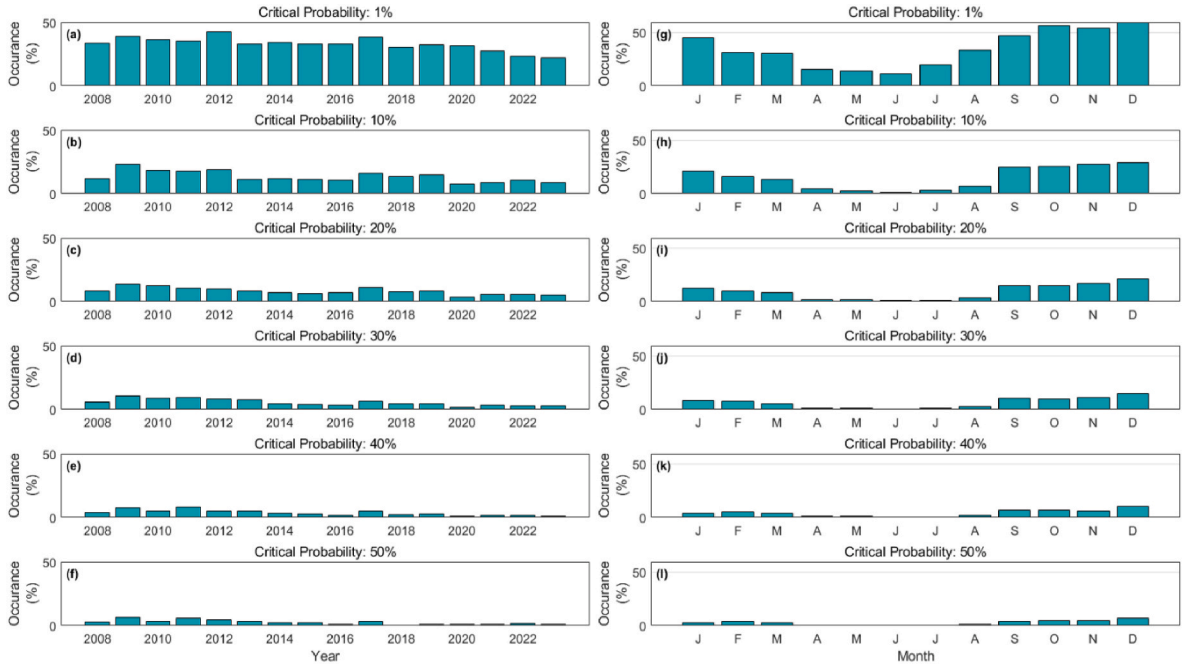


Fig. 8. Percentage occurrence of annual (a–f) and monthly (g–l) False Alarm outcomes from the critical probability sensitivity tests.

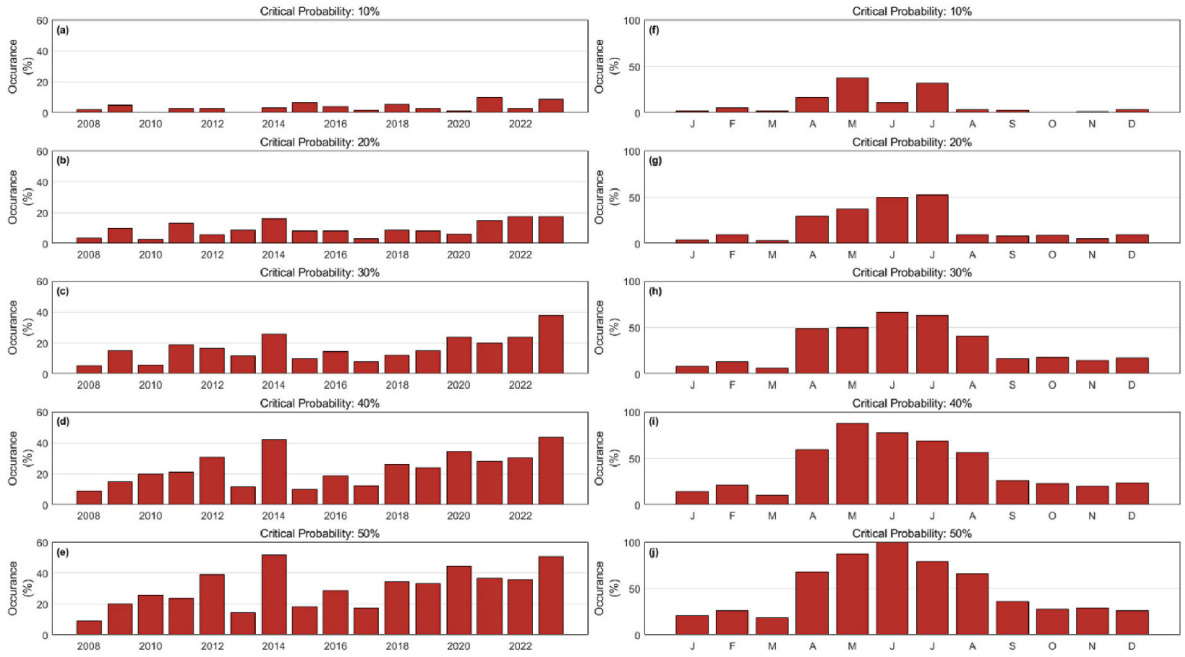


Fig. 9. Percentage occurrence of annual (a–e) and monthly (f–j) Misses from the critical probability sensitivity tests.

test run at a particular critical probability, it will also be present in the successive runs with a higher critical probability value.

The annual distribution of Misses reveals an increasing trend over the 16 years of the hindcast as seen most noticeably in Fig. 9e. The years 2014, 2020 and 2023 stand out as having large numbers of model outcomes classified as Misses, although the reason for this is not yet known. The 30 % critical probability sensitivity test has the greatest increase in Miss events.

The monthly distribution of Misses shows a clear seasonal pattern, with higher percentage occurrence in the months April to August compared to the rest of the year. As the critical probability value increases so does the percentage occurrence of Misses. In the 50 % critical probability test, June resulted in misclassification of all instances where the water levels were above the 170 cm threshold.

The following section describes the analysis of the difference between forecasted and measured water level for all instances where an incorrect model outcome occurred. The freeboard and threshold exceedance amount for the False Alarm and Miss events are shown in Fig. 10a and b respectively, these are presented in standard box plots. Summary statistics are presented in Supplementary Table 4.

The maximum value of the freeboard decreases as critical probability value increases, the interquartile range also decreases from 15 cm at 1 % critical probability to 9 cm at 50 % critical probability. The opposite trend is seen in threshold exceedance, which increases with increasing

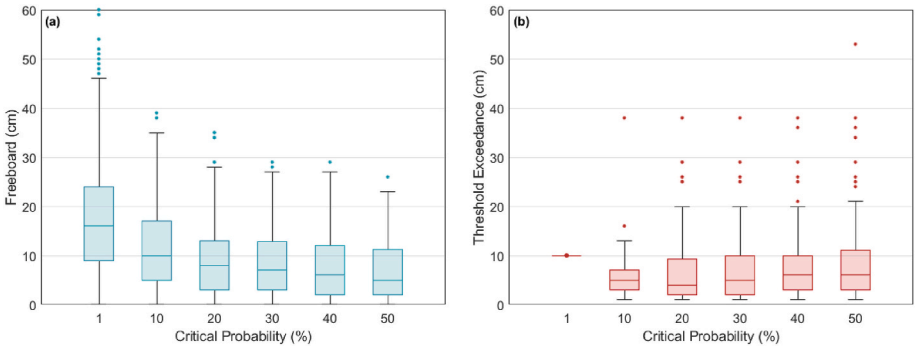


Fig. 10. Standard box plot of (a) False Alarm freeboard extent (b) Miss threshold exceedance amount for the critical probability sensitivity tests.

critical probability value. The interquartile range increases between the 10 % and 20 % critical probability tests and then remains stable for the other tested values.

4.3.2. *Water level threshold (H_{op})*

The annual distribution of False Alarms for different values of H_{op} is shown in Fig. 11a–f and the monthly distribution is shown in Fig. 11g–l. Overall, the results show that the percentage occurrence of False Alarms decreases with increasing water level thresholds and exhibit stronger seasonal patterns at higher water level thresholds. A declining trend in False Alarm occurrence is most noticeable in the 150 cm sensitivity test. For the 110 cm water level threshold, a slight decline in recent years is shown but the percentage occurrence of False Alarms remains above 80 %. In the 150 cm sensitivity test a gradual declining trend is evident after 2015. The two highest water level threshold tests of 190 cm and 210 cm show minimal year to year variation in the proportion of False Alarms.

The seasonal pattern of False Alarms varies depending on the water level threshold value tested. At lower water levels of 110 cm and 150 cm there is higher occurrence of false alarms in the maintenance season (April–September) compared to the storm season (October–March). At the baseline level of 170 cm there is little variation. While at the higher threshold levels of 190 cm and 210 cm show higher False Alarm rates in the storm season compared to the maintenance season.

Summary statistics of freeboard for the false alarms from the water level threshold sensitivity tests are shown in Fig. 12 (and summarised in Supplementary Table 5).

As the water level threshold increases, freeboard also increases. At the water level threshold of 210 cm there were the fewest number of False Alarms, however these events had higher freeboard values.

At three water level thresholds Misses occurred, these are 150 cm, 170 cm and 190 cm, with 2, 1 and 8 instances respectively. These events are summarised in Supplementary Table 6, which indicates the threshold water level of each event the date it occurred and the extent to which the threshold was exceeded. At 150 cm the two Miss events occurred in the last 2 years and only exceeded the threshold by 1 cm. As mentioned previously, the baseline model had one Miss which exceeded the threshold by 10 cm. The Misses that occurred at 190 cm, have all been since 2018 which is the latter third of the hindcast period. Six of the Misses occurred in pairs where consecutive days resulted in incorrect outcomes, these are November 2022, October and November 2023. Not all Misses are equally critical, as their severity depends on how much they exceed the threshold value by.

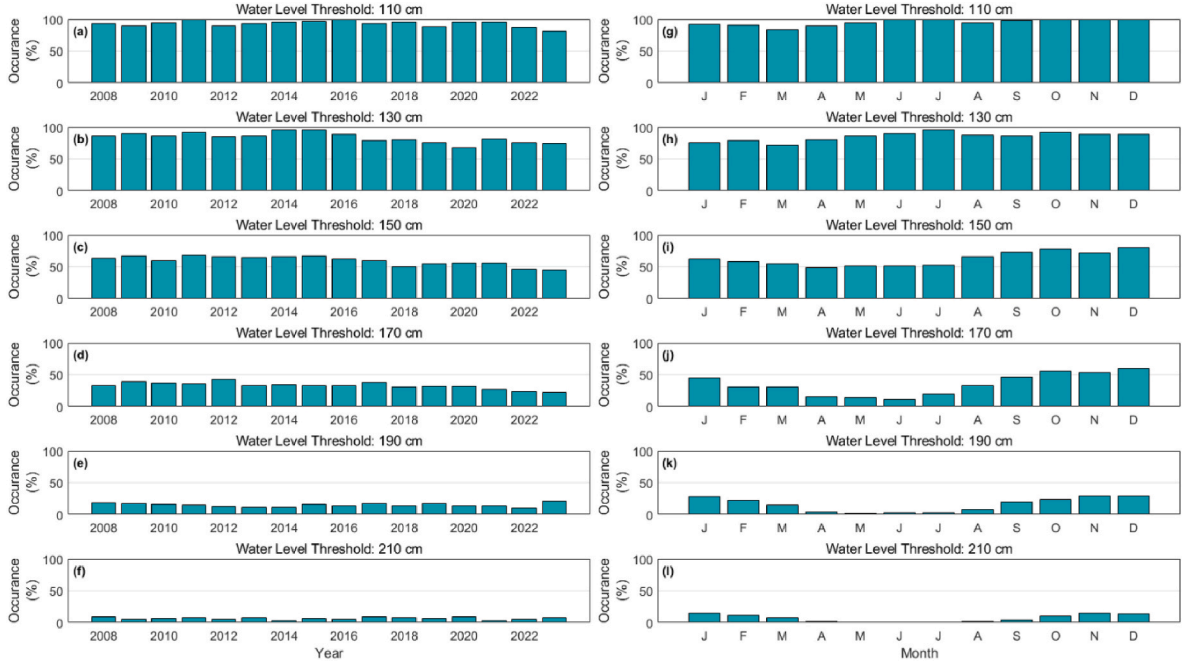


Fig. 11. Distribution of (a–f) annual and (g–l) monthly False Alarms from water level threshold sensitivity tests.

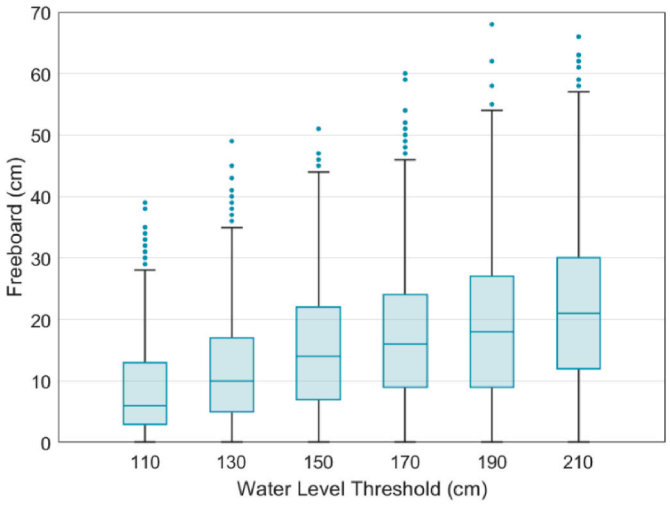


Fig. 12. Standard box plot of False Alarm freeboard extent for the water level threshold tests.

5. Discussion

The aim of this paper was to describe and evaluate a probabilistic model that is used to aid decision making regarding maintenance at storm surge barriers. To address this aim three objectives were defined. Findings from these objectives are discussed in the following sections, with the final part addressing the wider implications of this research and future work.

5.1. Baseline

Of the 5802 days analysed, observations showed that water levels were below the maintenance threshold on 4910 (85 %) days meaning conditions were safe for work to be carried out (the negative class). The remaining 892 (15 %) days had water levels above the 170 cm maintenance threshold meaning it could have been dangerous for staff working at the barrier (the positive class). Therefore, this classification had unbalanced classes as the number of days within each class differ by an order of magnitude or more (Starovoitov and Golub, 2020). This had implications for the metrics that were used to analyse the performance of the probabilistic model, as the high number of negative class outcomes could mask the small number of positive class outcomes. For this reason, Balanced Accuracy is used as a combined measure of model performance.

Performance analysis of the baseline probabilistic model, with values of 1 % for critical probability and 170 cm for water level threshold, illustrated that these parameters were highly conservative. As 33.1 % of model outcomes were False Alarms. On the other hand, of the 892 days when observed water levels were above the maintenance threshold, the model correctly gave the outcome that the safety criterion had not been met on 891 of these days. So, there was only one instance when the model incorrectly returned an outcome that the safety criterion had been met. This showed that the model was very good at avoiding the "increased risk" situations. However, the implications of this for maintenance work at the barrier were that the model underestimates the amount of time when water levels were below the threshold meaning conditions were safe for work to be carried out. Improving the model performance relates to increasing the amount of time when maintenance work can safely be carried out at the barrier. To achieve this, the number of False Alarms need to be decreased, while on the other hand, it is also important to mitigate the number of Misses as these can pose a risk to health and safety. As a means to achieve this, sensitivity tests were conducted to analyse the impact changes in baseline parameters had on model performance.

5.2. Sensitivity testing

Results for the critical probability and water level threshold sensitivity tests are discussed in sub-sections 5.2.1 and 5.2.2, respectively.

5.2.1. Critical probability (P_{crit})

Hindcast runs of the probabilistic model with adjusted parameter values allowed the performance of the modified model to be evaluated. Increasing the critical probability value (P_{crit}) from 1 % to 50 % reduced the number of False Alarms (Fig. 6a), as shown by the increase in Specificity (Fig. 6b), but this change resulted in a simultaneous increase in the number of Misses as indicated by reduction in model Recall. Critical probability values greater than 50 % were not tested as this scenario and the 40 % critical probability resulted in larger numbers of Misses than False Alarms, suggesting an optimum value of critical probability had been exceeded. The rate of change in False Alarms and Misses differed between the critical probability tests. The largest increase in Specificity was seen between critical probability test of 1 % and 10 %. This was due to the greatest decrease in number of False Alarm events (959). The rate of increase then reduced with each successive critical probability value tested. On the other hand, the decrease in Recall was almost linear between critical probability tests of 10 % – 50 %, with the exception between the 1 % and 10 % critical probability values which had the smallest increase in Miss events and therefore the lowest rate of change. The values of Specificity and Recall intersected at a critical probability slightly less than 20 %. The highest balanced accuracy occurred at the 10 % critical probability test with 91.2 %. At higher critical probability values the balanced accuracy decreased, which was due to the greater increase in number of Misses than the decrease in False Alarms. The choice of operational critical probability value therefore depends on how model performance is measured, and thus which metric is used to determine this.

Due to the nature of binary classification, adjustments to the probabilistic model parameters which decrease False Alarms will increase the number of Misses. This can be visualised by a shift in the number of events from the right-hand column to the left-hand column of the confusion matrix (Fig. 4). This means any tuning of model parameters, result in a trade-off between the risk willing to be taken and the loss of maintenance time due to False Alarms. A way to overcome this, is to mitigate the risks posed by the Misses, so that the number of False Alarms can be kept low while ensuring health and safety standards are sufficient. At the Maeslant Barrier this is addressed by sounding an alarm when the water level reaches the maintenance threshold indicating that any ongoing work has to stop.

5.2.2. Water level threshold (H_{op})

In all water level threshold tests the Recall values were above 97 % which was due to the low critical probability value of 1 % used in the tests. This indicated that there were very few Miss outcomes from the model. On the other hand, Specificity increased with increasing water level threshold as the proportion of False Alarms decreased. This was linked to the prevalence which decreased due to a reduction in the number of days when water levels where above the threshold. This can be considered as a shift in the number of events from the bottom row to the top row of the confusion matrix (Fig. 4). This trend resulted in higher balanced accuracy as the ratio of Correct model outcomes in the positive and negative classes improved. This was due to the constantly low number of Misses while the False Alarms became a smaller proportion of the negative class outcomes. This highlighted that at a given critical probability value, the model had more outcomes where the safety criterion was met at higher water level thresholds, which would have indicated more safe weather windows when maintenance work could be carried out. This confirmed that the water level of the maintenance threshold was fundamental to the amount of time available to carry out maintenance safely, highlighting the challenge posed by increasing water levels due to sea-level rise.

In practice, increasing the water level threshold is not as straightforward as just changing the value, other measures need to be taken to ensure workability and safety under such conditions (Trace-Kleeberg et al., 2023). For example, procedural changes could be introduced, such as assessing health and safety risks on a job-specific basis and assigning individual water level thresholds that dictate when work must stop. This would allow tasks in lower-risk areas, such as the control room, to continue at higher water levels compared to more exposed locations. Alternatively, physical modifications, such as retrofitting the barrier to prevent flooding around the ball joint and on the terrain or overtopping of the dock door, could be considered. However, such interventions could involve significant financial and constructional considerations and must ensure that the barrier's operational integrity is not compromised.

5.3. Analysis and interpretation of incorrect model outcomes

The critical probability tests showed a decreasing trend in annual distribution of False Alarms (Fig. 8a–f) and increasing trend in Misses (Fig. 9a–e). Further analysis is required to understand the drivers of the identified trends.

Only counting the number of incorrect model outcomes did not entirely quantify the risk associated with these events. To determine by how much the outcome was incorrect, values of freeboard and threshold exceedances were calculated (Figs. 10 and 12). The results from critical probability sensitivity testing showed a clear trade-off between freeboard and threshold exceedance. As critical probability increased from 1 % to 50 %, freeboard values decreased while the threshold exceedance slightly increased. Such that higher critical probabilities were associated with lower freeboard values but slightly higher threshold exceedances, thus emphasising the possible increase in risk associated with incorrect model outcomes at higher critical probability values.

Results in the water level threshold sensitivity tests showed that False Alarms were associated with increasing freeboard values as water level threshold increased. This trend was accompanied by greater variability and more extreme outliers at higher thresholds. So, although there were fewer False Alarms at higher water level thresholds the difference between the forecast and observed water levels for those events was greater, this demonstrates a larger risk associated with those incorrect model outcomes.

The Miss that resulted in a 38 cm threshold exceedance, corresponding to a water level of 208 cm, occurred on April 28, 2018. The probabilistic model outcome was that the safety criterion had been met which indicated upcoming water levels would be below the maintenance threshold. However, observations showed that the 6th highwater (i.e., just under 3 days from the forecast run time) the maintenance threshold was exceeded resulting in the Miss outcome. This instance occurred because the forecasted highwaters for 28 April did not capture the upcoming event. However, for the 29 April the water level forecasts adjusted, and the probabilistic model outcome changed to safety criteria not met which matched the observed water levels. The same thing occurred in the maximum threshold exceedance event at 50 % critical probability value. where on the forecast run of November 28, 2020 the threshold exceedance was 53 cm indicating a water level of 223 cm. This again corresponded to the 6th highwater, a lead time of 74 h, which by the next forecast run had adjusted to result in a correct probabilistic model outcome of safety criterion not met. This illustrates the importance of accurate water level forecasts as input to the probabilistic model to increase the chance of correct model outcomes with regard to safety criterion and therefore supporting the decision making at the barrier whether it is safe or not for maintenance to be carried out.

Even though the total number of events in the positive class was smaller between April and August the model is not able to correctly classify these events resulting in the incorrect model outcomes and possibly decisions on whether maintenance could be carried out. The reason for the shift in distribution of summer and winter False Alarms (Fig. 11h–l) has not yet been identified. Such investigation and a more detailed assessment of the meteorological and hydrodynamic conditions of incorrect model outcomes could form the basis for a more comprehensive future study.

5.4. Wider implications and further work

To the authors' knowledge, the probabilistic model described and evaluated in the study is the first to use ensemble forecasts to aid decision making at a storm surge barrier on whether upcoming water levels are safe for maintenance work to be carried out. The evaluation of model performance conducted in this study, is a starting point for further development. Such development could be in any of the system steps as outlined in Fig. 2, namely the meteorological model, hydrodynamic model, probabilistic model or the end user decision.

The combination of meteorological and hydrodynamic models used in this study only captures the uncertainty in initial weather conditions through the European Centre for Medium-range Weather Forecast ensemble prediction system. It does not capture uncertainties in the boundary conditions of the hydrodynamic model such as ocean currents, or river discharge. This can lead to differences between the forecast and observed water levels which could result in incorrect model outcomes and therefore impact the False Alarm or Miss situations.

The model combination used in this paper namely the ECMWF EPS/DCSMv5 models are no longer used operationally. Since November 2023, newer models of both the meteorological conditions and hydrodynamics are used for water level forecasting. This presents an opportunity to apply the evaluation method of this paper to another case study as the new dataset extends.

The probabilistic model analysed here, solely examines highwater levels to determine the risk to upcoming maintenance work. At the Maeslant barrier this is sufficient, but to extend the model to other barriers it may need to account for other variables. As mentioned in the introduction, maintenance at the Eastern Scheldt is also governed by wave height and in London at the Thames barrier river discharge in addition to water levels determine whether the barrier will need to close and so limit maintenance work. The other requirement for extending the application of the probabilistic model is the existence of ensemble forecasts of the required parameters in the relevant locations. In addition, the relative contributions of tidal and surge components to total water levels could be investigated in detail to better understand their respective roles, particularly in the model's performance for different water level thresholds. Furthermore, there is considerable uncertainty in future changes in intensity, duration and tracks of storms, which drive storm surges, and the implications for this could be considered in future assessment of the probabilistic model.

The probabilistic model is a decision support tool and requires knowledgeable people at the barrier to interpret the model output and make informed decisions about the risk posed by upcoming water levels to determine if short term maintenance can be done. Additional research could be undertaken to understand how the probabilistic model output is interpreted and used by barrier staff and contractors to aid their decision making. It is likely that other factors influence the decision making on maintenance work, such as criticality, complexity and duration of the planned work which all affect the risk associated with the maintenance. Systematically documenting this could allow the output of the probabilistic model to be better tailored to the end user's needs and build confidence in the teams of the model's output.

Probabilistic models such as the one used in this study provide a tool for addressing the challenge of completing the required maintenance work in reducing periods of time. This is one option to address the challenge posed by a changing climate and sea-level rise, but is unlikely to be the sole solution. Rather a combination of methods will be needed to alleviate the challenge of completing all required maintenance in the available safe working windows (Trace-Kleeberg et al., 2023). It is also likely that other barriers around the world, where set threshold values -

based on water levels or other drivers - are in operation can benefit from a probabilistic model as presented in this paper. The international knowledge sharing network I-STORM (www.i-storm.org) can support the expansion of this decision support tool, both in terms of further method development and application at other locations, through sharing expertise and experience between its members.

6. Conclusions

This paper has described and evaluated a probabilistic model used to aid decision making to determine if the upcoming high-water levels are safe for maintenance work to be carried out at the Maeslant storm surge barrier or not. The probabilistic model performance was evaluated by comparison of model outcome with observations over a 16-year period (2008–2023). The baseline configuration with a critical probability of 1 % and a water level threshold of 170 cm. demonstrated strong performance but was conservative, with 33.1 % of model outcomes resulting in a False Alarm. Conversely, only one miss occurred during the hindcast period. Sensitivity analysis showed that adjusting the critical probability could improve model performance, reducing False Alarms but increasing in the number of Misses. This trade-off underscores the importance of balancing safety and efficiency in the model's configuration. Additionally, water level threshold sensitivity tests revealed that sea-level rise could eventually limit the ability to carry out maintenance, highlighting a potential tipping point where operation of the barrier may be jeopardised. Ultimately, the probabilistic model provides valuable insight into the likelihood of exceeding upcoming water level thresholds to aid decision making regarding the safety of maintenance. However, ongoing evaluation and model parameter tuning is needed, as well as a better understanding of external risk mitigating strategies. Parallel studies are underway to address these knowledge gaps.

Beyond the Maeslant barrier, the probabilistic model presented in this study offers a promising tool for assessing safe maintenance across other storm surge barriers, provided ensemble forecasts and clear operational thresholds are available. The model's approach, incorporating binary classification and sensitivity analysis for evaluation, could be adapted to other coastal infrastructure that relies on similar maintenance protocols. To ensure effective use, the probabilistic model must be tailored to specific contexts with careful consideration of risk tolerance and mitigation measures. Furthermore, training personnel to interpret the model's output is crucial for successful integration into decision-making processes. Given its flexibility, the model could serve as a valuable resource for improving maintenance and operational safety in other locations vulnerable to storm surges in order to contribute to enhanced resilience in the face of climate change induced sea-level rise.

CRediT authorship contribution statement

Sunke Trace-Kleeberg: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Krijn Saman:** Writing – review & editing, Software, Conceptualization. **Robert Vos:** Writing – review & editing, Software. **Elja Huibregtse:** Writing – review & editing, Methodology. **Ivan D. Haigh:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Marc Walraven:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Annette Zijderveld:** Writing – review & editing. **Susan Gourvenec:** Writing – review & editing, Supervision, Funding acquisition.

Code availability

Code and data are available on request from Sunke Trace-Kleeberg.

Funding

Sunke Trace-Kleeberg was funded through the INSPIRE Doctoral Training Partnership by the Natural Environment Research Council (NERC) (NE/S007210/1), and the University of Southampton Marine & Maritime Institute, and co-sponsored by Rijkswaterstaat the Dutch Ministry of Infrastructure and Water Management. Ivan Haigh's time on this projected was funded by a UK Natural Environment Research Council Knowledge Exchange fellowship grant (NE/V018655/1). The time of Krijn Saman, Robert Vos, Elja Huibregtse, Marc Walraven and Annette Zijderveld was supported by Rijkswaterstaat. Susan Gourvenec's time on this projected was supported through the Royal Academy of Engineering Chairs in Emerging Technologies scheme.

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 authors would like to thank Peter Oskam for his collaboration and knowledge sharing about the maintenance of the Maeslant storm surge barrier. The authors are grateful for two anonymous reviewers and the editor for their helpful comments and suggestions which have strengthened this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.coastaleng.2025.104766.

Data availability

Data will be made available on request.

References

Aerts, J.C.J.H., Botzen, W.J.W., Emanuel, K., Lin, N., De Moel, H., Michel-Kerjan, E.O., 2014. Climate adaptation: evaluating flood resilience strategies for coastal megacities. Science. https://doi.org/10.1126/science.1248222, 1979.

Bol, R., 2005. Operation of the "maeslant barrier": (storm surge barrier in the Rotterdam new waterway). In: Flooding and Environmental Challenges for Venice and its Lagoon: State of Knowledge, pp. 311–315.

Brown, S., Nicholls, R.J., Woodroffe, C.D., Hanson, S., Hinkel, J., Kebede, A.S., Neumann, B., Vafeidis, A.T., 2013. Sea-level rise impacts and responses: a global perspective. In: Finkl, C.W. (Ed.), Coastal Hazards. Springer, Dordrecht, pp. 117–149. https://doi.org/10.1007/978-94-007-5234-4_5.

Buizza, R., 2006. The ECMWF ensemble prediction system. In: Predictability of Weather and Climate. Cambridge University Press, pp. 459–488. https://doi.org/10.1017/CBO9780511617652.018.

Buizza, R., Hollingsworth, A., Lalaurette, F., Ghelli, A., 1999. Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System.

Chen, Z., Orton, P., Wahl, T., 2020. Storm surge barrier protection in an era of accelerating seasea-level rise: quantifying closure frequency, duration and trapped river flooding. J. Mar. Sci. Eng. 8, 725. https://doi.org/10.3390/jmse8090725.

Del-Rosal-Salido, J., Folgueras, P., Bermúdez, M., Ortega-Sánchez, M., Losada, M.Á., 2021. Flood management challenges in transitional environments: assessing the effects of sea-level rise on compound flooding in the 21st century. Coast. Eng. 167, 103872. https://doi.org/10.1016/j.coastaleng.2021.103872.

Deltares, 2018. Overview storm surge barriers [WWW Document]. Deltares. URL. https://kyst.dk/media/80421/deltares_2018_overview_storm_surge_barriers_komprimeret.pdf.

de Vries, H., 2008. Probability Forecasts for Water Levels at the Coast of The Netherlands. ECMWF Newsletter. https://doi.org/10.21957/gpsn56s02c.

de Vries, H., 2009. Probability forecasts for water levels at the coast of The Netherlands. Mar. Geod. 32, 100–107. https://doi.org/10.1080/01490410902869185.

Fahmy, M., 2022. Confusion matrix in binary classification problems: a step-by-step tutorial. Journal of Engineering Research 6 (5). Available at: https://digitalcommons.aaru.edu.jo/erjeng/vol6/iss5/1/.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.
Flood Protection Authority East, 2014. Hurricane storm damage risk reduction system [WWW Document]. URL. https://storymaps.arcgis.com/stories/4c5f9679cb1d40578d5f575d81f972da.
Flowerdew, J., Horsburgh, K., Wilson, C., Mylne, K., 2010. Development and evaluation of an ensemble forecasting system for coastal storm surges. Q. J. R. Meteorol. Soc. 136, 1444–1456. https://doi.org/10.1002/qj.648.
Fox-Kemper, B., Hewitt, H.T., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S.S., Edwards, T.L., Golledge, N.R., Hemer, M., Kopp, R.E., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I.S., Ruiz, L., Sallée, J.-B., Slangen, A.B.A., Yu, Y., 2021. Ocean, cryosphere and sea level change. In: Climate Change 2021: the Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, p. 1211. https://doi.org/10.1017/9781009157896.011.
Gerritsen, H., de Vries, H., Philippart, M., 1995. The Dutch continental Shelf model. In: Lynch, D.R., Davies, A.M. (Eds.), Quantitative Skill Assessment for Coastal Ocean Models. American Geophysical Union, pp. 425–467. https://doi.org/10.1029/CE047p0425.
Haigh, I.D., D'Arcy, E., Brand, J., Inayatillah, A., Trace-Kleeberg, S., Walraven, M., Saman, K., Batchelor, A., Lewis, C., Barlow, N.L.M., Thompson, P., O'Brien, P., Marzion, R., 2024. Rapid Acceleration in the Number of Closures of Storm Surge Barriers in the Future: A New Tool for Estimating Barrier Closures. https://doi.org/10.20944/preprints202410.2298.v1.
Hamerslag, E.J.F., Bakker, A.M.R., 2023. Embedding functional performance in asset management of hydraulic structures. In: Life-Cycle of Structures and Infrastructure Systems. CRC Press, pp. 2591–2597. https://doi.org/10.1201/9781003323020-315.
Haasnoot, M., Bouwer, L., Diermanse, F., Kwadijk, J., van der Spek, A., Oude Essink, G., Delsman, J., Weiler, O., Mens, M., ter Maat, J., Huismans, Y., Sloff, K., Masselman, E., 2018. Mogelijk gevolgen van versnelde zeespiegelstijging voor het Deltaprogramma. Een verkenning. Deltares rapport 11202230-005-0002. https://www.deltares.nl/nl/publication/mogelijke-gevolgen-van-versnelde-zeespiegelstijging-voor-het-deltaprogramma-een-verkenning/.
Hinkel, J., Lincke, D., Vafeidis, A.T., Perrette, M., Nicholls, R.J., Tol, R.S.J., Marzeion, B., Fettweis, X., Ionescu, C., Levermann, A., 2014. Coastal flood damage and adaptation costs under 21st century sea-level rise. Proc. Natl. Acad. Sci. U. S. A. 111, 3292–3297. https://doi.org/10.1073/PNAS.1222469111/SUPPL_FILE/PNAS.201222469SI.PDF.
ICE, 2020. Boston barrier scheme [WWW Document]. URL. https://www.ice.org.uk/what-is-civil-engineering/what-do-civil-engineers-do/boston-barrier-scheme, 1.21.22.
Irazoqui Apecechea, M., 2018. HATYAN technical design. Available at: https://github.com/Deltares/hatyan/blob/main/docs/11202223-000-ZKS-0002%20-%20HATYAN%20technical%20design_maart2018.pdf.
Jan De Nul, 2024. Storm surge barrier protects Belgium from natural disasters [WWW Document]. URL. https://www.jandenul.com/storm-surge-barrier-protects-belgium-natural-disasters, 4.25.24.
Jonkman, S.N., Hillen, M.M., Nicholls, R.J., Kanning, W., Van Ledden, M., 2013. Costs of adapting coastal defences to sea-level rise - new estimates and their implications. J. Coast Res. 29, 1212–1226. https://doi.org/10.2112/JCOASTRES-D-12-00230.1.
Kirshen, P., Borrelli, M., Byrnes, J., Chen, R., Lockwood, L., Watson, C., Starbuck, K., Wiggin, J., Novelly, A., Uiterwyk, K., Thurson, K., McMann, B., Foster, C., Sprague, H., Roberts, H.J., Bosma, K., Jin, D., Herst, R., 2020. Integrated assessment of storm surge barrier systems under present and future climates and comparison to alternatives: a case study of Boston, USA. Clim. Change 162, 445–464. https://doi.org/10.1007/s10584-020-02781-8.
Kluijver, M., Dols, C., Jonkman, S.N., Mooyaart, L., 2019. Advances in the planning and conceptual design of storm surge barriers – application to the New York metropolitan area, 326–336. https://doi.org/10.18451/978-3-939230-64-9.
Knoester, M., Visser, J., Bannink, B.A., Colijn, C.J., Broeders, W.P.A., 1984. The eastern Scheldt project. Water Sci. Technol. 16, 51–77. https://doi.org/10.2166/wst.1984.0044.
Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. J. Comput. Phys. 227, 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014.
Merrell, W.J., Reynolds, L.G., Cardenas, A., Gunn, J.R., Hufton, A.J., 2011. The ike dike: a coastal barrier protecting the houston/galveston region from Hurricane storm surge. In: Environmental Science and Engineering. Springer Science and Business Media Deutschland GmbH, pp. 691–716. https://doi.org/10.1007/978-3-642-14779-1_31.
Mooyaart, L., Jonkman, S., De Vries, P., Van der Toorn, A., Van Ledden, M., 2014. Storm surge barrier: overview and design considerations. Coastal Engineering Proceedings 1, 45. https://doi.org/10.9753/icce.v34.structures.45.
Mooyaart, L.F., Jonkman, S.N., 2017. Overview and design considerations of storm surge barriers. J. Waterw. Port, Coast. Ocean Eng. 143, 06017001. https://doi.org/10.1061/(asce)ww.1943-5460.0000383.
Munaretto, S., Vellinga, P., Tobi, H., 2012. Flood protection in venice under conditions of seasea-level rise: an analysis of institutional and technical measures. Coast. Manag. 40, 355–380. https://doi.org/10.1080/08920753.2012.692311.
Smaling, D., 2024. "3D storm surge barrier map" [web map]. Scale Not Given. https://i-storm.maps.arcgis.com/apps/instant/3dviewer/index.html?appid=cfdeabee7cc2472eb76d177d5c8e2051. (Accessed 24 March 2025).
Somerset Council, 2024. Bridgwater tidal barrier [WWW Document]. URL. https://www.somerset.gov.uk/beaches-ports-and-flooding/bridgwater-tidal-barrier/. ed 4.25.24.
Starovoitov, V., Golub, Y., 2020. Comparative study of quality estimation of binary classification. https://doi.org/10.13140/RG.2.2.13697.28000.
Stephenson, D.B., Coelho, C.A.S., Doblas-Reyes, F.J., Balmaseda, M., 2005. Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. Tellus Dyn. Meteorol. Oceanogr. 57, 253. https://doi.org/10.3402/tellusa.v57i3.14664.
Trace-Kleeberg, S., Haigh, I.D., Walraven, M., Gourvenec, S., 2023. How should storm surge barrier maintenance strategies be changed in light of sea-level rise? A case study. Coast. Eng. 184. https://doi.org/10.1016/j.coastaleng.2023.104336.
US Army Corps of Engineers, 2015. New Bedford Hurricane protection barrier [WWW Document]. https://www.nae.usace.army.mil/Missions/Civil-Works/Flood-Risk-Management/Massachusetts/New-Bedford/, 21.10.24.
Van Dorland, R., Beersma, J., Bessembinder, J., Bloemendaal, N., Van Den Brink, H., Brotons Blanes, M., Drijfhout, S., Groenland, R., Haarsma, R., Homan, C., Keizer, I., Krikken, F., Le Bars, D., Lenderink, G., Van Meijgaard, E., Meirink, J.F., Overbeek, B., Reerink, T., Selten, F., Severijns, C., Siegmund, P., Sterl, A., De Valk, C., Van Velthoven, P., De Vries, H., Van Weele, M., Wichers Schreur, B., Van Der Wiel, K., 2024. KNMI national climate scenarios 2023 for The Netherlands, 2023. Scientific report. De Bilt. Available at: https://www.knmi.nl/kennis-en-datacentrum/publicatie/knmi-national-climate-scenarios-2023-for-the-netherlands, 6.14.24.
Wagenaar, N., 2018. Verification of WAQUA/DCSMv5's operational water level probability forecasts. De Bilt. Available at: https://cdn.knmi.nl/knmi/pdf/bibliotheek/knmipubIR/IR2018-01.pdf, 6.14.24.
Walraven, M., Vrolijk, K., Kothuis, B.B., 2022. Design, maintain and operate movable storm surge barriers for flood risk reduction. In: Coastal Flood Risk Reduction. Elsevier, pp. 271–286. https://doi.org/10.1016/B978-0-323-85251-7.00020-2.
Wilkes, D., Lavery, S., 2005. The Thames Barrier - now and in the future. In: Fletcher, C. A., Spencer, T. (Eds.), Flooding and Environmental Challenges for Venice and its Lagoon: State of Knowledge. Cambridge University Press, pp. 287–294.
Zhong, H., Overloop, P.-J. Van, Gelder, P. Van, Rijcken, T., 2012. Influence of a storm surge barrier's operation on the flood frequency in the rhine delta area. Water (4), 474–493. https://doi.org/10.3390/W4020474, 2012, Vol. 4, Pages 474-493.
Zijderveld, A., Netel, S., Verboeket, R., Driebergen, J., van Schaik, M., van Galen, L., 2024. Stormsurge tides storm pia from 21 to 22 december 2023 (in Dutch: Stormvloed tijden storm Pia van 21 en 22 december 2023). https://waterberichtgeving.rws.nl/data/579-25-sr101_stormvloedrapport_pia_met_bijlagen.pdf.
Zijl, F., Verlaan, M., Gerritsen, H., 2013. Improved water-level forecasting for the Northwest European Shelf and North Sea through direct modelling of tide, surge and non-linear interaction. Ocean Dyn. 63, 823–847. https://doi.org/10.1007/s10236-013-0624-2.