TECHNICAL UNIVERSITY OF DENMARK AND POLITECNICO DI TORINO

Master's Degree in Environmental and Land Engineering



Master's Thesis

Correcting Hydrodynamic Model Biases in Western Baltic Sea Storm Surges via Neural Networks

Supervisors

Prof. Vincenzo NAVA

Prof. Karina NIELSEN

Dr. Jun SHE

Dr. Morten Andreas Dahl LARSEN

Dr. Vahidreza JAHANMARD

Candidate

Nicolo' CARON

September 2026

Abstract

Hydrodynamic sea-level models provide spatially continuous information but retain local, time-varying errors that can become important during high-water events. This thesis investigates whether the dynamic DKSS error can be learned across the Danish and western Baltic tide-gauge network and transferred to stations excluded from model development.

Hourly data from 2013–2022 combine 28 tide gauges, DKSS sea level and atmospheric forcing. After removing a station-dependent static bias, a causal Temporal Convolutional Network predicts the remaining dynamic residual from atmospheric and hydrodynamic histories. Fifteen stations are used for training, six for validation and seven for independent testing.

Across the test stations, the correction reduces RMSE from 8.39 to 5.37 cm relative to bias-corrected DKSS, corresponding to a 35.9% improvement, while also reducing the variability of the error between stations. For 45 registered high-water events, RMSE over the observed peak ±60 hours decreases from 11.04 to 7.51 cm and mean absolute error at the observed peak from 12.07 to 8.25 cm. The correction is particularly effective for strongly underestimated high-water peaks, substantially reducing the systematic negative error of the hydrodynamic model, although some individual events still deteriorate.

Overall, the results show that a substantial component of the dynamic DKSS error is predictable from atmospheric forcing and recent hydrodynamic behaviour and can be transferred across geographically distinct stations using a single shared correction model. The remaining errors and event-to-event variability indicate that the learned correction complements, rather than replaces, the underlying hydrodynamic model.

Table of Contents

Abstract **I**

Abbreviations **VIII**

1 Introduction **1**
1.1 Coastal water level as a decision variable 1
1.1.1 Underestimating a peak means underestimating the risk . . . 2
1.2 The error of the hydrodynamic model 2
1.3 Correcting the error where the gauge cannot 3
1.4 Residual learning as the correction strategy 6
1.5 Research question, objectives and contribution 6
1.6 Structure of the thesis 8

2 Dataset and target variable **9**
2.1 Data foundation of the correction problem 9
2.2 Raw data sources 9
2.3 Station network 10
2.4 Tide-gauge observations 12
2.5 Hydrodynamic model data 13
2.6 Atmospheric forcing data 13
2.7 Quality control and despiking 15
2.7.1 Tide-gauge and target checks 15
2.7.2 Hydrodynamic-model and atmospheric checks 16
2.8 Target variable: the dynamic residual 17
2.8.1 Dynamic-residual sign and distribution over the complete record 19
2.8.2 Frequency structure of the target 20
2.9 Summary 23

3 Physical interpretation of hydrodynamic-model error **24**
3.1 Mechanisms behind the residual 24
3.2 Physical signatures in the dynamic residual 25
3.3 Wind direction and coastal geometry 28
3.4 Observed memory and frequency structure 29
3.5 KDI registered high-water events 31
3.5.1 What the register shows about the events 33

3.6 Domain-scale forcing modes . . . 34
3.7 Related residual-correction studies . . . 36
3.8 From diagnosis to model design . . . 37

4 Methodology **38**
4.1 Residual post-processing formulation . . . 39
4.2 Time-ordered samples and predictor windows . . . 39
4.3 Predictor representation . . . 40
4.3.1 Standardization and calibration scope . . . 42
4.4 Shared-weight TCN architecture . . . 42
4.4.1 Causal dilated convolutions and receptive field . . . 43
4.4.2 Residual block . . . 44
4.4.3 Prediction head and output . . . 44
4.5 Extreme-oriented training objective . . . 45
4.6 Optimization and checkpoint selection . . . 47
4.7 Spatial experimental protocols . . . 47
4.7.1 Primary station-split evaluation . . . 47
4.7.2 Independent input-group ablation . . . 48
4.8 Evaluation metrics . . . 48

5 Results and discussion **51**
5.1 Finding 1: The dynamic residual is learnable and transferable . . . 51
5.1.1 What the baseline already accounts for . . . 51
5.1.2 Pooled and station-level transfer . . . 52
5.2 Finding 2: High-water skill depends on the evaluation sample . . . 54
5.2.1 Peak errors contract towards a common residual scale . . . 57
5.3 Secondary robustness example: regional specialisation . . . 58
5.4 What the model uses and what it removes . . . 60
5.4.1 Which inputs carry the transferable signal? . . . 60
5.4.2 Remaining frequency structure . . . 61
5.5 Finding 3: Post-processing remains available during gauge outages . 62
5.6 Synthesis of the findings . . . 64

6 Conclusions **65**
6.1 Main findings . . . 65
6.2 Scientific and methodological contribution . . . 66
6.3 Operational interpretation . . . 67
6.4 Limitations . . . 67
6.5 Future work . . . 68

References **70**

List of Figures

1.1 Population density in Denmark 1
1.2 The Danish and western Baltic study area 4
1.3 Post-processing rather than physical-model replacement 5
1.4 End-to-end workflow of the thesis 7

2.1 The 28-station tide-gauge network 11
2.2 Observed mean sea level and variability at the 28 stations 12
2.3 Median modelled sea level over the DKSS grid 14
2.4 Verified tide-gauge quality-control example 16
2.5 Station reference-level bias 18
2.6 Residual post-processing chain 19
2.7 Dynamic-residual distributions at Grenaa, Faaborg and Drogden-Fyr 20
2.8 Spread and tail asymmetry across all 28 stations 21
2.9 Dynamic-residual time series and spectra at Grenaa, Faaborg and Drogden-Fyr 22

3.1 Storm-surge drivers and hydrodynamic-model error during Malik . . 26
3.2 Representative hydrodynamic-model discrepancies before neural correction 27
3.3 Observed sea level and raw DKSS error at three reference stations . 27
3.4 Dynamic residual by wind direction at six stations 29
3.5 Autocorrelation of the hydrodynamic-model error 30
3.6 Median error spectrum across the 28-station network 31
3.7 Official high-water statistics datasheet (Bagenkop) 33
3.8 Leading atmospheric EOF patterns and their scores 35

4.1 Architecture of the final 16-channel shared-weight TCN 43
4.2 Shape and asymmetry of the extreme-oriented loss 46
4.3 Spatial train–validation–test split 48

5.1 Full-record RMSE for raw DKSS and the post-processed output . . . 53
5.2 Full-record error spread for raw DKSS and the post-processed output 53
5.3 Skill across the complete record and registered events 54
5.4 Raw DKSS and post-processed errors during one registered event . . 55
5.5 Contrasting event examples for the final post-processing model . . . 56

5.6 Raw and post-processed errors at time-separated observed peaks . . 57
5.7 Network-wide and regional training on identical test observations . . 59
5.8 Validation-first leave-one-group-out ablation 60
5.9 Residual spectra before and after correction 61
5.10 Gap-filling cards during ordinary and storm conditions 63

List of Tables

2.1 Main data sources used to construct the residual post-processing dataset. 10
2.2 Tide-gauge diagnostics of the 28 stations 11

3.1 Major storm-surge episodes 2013–2022 34

4.1 Sign convention of the extreme-oriented loss 45

5.1 Pooled performance across the main evaluation settings 52
5.2 Highest registered event by region 58
5.3 Pooled network-wide and regional performance 59
5.4 Validation-ranked input-group ablation 61
5.5 Masked-window reconstruction performance 62

Abbreviations

ACF	autocorrelation function
DMI	Danish Meteorological Institute
DKSS	DMI operational sea-level product for Danish waters
DVR90	Dansk Vertikal Reference 1990
EOF	empirical orthogonal function
EoL	extreme-oriented loss
HDM	hydrodynamic model
KDI	Danish Coastal Authority (Kystdirektoratet)
MAE	mean absolute error
RMSE	root-mean-square error
SLP	sea-level pressure
TCN	temporal convolutional network
TG	tide gauge

Chapter 1

Introduction

1.1 Coastal water level as a decision variable

Denmark is a low-lying maritime country whose population is not concentrated around a single inland centre. Settlements, harbours, transport links and urban infrastructure are distributed along the shores of the Kattegat, the Danish Straits and the western Baltic, often directly at the waterline (Figure 1.1). Coastal water level is therefore not only a geophysical quantity to be described. It is the variable on which harbour operations are scheduled, storm-surge warnings are issued, temporary flood barriers are deployed and permanent coastal protection is dimensioned.

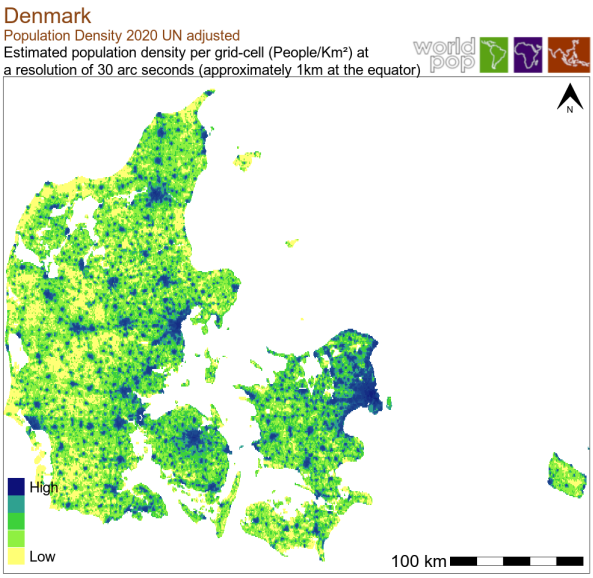


Figure 1.1: Population density in Denmark. The source and licence are printed in the image: WorldPop, 2020 UN-adjusted population-density product, distributed under CC BY 4.0. The figure documents the coastal distribution of exposure; no population variable enters the analysis.

Those decisions are threshold decisions, and the thresholds are published. The

Danish Coastal Authority (Kystdirektoratet, KDI) maintains official high-water statistics for the national tide-gauge network: for every station it prints a return-level curve, the design levels of the 20-, 50- and 100-year events, and a dated register of the extreme water levels actually recorded (Poulsen et al. 2024). These levels are the reference against which coastal protection is designed and climate-adaptation measures are planned. A predicted water level is useful in proportion to how well it places an approaching event relative to them.

1.1.1 Underestimating a peak means underestimating the risk

This is what makes a centimetre-scale error consequential. Across the 23 stations whose published parameters reproduce their printed 100-year level to within 1.5 cm, the median separation between the 20- and 100-year design levels is only 18 cm, and just 11 cm at Køge and Drogden Fyr (Poulsen et al. 2024). The register and its consistency checks are described in Section 3.5. Chapter 5 shows that bias-corrected DKSS underestimates 350 selected high-water peaks across the seven test stations by a median of 8.5 cm. In KDI's return level equations, this would make a true 100-year level appear as a 45-year event, and a 20-year level as an 11-year event. Errors of only a few centimetres can therefore shift an event substantially within the official return period scale used for planning and warning.

The two error directions are not equivalent in consequence. Overestimation may trigger unnecessary precautionary action, whereas underestimation can leave warnings, barriers or preparedness below the level required by the event that actually occurs. This asymmetry is therefore treated as a design constraint and reappears in Chapter 4 through the training objective.

The events are not hypothetical. During 2013 to 2022, Bodil and the January 2017 western Baltic surge each reached the reproduced 100-year class at one or more stations, while the widest episodes were recorded simultaneously at 19 to 21 gauges (Poulsen et al. 2024). The January 2017 event is particularly instructive because it was a "silent" surge: exceptional water levels developed without equally exceptional local winds (She and Nielsen 2019; Andrée et al. 2023). This illustrates why extreme coastal water levels can remain difficult to predict from the instantaneous meteorological signal alone.

Improving extreme sea level prediction in these waters therefore remains a problem in which errors of only a few centimetres can have practical consequences at the coast.

1.2 The error of the hydrodynamic model

Hydrodynamic models provide the physically consistent estimate from which such a prediction starts. They solve discretized forms of the governing ocean equations and combine bathymetry, coastline geometry, atmospheric forcing and boundary conditions into a sea-level field that is continuous over the whole domain. This is

what a tide gauge cannot deliver: the gauge measures the water level accurately, but at one point on the coast. Figure 1.2 shows the domain over which that field is produced: the transition between the North Sea and the Baltic, where the external signal enters through the Skagerrak and the Kattegat and is filtered by the Danish Straits.

No such model reproduces the coastal water level exactly. Its error can originate in the atmospheric forcing, in imperfect bathymetry, in insufficient resolution near the coast, in simplified parameterizations, in the open-boundary conditions, in datum inconsistencies or in local processes the grid does not resolve. During a storm surge these deficiencies surface as a peak of the wrong amplitude, a peak at the wrong time, or a build-up and relaxation that do not follow the observed event, exactly the situations described in Section 1.1.1, in which the predicted level is compared with a threshold.

The practical difficulty is that this error is only visible where an observation exists. Tide gauges deliver high-quality coastal records, but their spatial coverage is sparse and their series contain gaps, datum changes and instrumental anomalies. Away from the gauges, neither the magnitude nor the temporal structure of the model error is directly known. The model is continuous but imperfect; the observations are accurate but local. A correction method should exploit both rather than substitute one for the other.

The error is also not one homogeneous quantity. Part of it is a persistent offset that reflects the difference between the vertical reference of the gauge and that of the model, and part of it varies in time with the meteorological and hydrodynamic state. This thesis calibrates the first and learns the second, hereafter the *dynamic residual*. Chapter 2 defines both terms formally from the collocated records. The separation fixes the baseline of the whole thesis: every improvement reported later is measured against the hydrodynamic model after the persistent offset has already been removed, so it quantifies the correction of the time-varying error and never the removal of a constant. This convention follows established Baltic sea-level practice (Jahanmard et al. 2023).

The dynamic residual is only partly predictable. It combines observation noise, collocation error and genuinely unpredictable variability with a component driven by known predictors, namely atmospheric forcing and recent hydrodynamic variability. The correction developed here does not attempt to reproduce every fluctuation; it estimates the predictable part from predictor histories that end at the time being estimated. The network creates no new ocean physics. It estimates a post-processing correction to the output of a physical model, which places the problem in the domain of time-series modelling (Madsen 2007).

1.3 Correcting the error where the gauge cannot

An operational sea-level service already combines these two ingredients: a deterministic hydrodynamic product covering the domain, and a sparse network of gauges

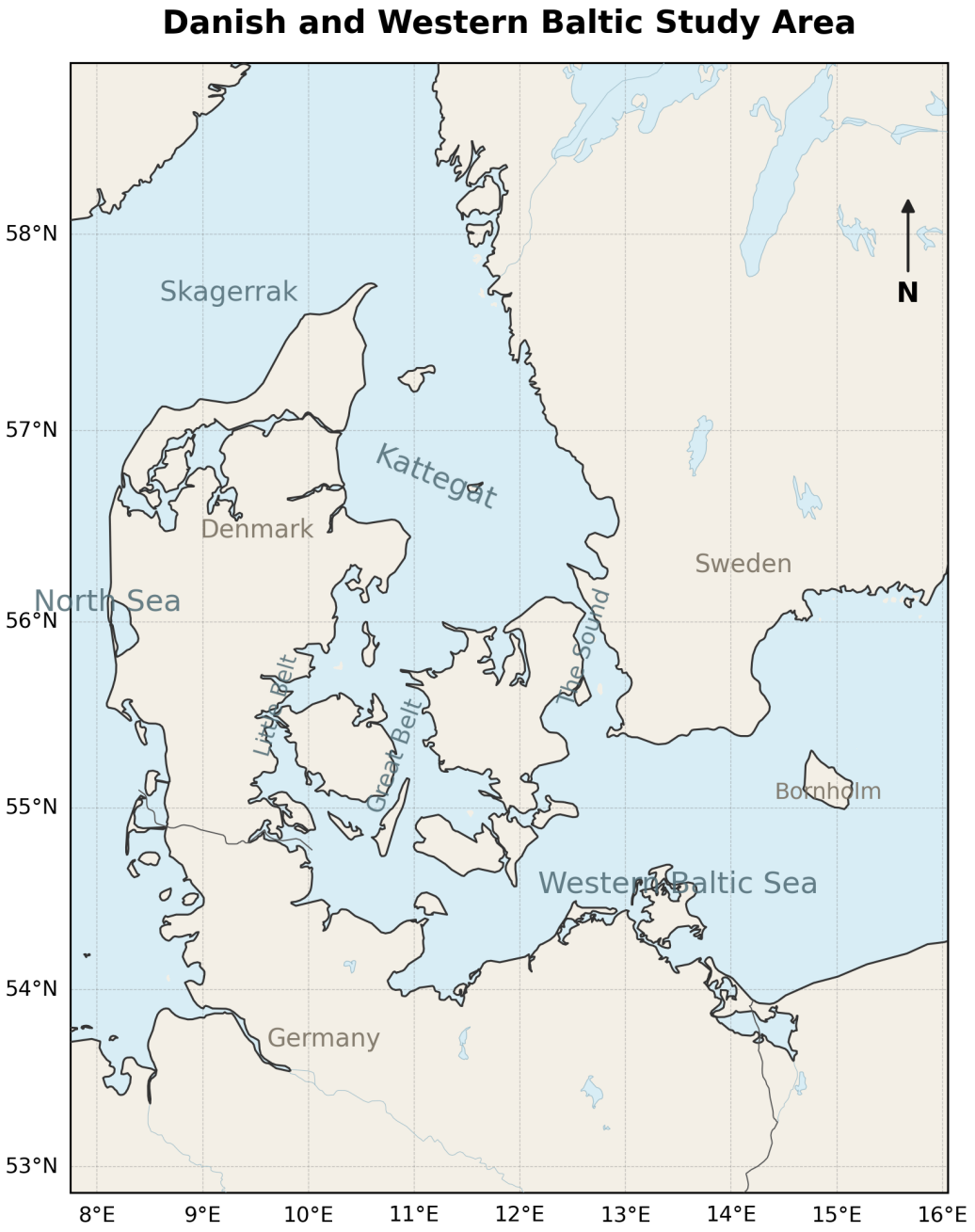


Figure 1.2: The Danish and western Baltic study area. The domain is the transition between the North Sea and the Baltic Sea: the external signal enters through the Skagerrak and the Kattegat and is filtered by the Danish Straits before reaching the western Baltic.

used for monitoring and local correction. A disagreement between them may indicate model error, an observational problem or unresolved local variability, and the distinction matters most during a high-water event or a gauge interruption. The same storm can also produce very different agreement between model and gauge at neighbouring stations, so a network-wide average says little about local reliability.

This is the operating condition that shapes the design of the correction. A method that relies on the recently observed error at the target station can follow event-scale variation, but it stops working exactly when the gauge does. The model studied here instead reads atmospheric and hydrodynamic histories ending at the prediction time and uses neither the contemporaneous gauge value nor a station identifier (Figure 1.3). Two consequences follow directly. The correction remains computable while a gauge is out of service, which makes masked-gap reconstruction a testable secondary application; and the same learned weights can be applied at stations whose observations never entered the fitting, which makes spatial transfer a testable claim rather than an aspiration.

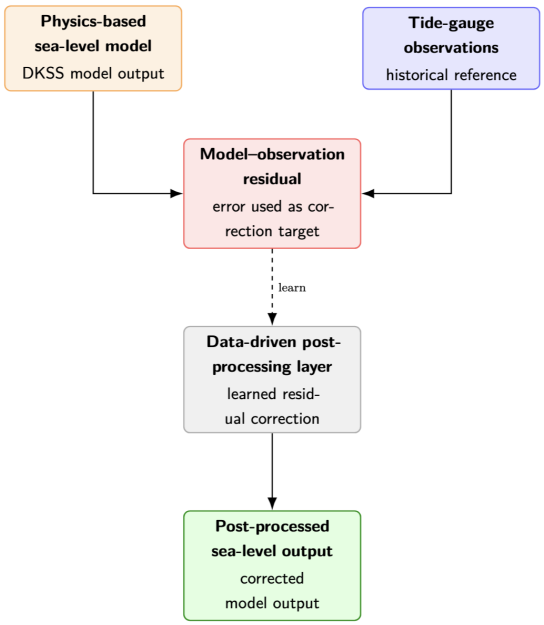


Figure 1.3: High-level post-processing concept. Tide-gauge observations define the historical reference for the model error; a data-driven layer learns its predictable component and returns a post-processed version of the hydrodynamic output.

The experiments use archived hindcast-quality inputs, so they constitute an operationally motivated research assessment and not a deployed forecast service. The correction is evaluated inside an existing gauge network: the persistent reference-level offset must still be known at each corrected location, which is why the results demonstrate transfer to stations excluded from fitting rather than deployment at a coast with no observational history. Within that scope the thesis asks how much of the dynamic residual can be recovered.

1.4 Residual learning as the correction strategy

Residual learning has become the natural formulation for this problem. Instead of predicting the complete water level, the data-driven component predicts the error of an existing hydrodynamic forecast. The physical model remains the backbone of the prediction system, and the statistical model faces a smaller, better-posed target: the part of the error that is related to known predictors and past states. Residual correction of operational sea-level forecasts has already proved effective in nearby basins (Tedesco et al. 2024; Jahanmard et al. 2023). This thesis follows the same philosophy but concentrates on the error during extreme sea levels, on a lightweight temporal architecture, and on spatial transfer across the Danish network; the detailed comparison with previous work is developed in Chapter 3.

Two design choices follow from the nature of the target. First, the correction is posed as a time-series problem: the residual carries memory over several days, so the model must read a history rather than an instantaneous state. Second, spatial context enters through domain-scale atmospheric modes rather than through station coordinates. Recurring large-scale pressure, temperature and wind fields are compressed into a small number of temporal scores by empirical orthogonal function (EOF) analysis, which for readers from a machine-learning background is the same linear decomposition commonly called principal component analysis. Combining those shared scores with station-local histories tests whether the transferable part of the residual can be recovered without station identity and without static local descriptors.

Before any model is trained, the residual is characterized directly: its magnitude, memory, spectrum, spatial coherence and dependence on wind. That diagnosis is what justifies the representation.

1.5 Research question, objectives and contribution

The thesis is organised around one research question:

> *Can a single residual post-processing model reduce the sea-level error remaining after reference-level calibration at test stations outside the training population, both over the complete record and during independently registered high-water events, using only past and current atmospheric and hydrodynamic predictors and neither station identity nor the contemporaneous tide-gauge value?*

There is one target, the dynamic residual, and one primary transfer test: the same learned weights are evaluated at stations whose residuals entered neither fitting nor model selection. Complete-record error measures average post-processing skill, the station map shows where that skill transfers, and the independent KDI register tests event windows and observed peaks.

The specific objectives are:

1. to define a physically interpretable target by separating the station-dependent static bias from the time-varying hydrodynamic modelling error;
2. to characterize that target before any learning, using station-wise statistics, autocorrelation, spatial correlations, event classification and wind diagnostics, and to compress the gridded atmospheric fields with EOF analysis;
3. to develop a temporal neural-network correction that uses no predictor value later than the prediction time;
4. to quantify how much of the dynamic residual the correction recovers at test stations outside the training population, using global, per-station and extreme-event metrics;
5. to assess whether the resulting framework can support future operational applications, including post-processing and gap filling of interrupted gauge records.

The contribution is therefore a complete workflow: definition of the target, physical diagnosis of its spatial, temporal and frequency structure, time-ordered learning with causal left-padded convolutions, and evaluation on the registered high-water events of the Danish Coastal Authority. Taking extremes from the authority's register, each event already carrying an official return-period class, is preferable here to fitting an extreme-value distribution to a ten-year record, and it keeps the evaluation in the same units as the risk argument of Section 1.1.1. A full-series root-mean-square error (RMSE) alone would not establish a good storm-surge correction; the evidence must also show what the model does during the events that already carry an authoritative coastal-risk interpretation.

Figure 1.4 gives the end-to-end reading of the work. It separates the construction and diagnosis of the dynamic residual from the temporal regression step and from application at test stations, and it marks where the validation loop ends: validation observations select the training procedure, whereas the reconstruction and evaluation blocks use the frozen final model.

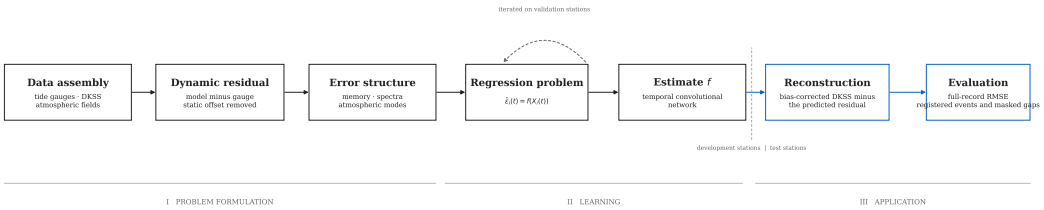


Figure 1.4: End-to-end workflow from data assembly to post-processed sea level and evaluation. The first three blocks formulate and diagnose the dynamic-residual problem; the temporal convolutional network estimates the residual from histories ending at the prediction time; reconstruction subtracts that estimate from the bias-corrected hydrodynamic level. The dashed boundary distinguishes development from the test stations.

1.6 Structure of the thesis

Chapter 2 defines the input data and the supervised target: the 28-station tide-gauge network, the Danish Meteorological Institute (DMI) operational sea-level product DKSS, the DMI HIRLAM/HARMONIE atmospheric forcing, the quality-control rules, and the separation of the raw model-minus-gauge residual into a static reference-level bias and the dynamic residual.

Chapter 3 interprets those records physically. It connects wind, storm evolution, temporal memory and domain-scale forcing modes to the measured residual, introduces the KDI registered-event archive, and positions the problem relative to earlier correction studies.

Chapter 4 presents the methodology: the input windows ending at the prediction time, the predictor representation, the point-output architecture, the extreme-oriented loss that encodes the asymmetry of Section 1.1.1, and the spatial train–validation–test protocol.

Chapter 5 reports and discusses the results. Complete-record performance is read together with the station-level improvement map and the registered KDI events. Gap filling and geographical specialisation appear as secondary applications on the same test observations.

Chapter 6 summarizes the findings, discusses the limitations of the approach and outlines future work.

Chapter 2

Dataset and target variable

2.1 Data foundation of the correction problem

Three sources are combined here: tide gauges provide the local observational reference, the hydrodynamic model provides the physically based sea-level estimate to be corrected, and the atmospheric fields describe the forcing under which the discrepancy between them evolves. This chapter specifies their temporal overlap, spatial collocation and quality control, and derives the supervised target from them.

One difficulty shapes every choice that follows: the three sources have different spatial supports. DKSS and the atmospheric archive are continuous gridded fields, while the model error can be observed only where a tide gauge keeps a quality-controlled point record. Two requirements follow. Unlike sources have to be collocated without implying that gauge information exists between stations, and a fixed reference-level difference has to be separated from the time-varying discrepancy that might transfer across the network. The diagnostics at the end of the chapter describe the target that results; Chapter 3 takes up its physical interpretation.

2.2 Raw data sources

The dataset combines tide-gauge observations, DMI operational hydrodynamic-model output (hereafter DKSS) and DMI HIRLAM/HARMONIE atmospheric forcing. The three differ in native period, resolution and spatial support: the gauges are coastal point records, while DKSS and HIRLAM/HARMONIE are gridded time-dependent fields. Each is resampled once onto a shared hourly axis and kept at full length. What this chapter describes is therefore a set of continuous records; the fixed-length windows the network actually consumes are formed from them later and are defined in Section 4.2.

The timeframe used in the thesis is

$$\text{2013-10-01 07:00} \leq t \leq \text{2022-12-31 23:00}. \tag{2.1}$$

The start is set by DKSS and the end by the local DMI HIRLAM/HARMONIE

archive. Over that interval the aligned record covers 81 089 hourly timestamps, with DKSS sea level on 81 002 finite sea cells and four HIRLAM/HARMONIE fields on a 113×151 crop. The source archives extend beyond this common interval, as summarized in Table 2.1; only their overlap enters the model dataset.

Table 2.1: Main data sources used to construct the residual post-processing dataset.

Layer	Source	Native availability	Resolution	Variables and units
Tide gauges	Danish water-level archive	Mostly 1 Jan 2013–1 Jan 2023 boundary time	10 min	Observed sea level [cm], converted to metres
HDM	DKSS2019 and DKSS2020	1 Oct 2013 07:00–11 Nov 2021 06:00; 1 Oct 2021 07:00–17 Jun 2024 00:00	Hourly	Modelled sea level, p82 [m]
Atmosphere	DMI HIRLAM/HARMONIE archive	1 Jan 2013 00:00–31 Dec 2022 23:00	Hourly	SLP [Pa], t2m [K], u_{10} and v_{10} [$\mathrm{m\,s^{-1}}$]

2.3 Station network

The experimental network is 28 tide gauges in the inner Danish waters and the western Baltic, drawn from a national inventory of roughly 70. A station is kept only if it lies inside the study domain and carries at least 90% valid hourly values. Records are excluded if they lie outside the study area (Bornholm, the Skagerrak transition and the North Sea coast), contain multi-year gaps, or duplicate a retained record. Unless stated otherwise, every station-based figure and table in this thesis refers to these 28 stations.

Their vertical reference is common. Since 2002 every Danish tide gauge has been referred to the national height reference DVR90 (Dansk Vertikal Reference 1990), a single levelling datum tied to the 1990 mean sea level (Poulsen et al. 2024), so the records do not sit on arbitrary per-station zeros. The hydrodynamic model keeps its own vertical reference, and the offset between the two is what the per-station constant β_i of Section 2.8 absorbs. It is removed before the dynamic correction is learned.

Figure 2.1 shows the complete retained network without imposing an additional geographical classification. The mutually disjoint training, validation and test roles are introduced only with the experimental protocol in Chapter 4.

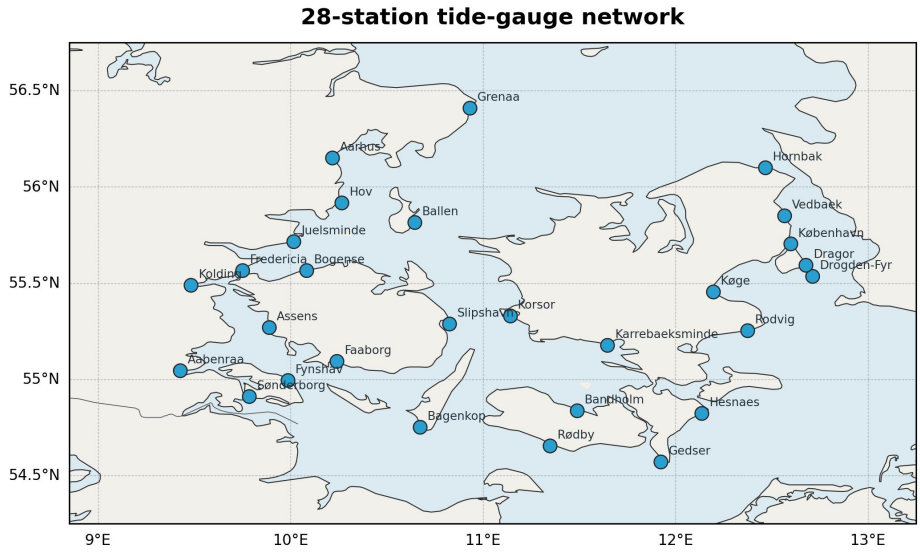


Figure 2.1: The 28-station experimental network used throughout the thesis. Uniform circular markers show the retained tide gauges; station names make the subsequent train–validation–test split and station-level results directly traceable.

Table 2.2: Summary diagnostics of the 28 tide-gauge stations. Coverage is the fraction of hourly collocated targets retained after target quality control. Mean raw residual, RMSE and MAE use $r = \eta_{\mathrm{HDM}} - \eta_{\mathrm{TG}}$ before removal of the reference-level bias; a positive value indicates that the model is high relative to the gauge. KDI events are the registered extremes listed by Kystdirektoratet during 2013–2022. A dash denotes a station that is not covered by the authority's report, not a zero event count.

Station	ID	Coverage [%]	Mean raw residual [cm]	RMSE [cm]	MAE [cm]	KDI events
Grenaa	22121	98.2	9.2	11.9	10.0	12
Aarhus	22331	99.3	3.3	9.4	7.5	4
Hov	22598	98.7	7.6	11.1	9.1	–
Juelsminde	23132	92.5	1.7	12.7	10.3	21
Fredericia	23293	99.0	-4.0	11.9	9.4	4
Ballen	27084	97.2	6.5	10.6	8.7	11
Bogense	28003	96.1	6.0	10.5	8.5	19
Kolding	23322	91.7	-0.3	12.3	8.9	9
Assens	28366	93.9	4.3	7.8	6.3	11
Faaborg	28397	95.6	3.8	8.1	6.6	8
Fynshav	26457	99.4	6.8	9.7	8.0	3
Sønderborg	26473	96.5	2.7	7.6	6.1	10
Aabenraa	26239	97.5	2.8	8.0	6.4	8
Bagenkop	28548	97.1	1.0	7.8	6.2	8
Rødby	31573	99.6	7.0	9.6	8.0	3
Hornbæk	30017	98.8	12.5	15.4	13.3	5
Vedbæk	30202	93.3	11.9	13.8	12.1	–
København	30336	98.6	9.0	13.0	11.2	5
Dragør	30361	98.9	9.3	15.8	12.2	–
Drogden-Fyr	30357	99.1	11.2	16.3	13.0	6
Køge	30478	98.5	9.9	11.8	10.3	3
Rødvig	31063	99.1	8.6	10.6	9.1	10
Gedser	31616	96.4	5.9	9.2	7.4	2
Hesnæs	31494	97.2	5.4	14.9	11.4	8
Slipshavn	28234	98.5	-1.9	8.3	6.6	5
Korsør	29393	99.7	1.6	8.3	6.5	5
Karrebæksminde	31171	98.6	9.6	13.4	11.1	12
Bandholm	31342	97.3	7.0	10.8	8.9	–

Table 2.2 summarizes the 28 stations through hourly target coverage and raw model-observation residual statistics. Its last column records the number of 2013–2022 events present in Kystdirektoratet's station sheet; a dash indicates that the report does not cover that station.

2.4 Tide-gauge observations

The observed sea level at a tide-gauge (TG) station i is denoted by

$$\eta_i^{\mathrm{TG}}(t). \tag{2.2}$$

The national files provide water level in centimetres at 10 min resolution over the calendar years 2013–2022, with 1 January 2023 retained only as the closing boundary timestamp. Values are converted to metres before the residual is formed. Each hourly value is taken from the exact timestamp where one exists; failing that, from $t - 10\,\mathrm{min}$ and then from $t + 10\,\mathrm{min}$. A provenance flag records which of the three was used, and an hour with none of them remains explicitly missing.

Observations are never averaged between gauges or interpolated in space to fill a missing target. Differences in record completeness are carried through in the station-specific masks and are visible in Table 2.2.

The observations vary systematically across the network. Figure 2.2 maps two complementary station summaries over all quality-controlled overlapping hours. The mean is expressed in the common DVR90 reference and must therefore not be interpreted as a present-day mean dynamic topography, because DVR90 is tied to the 1990 mean sea level and still incorporates the geoid and local vertical land motion. The standard deviation is invariant to a constant datum offset and shows the spatial variation in observed sea-level variability.

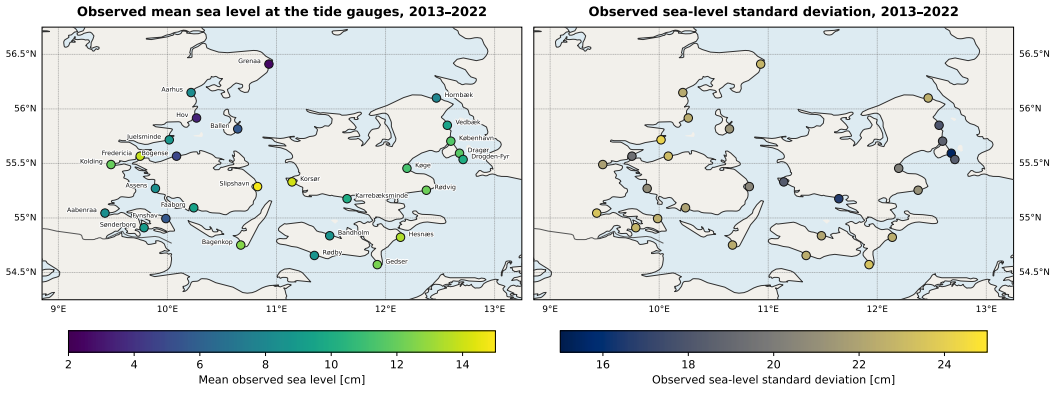


Figure 2.2: Observed sea-level statistics at the complete 28-station network over 2013–2022, using all quality-controlled hours in the common dataset interval. Left: station mean sea level in the Danish national height reference DVR90, to which all Danish tide-gauge records are referred. Right: station standard deviation, which is independent of a constant datum offset.

2.5 Hydrodynamic model data

The hydrodynamic-model sea level is denoted by

$$\eta^{\mathrm{HDM}}(\phi, \lambda, t), \tag{2.3}$$

where ϕ and λ are latitude and longitude. Two hourly archives of the same DKSS grid and variable are joined: DKSS2019 supplies the record from October 2013 through its end in November 2021, and DKSS2020 supplies the later period. DKSS2019 has priority where the two archives overlap, and a provenance flag preserves which model version supplied each hour. Within the thesis interval this gives 71 076 DKSS2019 hours, 9 941 DKSS2020 hours and 72 hours with no model field.

The native grid has shape 482×396, with a resolution of about $30''$ in latitude and $50''$ in longitude, and covers approximately 53.59–57.60°N and 9.34–14.83°E. Land and inactive cells are not treated as zero sea level. The fixed wet mask leaves 81 002 finite sea cells, stored as

$$\eta_c^{\mathrm{HDM}}(t), \qquad c = 1, \ldots, 81\,002, \tag{2.4}$$

where c indexes finite sea cells rather than the full land–sea grid. This representation avoids storing land cells while preserving the full-domain sea-level field.

At a tide-gauge station, the model value is obtained by spatial collocation. The station coordinates are first compared with the wet HDM cells using geographical distance. In practice this is a nearest-wet-cell operation based on spherical distance, followed by a 3×3 finite-cell median around the selected wet cell:

$$\eta_i^{\mathrm{HDM}}(t) = \mathrm{median}\Big\{\eta_c^{\mathrm{HDM}}(t) : c \in \mathcal{N}_{3\times3}(i),\ \eta_c^{\mathrm{HDM}}(t) \text{ finite}\Big\}. \tag{2.5}$$

A single coastal grid cell can be unduly sensitive to land-mask details and coastline geometry, which is why a local median is taken instead. At a nominal grid spacing near 0.9 km the operation stays local, but one marginal wet cell no longer decides the value.

Figure 2.3 shows the field on the native wet grid. Set beside the station map in Figure 2.1, it makes the support mismatch plain: a continuous two-dimensional product on one side, 28 coastal points on the other. The model can be evaluated against observations only at those points, through the collocation of Equation 2.5.

2.6 Atmospheric forcing data

The raw DMI HIRLAM/HARMONIE archive consists of 120 monthly files covering January 2013 through December 2022. Four hourly near-surface fields are retained:

$$\mathbf{F}(t) = [p_{\mathrm{SL}}(t), T_{2\mathrm{m}}(t), u_{10}(t), v_{10}(t)]\,. \tag{2.6}$$

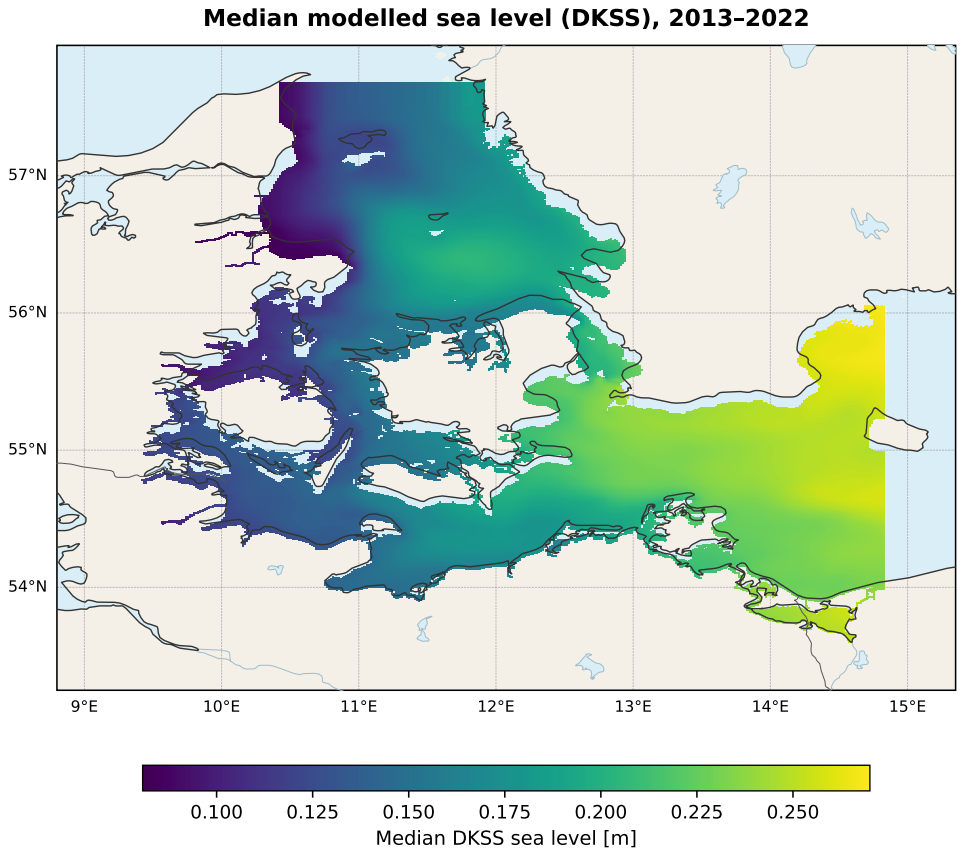


Figure 2.3: Temporal median of the hourly DKSS sea-level field over 2013–2022. Colour encodes the modelled value at each of the 81 002 wet grid cells.

Here p_{SL}, $T_{2\mathrm{m}}$, u_{10} and v_{10} denote sea-level pressure, 2-m air temperature, and the zonal and meridional 10-m wind components, respectively. The retained atmospheric domain is a 113×151 crop covering approximately 53.03–58.63°N and 8.33–15.83°E. This is wider than the DKSS grid, so approaching pressure systems and wind fields remain represented before they reach the coastal stations.

These four fields enter the analysis through two spatial supports: they are compressed into domain-scale EOF scores (Section 3.6) and are also extracted at the atmospheric grid cell nearest each tide gauge.

Sea-level pressure is converted from pascals to an anomaly relative to 101 325 Pa for compact storage, whereas 2-m temperature remains in kelvin and the wind components in metres per second. Retaining u_{10} and v_{10} rather than wind speed alone preserves direction. From them, wind speed and wind-stress quantities can be written as

$$|\mathbf{U}_{10}(t)| = \sqrt{u_{10}(t)^2 + v_{10}(t)^2}, \tag{2.7}$$

and

$$\boldsymbol{\tau}(t) = \rho_a C_D |\mathbf{U}_{10}(t)|\, \mathbf{U}_{10}(t), \tag{2.8}$$

where ρ_a is air density and C_D a drag coefficient. Neither derived quantity is supplied directly to the network, which receives u_{10} and v_{10} themselves. They are

introduced here only because wind magnitude and orientation are useful for the physical diagnostics developed in Chapter 3. Raw-channel and missing-value checks follow in Section 2.7.

2.7 Quality control and despiking

Quality control is source specific. A missing DKSS hour, an isolated gauge failure and an atmospheric-channel error cannot be treated by the same rule. The procedure therefore checks the model, observations and atmospheric predictors separately and subsequently combines them through the supervised target mask.

A further difficulty is that genuine high-water events and non-physical outliers both lie in the tails of the sea-level distribution. A simple removal based on fixed thresholds or upper and lower quantiles would therefore risk censoring precisely the storm-surge conditions that the model is intended to learn. Magnitude alone is not sufficient to distinguish the two: an instrumental spike can reach the same water level as a real storm-surge peak. Their temporal structure is different, however. Genuine surges generally develop and relax coherently over several consecutive hours, whereas instrumental outliers are more often characterized by an abrupt departure from neighbouring values followed by a rapid return.

For this reason, the tide-gauge quality control is designed around local temporal consistency rather than the absolute magnitude or quantile of an observation. Extreme values are retained when they form part of a physically coherent event trajectory, while short isolated departures or jump-and-return patterns are treated as candidate artefacts.

2.7.1 Tide-gauge and target checks

An hourly target is initially valid only when the selected tide-gauge value and the collocated DKSS value are both finite and the stored raw residual $r_i(t) = \eta_i^{\mathrm{HDM}}(t) - \eta_i^{\mathrm{TG}}(t)$ agrees numerically with $\eta^{\mathrm{HDM}} - \eta^{\mathrm{TG}}$. Provenance flags, leaked missing-value sentinels and hard physical ranges are checked before statistical despiking.

The Hampel component tests whether a value is locally isolated. Here x denotes the series being checked, either the tide-gauge value or the residual. For the valid neighbours within $k = 6$ hours of t, excluding the centre point itself, define

$$W_t^{(-)} = \{x_{t-k}, \ldots, x_{t-1}, x_{t+1}, \ldots, x_{t+k}\}, \tag{2.9}$$

and

$$m_t = \mathrm{median}(W_t^{(-)}), \qquad S_t = 1.4826\ \mathrm{median}_{x \in W_t^{(-)}} |x - m_t|. \tag{2.10}$$

The implemented acceptance threshold is

$$|x_t - m_t| < \max(0.25\,\mathrm{m},\, 6S_t)\,, \tag{2.11}$$

and only flagged runs of at most three hours are removed. The 1.4826 factor places

the median absolute deviation (MAD) on a Gaussian-equivalent scale; it does not assume that the sea-level distribution itself is Gaussian.

Hampel is complemented by a station-adaptive jump test applied to both the gauge and residual. If $|\Delta x|$ denotes the valid one-hour jumps at one station, its threshold is

$$\theta_s = \max(0.4\,\mathrm{m},\ \mathrm{median}(|\Delta x|) + 8 \cdot 1.4826\,\mathrm{MAD}(|\Delta x|),\ 2q_{0.999}(|\Delta x|))\,. \qquad (2.12)$$

A one-hour value is rejected only if it jumps away from both neighbours by at least θ_s while those neighbours agree within $\max(0.30\,\mathrm{m}, 0.25\theta_s)$, and the same return-to-baseline logic covers coherent two-hour bursts. What marks the failure is the short symmetric departure and return, never the magnitude of the water level.

Across the whole record the combined rules mask 101 records inside the final 28-station population. The masked fraction is negligible. Raw arrays are left unchanged and affected targets simply become unavailable; nothing is interpolated into their place.

Figure 2.4 illustrates the outcome at Grenaa. A one-hour value reaching 1.17 m is masked because its departure from the local median exceeds the production Hampel threshold.

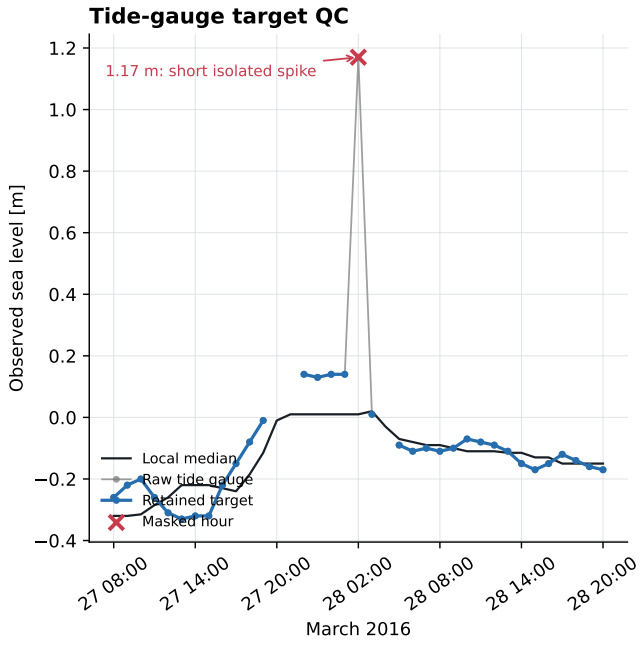


Figure 2.4: Tide-gauge quality-control example at Grenaa. The observation at 02:00 on 28 March 2016 reaches 1.17 m for one hour and is masked by the production Hampel test. Grey is the original gauge series, black its local median, blue the retained target and the red cross the masked hour.

2.7.2 Hydrodynamic-model and atmospheric checks

A provenance value of 255 marks the 72 hours without a DKSS field; these are never bridged by temporal interpolation. Spatial NaNs over land are handled by the fixed

wet mask rather than filled, and the station collocation uses only finite cells in the local 3×3 neighbourhood. This separates a missing model time from the ordinary land mask of a coastal grid.

DMI HIRLAM/HARMONIE required a different correction. The raw archive contained a documented interval in which sea-level pressure (SLP) and t2m were assigned to the wrong channels. Since pressure is stored in pascals and temperature in kelvin, the condition SLP < 1000 together with t2m > 1000 identifies the exchange without an empirical threshold fitted to the weather. The two values are swapped back before pressure is converted to an anomaly. The raw-data audit locates the affected station series from 5 April 2018 06:00 to 23:00, a total of 18 hours. After this repair, the atmospheric fields are audited against broad physical ranges: SLP 85 000–110 000 Pa, t2m 180–330 K and $|u_{10}|, |v_{10}| \leq 75\ \mathrm{m\,s^{-1}}$.

Remaining non-finite atmospheric cells are predictor gaps, not missing observations, and they are filled in three recorded passes: nearest finite spatial neighbour, linear interpolation through time, then a final spatial pass. The temporal pass bridges four six-hour gaps, 24 hours in total, using fields on either side. This is the only source in the thesis for which anything is imputed. It happens before EOF fitting and local extraction, and the atmospheric values extracted at the 28 stations are finite throughout.

No missing tide-gauge or DKSS target is filled in this way. Reaching across a gap is a small but real limitation for a strictly online reading of those 24 hours, and it is documented here.

2.8 Target variable: the dynamic residual

With the sources aligned and quality controlled, the supervised target can be defined. The raw residual at each valid station–time pair is

$$r_i(t) = \eta_i^{\mathrm{HDM}}(t) - \eta_i^{\mathrm{TG}}(t). \tag{2.13}$$

A positive value means that DKSS overestimates the tide gauge, while a negative value means that it underestimates it. The learning target is not $r_i(t)$ itself. The station-dependent static bias is first estimated as the mean raw residual over the quality-controlled record $\mathcal{T}_{\mathrm{bias},i}$ of that station,

$$\beta_i = \frac{1}{|\mathcal{T}_{\mathrm{bias},i}|} \sum_{t \in \mathcal{T}_{\mathrm{bias},i}} r_i(t), \tag{2.14}$$

and the dynamic residual is

$$\varepsilon_i(t) = r_i(t) - \beta_i. \tag{2.15}$$

The reference-level bias β_i is not learned. It is estimated once for each station from every quality-controlled residual available over the complete common 2013–2022 record (Jahanmard et al. 2023). Removing it isolates the time-varying residual

$\varepsilon_i(t)$ that the post-processing model predicts. The same β_i enters both the bias-corrected baseline and the post-processed level, so every improvement reported later compares the dynamic discrepancy after removal of the same static station-dependent component.

Figure 2.5 maps β_i across the network. The spread between stations rules out a single domain-wide constant: the reference-level difference is local, and it is estimated station by station before any network weight is fitted.

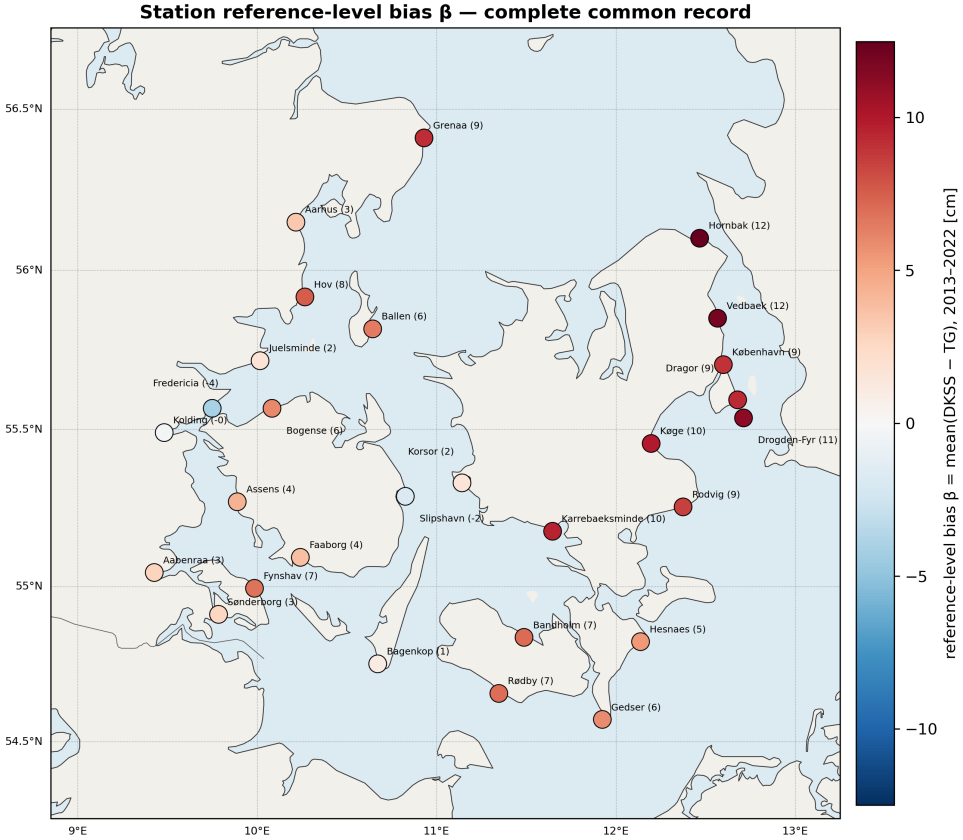


Figure 2.5: Reference-level bias β_i, defined as the quality-controlled mean DKSS–tide-gauge residual over the complete common 2013–2022 record. Positive values mean that the DKSS reference level is above DVR90 at the collocated station. This static calibration is removed from both the baseline and the final post-processed series.

A supervised target exists only where both inputs are finite and the point has passed the spike checks of Section 2.7. Writing $q_i(t) = 1$ for a point retained by quality control, the target mask is

$$\mathcal{M}_i(t) = \mathbb{1}\Big[\eta_i^{\mathrm{TG}}(t) \text{ finite } \wedge\ \eta_i^{\mathrm{HDM}}(t) \text{ finite}\Big]\, q_i(t). \tag{2.16}$$

Missing target values are never imputed; samples with masked targets are excluded from the loss computation rather than replaced by a synthetic value.

The final corrected sea level is

$$\hat{\eta}_i^{\mathrm{corr}}(t) = \eta_i^{\mathrm{HDM}}(t) - \beta_i - \hat{\varepsilon}_i(t), \tag{2.17}$$

so the baseline against which every result is measured is the bias-corrected model $\eta_i^{\mathrm{HDM}}(t) - \beta_i$, never the unadjusted hydrodynamic output.

Figure 2.6 gathers these definitions into one chain. Everything above the model box belongs to the target construction described in this chapter: the two sea-level series, their difference, and the decomposition into a station constant and a time-varying target. The predictor history and neural network are included only to make the target's later role explicit; they are specified in Chapter 4. The hydrodynamic level re-enters during reconstruction because the network corrects DKSS rather than replacing it.

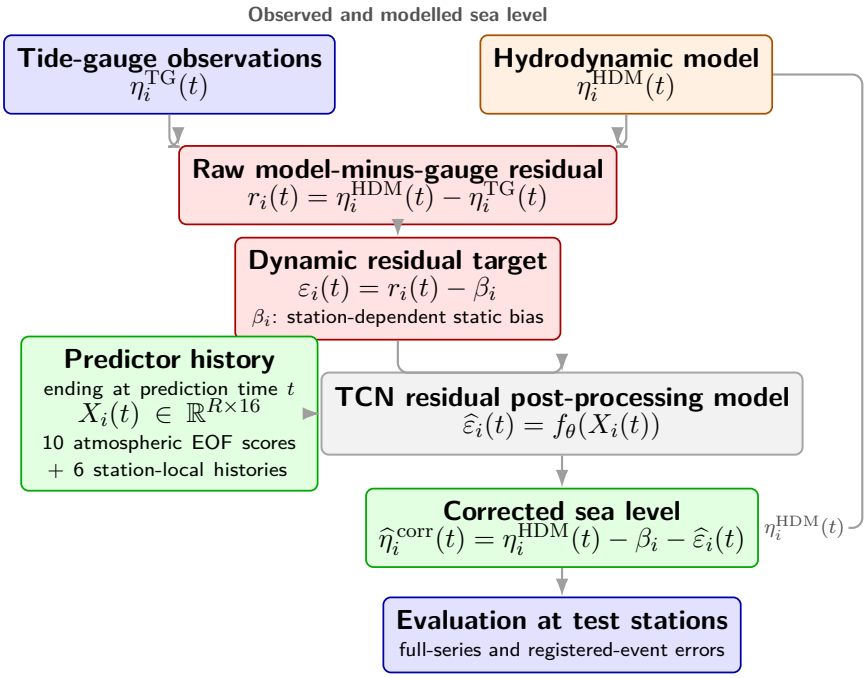


Figure 2.6: The residual post-processing chain. Tide-gauge observations and collocated DKSS output define the raw model-minus-gauge residual $r_i(t)$; the station mean β_i, estimated over the complete quality-controlled record, is removed to leave the dynamic target $\varepsilon_i(t)$. The 16-channel predictor history ends at prediction time t, the network estimates $\widehat{\varepsilon}_i(t)$, and reconstruction subtracts both the static and the predicted dynamic component from the hydrodynamic level before evaluation at the test stations.

2.8.1 Dynamic-residual sign and distribution over the complete record

A mean says nothing about spread or tails, so Figure 2.7 shows the full empirical distribution at Grenaa, Faaborg and Drogden-Fyr, three widely separated gauges that recur as reference stations throughout the thesis. Every quality-controlled value is included, and each panel annotates the full-record mean, standard deviation and 1st–99th percentile interval.

Figure 2.8 extends the comparison to all 28 stations without overlaying 28 histograms. The network is sorted by standard deviation; thin lines show the 1st–99th percentile range, thick lines the interquartile range and dots the mean. Standard deviation runs from roughly 6.3 to 13.8 cm, and the unequal left and right tails show

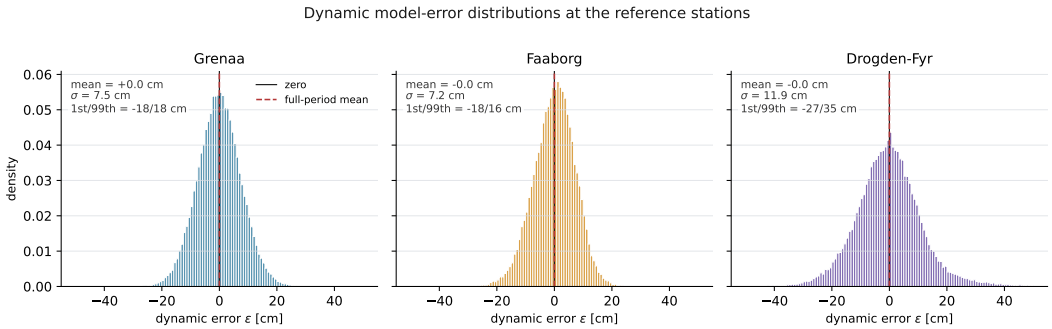


Figure 2.7: Complete empirical distribution of ε_i at Grenaa, Faaborg and Drogden-Fyr, shown in blue, orange and purple, respectively. The black line is zero and the dashed red line is the full-period mean. Each panel annotates the mean, standard deviation σ and 1st–99th percentiles.

station-dependent asymmetry even where the full-period mean sits close to zero. The constructed target varies across the network in spread and in tail shape, not only in level.

2.8.2 Frequency structure of the target

The same three reference stations provide a frequency-domain description of the constructed target in Figure 2.9. The upper panel shows the complete hourly dynamic-residual record, while the lower panel summarizes recurrent periods after de-meaning the longest segment whose gaps of at most 24 hours can be interpolated. Because subtraction of the station bias β_i affects only the zero-frequency component, all non-zero-frequency structure belongs to the dynamic target $\varepsilon_i(t)$.

The three stations show a common but station-dependent spectral structure. Narrow short-period peaks occur near semidiurnal and diurnal periods, while broader variability extends across multi-day time scales. Their amplitudes differ substantially between stations, consistent with the spatial differences already visible in the residual distributions.

At this stage the spectra document what the dataset construction produced, and nothing more. Their physical interpretation, including the relation to astronomical periods, synoptic forcing and residual memory, is discussed in Chapter 3, where the same structure is examined across the whole station network.

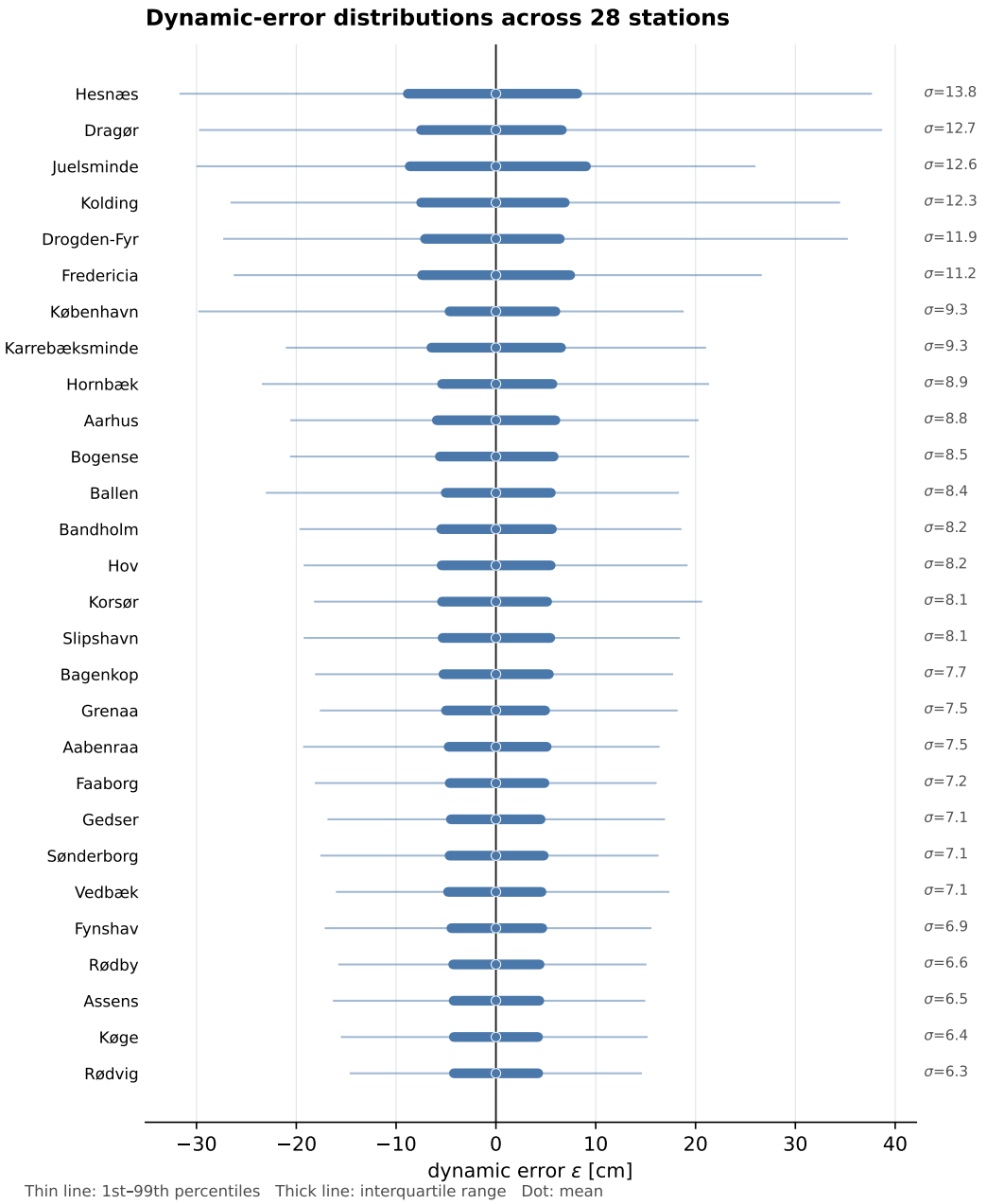


Figure 2.8: Distribution summary of the dynamic residual ε_i at all 28 stations, ordered by decreasing standard deviation. Thin lines span the 1st–99th percentiles, thick lines the interquartile range and dots the full-period mean. Uniform colour preserves the network-wide framing; the right-hand labels report σ.

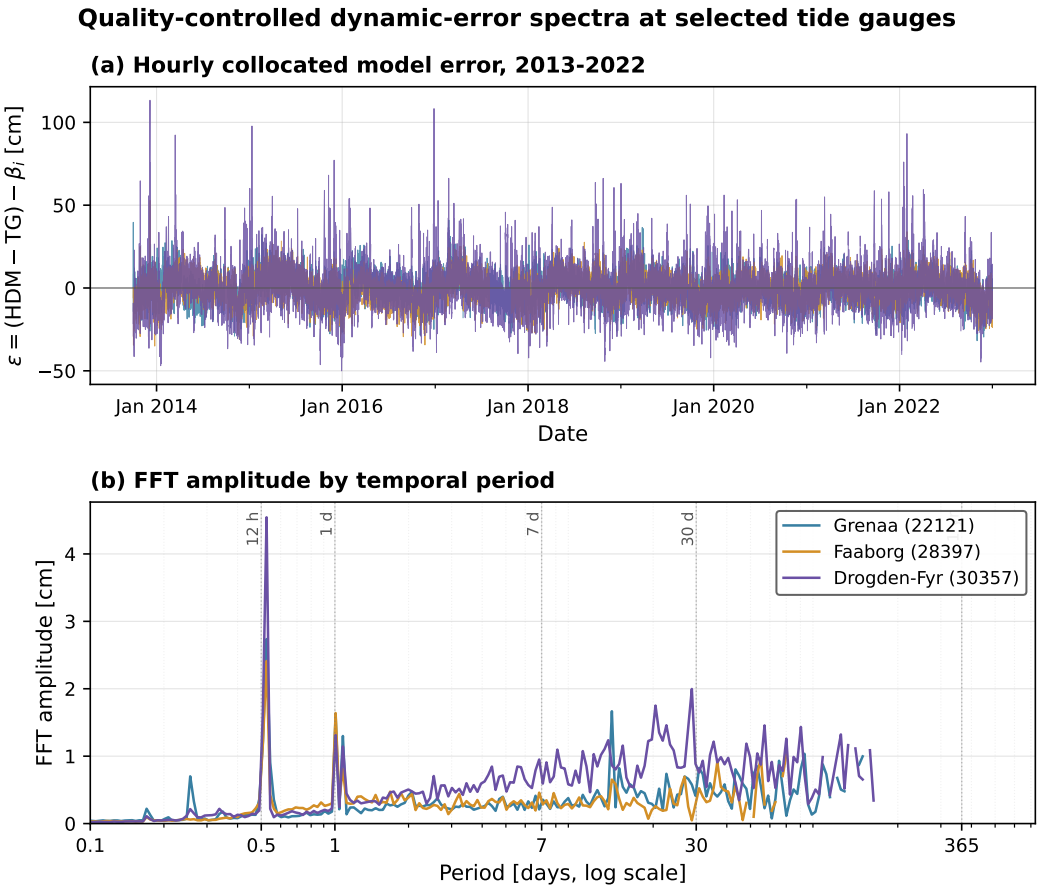


Figure 2.9: Quality-controlled dynamic residual and its Hann-windowed fast Fourier transform (FFT) amplitude at Grenaa, Faaborg and Drogden-Fyr, shown in blue, orange and purple, respectively. The lower panel expresses frequency as period on a logarithmic axis; guides mark 12 hours, 1, 7 and 30 days and 1 year. The figure documents shared short-period structure together with station-dependent variability at longer time scales.

2.9 Summary

The heterogeneous data sources are now aligned on a common hourly axis, DKSS is collocated to each tide gauge through local wet-cell aggregation, and isolated non-physical spikes are removed without censoring genuine high-water events. The supervised target is the station-bias-removed dynamic residual $\varepsilon_i(t)$. Where a gauge or hydrodynamic-model value is missing, the target remains missing: no interpolation manufactures a label and none is created between stations.

Removing the station-dependent mean does not leave a homogeneous target: spread, tails and temporal structure still differ from gauge to gauge. Chapter 3 takes up where that remaining structure comes from, what time scales it occupies, and which parts of it could plausibly be learned from the available hydrodynamic and atmospheric histories.

Chapter 3

Physical interpretation of hydrodynamic-model error

3.1 Mechanisms behind the residual

Chapter 2 built the dynamic residual $\varepsilon_i(t)$ by removing the station-dependent reference level. What it did not establish is whether anything physically interpretable survives that subtraction. This chapter asks which mechanisms can leave a signature in the remainder, on what time scales they act, and how strongly they depend on the meteorological and hydrodynamic state.

In the Danish waters, sea-level variability results from the combined effect of several processes. At short to synoptic time scales, the main contributions can be represented conceptually as

$$\eta(t) = f(\eta_{\mathrm{IB}}(t),\, \eta_{\mathrm{wind}}(t),\, \eta_{\mathrm{boundary}}(t),\, \eta_{\mathrm{tide}}(t),\, \eta_{\mathrm{local}}(t))\,, \tag{3.1}$$

where η_{IB} denotes the inverse-barometer response, η_{wind} the wind-driven setup or setdown, η_{boundary} the propagation of external sea-level signals from neighbouring basins, η_{tide} the astronomical tide, and η_{local} the influence of coastal geometry, bathymetry and unresolved local processes. To leading order these contributions may be viewed separately, but several interact nonlinearly, notably through tide–surge interaction and the dependence of wind setup on the antecedent basin state. For this reason, f is written as a general function rather than as a simple sum.

Equation (3.1) is not used as a numerical model. It is a way of organizing the mechanisms that can contribute to a time-dependent discrepancy between the hydrodynamic model and the observed coastal water level.

On storm-surge time scales, pressure and wind are the dominant atmospheric drivers. Under the static inverse-barometer approximation, a pressure decrease of 1 hPa corresponds to roughly 1 cm of sea-level rise. In shallow and semi-enclosed waters this is only a first-order reference: the actual response also depends on wind direction and duration, coastline orientation, basin geometry and the capacity of the system to accumulate or export water.

The Danish waters connect the North Sea and the semi-enclosed Baltic through the Kattegat, the Great Belt, the Little Belt and the Sound. Their narrow and shallow geometry modifies the propagation of large-scale signals and controls exchange between the two basins (Leppäranta and Myrberg 2009). Each of the 28 stations sits in its own combination of open-boundary influence, wind exposure, local bathymetry, strait geometry and basin state, so the same atmospheric forcing need not produce the same coastal response everywhere.

The response also depends on what happened before the hour in question. Persistent winds redistribute water over several days, and once displaced it can keep oscillating after the local forcing has weakened. Instantaneous atmospheric conditions do not fix the relevant state.

The January 2017 event introduced in Section 1.1.1 illustrates this preconditioning mechanism particularly clearly. Its exceptional coastal water levels developed from an elevated Baltic basin state rather than from an equally exceptional instantaneous local wind forcing (She and Nielsen 2019; Andrée et al. 2023). Here the event is relevant not as another risk example, but because it shows why coastal response may depend on forcing and hydrodynamic history over time scales longer than the local storm peak.

The astronomical tide is comparatively small at many western Baltic stations, but it is not absent from the residual dynamics. As shown later in this chapter, narrow spectral features remain close to astronomical periods even after the station-dependent mean has been removed. These periods are used only to interpret the residual spectrum; tidal information is not supplied explicitly as a predictor to the final post-processing model.

3.2 Physical signatures in the dynamic residual

The previous section listed mechanisms that can, in principle, leave a signature in $\varepsilon_i(t)$. Whether they actually do is an empirical question, and it is the narrower one this section addresses: does the remaining time-varying error carry structure that changes systematically with the evolving physical state, or is it indistinguishable from noise around the removed station constant?

The Danish Straits provide a demanding setting for this test. Their narrow and shallow geometry makes coastal response sensitive to forcing direction, propagation and local representation (Leppäranta and Myrberg 2009; Frishfelds et al. 2025). The diagnostics below do not attempt to attribute each residual to one unique modelling deficiency. Instead, they test whether the observed discrepancy evolves coherently through storm events and whether recurring error patterns can be identified before any neural correction is applied.

Storm Malik at Fredericia, a Little Belt station, provides a direct example. The episode is among the widest in the 2013–2022 register, so its coastal signal is well constrained by the observations, and the discrepancy that survives the removal of β_i can be followed hour by hour through the build-up and the crest.

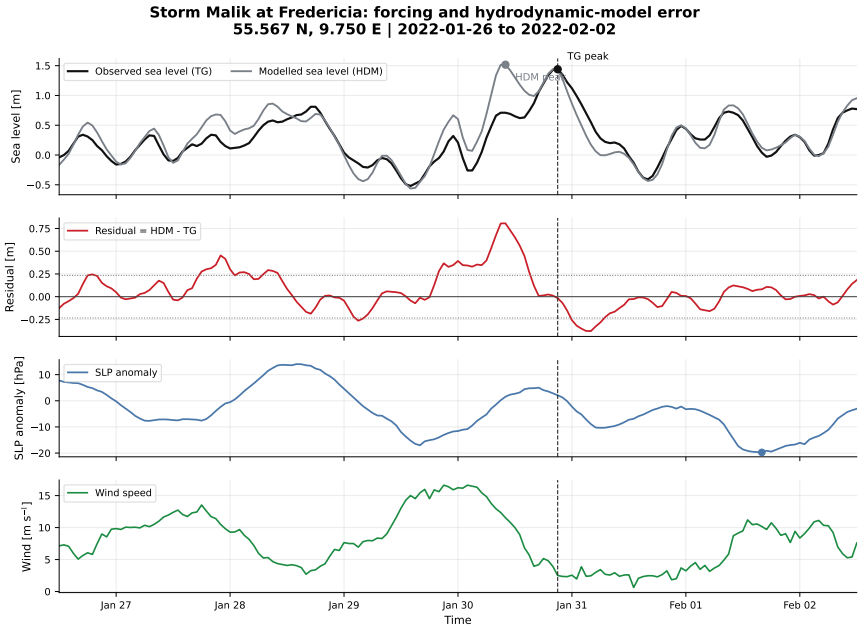


Figure 3.1: Observed forcing and hydrodynamic-model error at Fredericia during storm Malik in late January 2022. From top to bottom: observed and modelled sea level, model-minus-gauge residual, sea-level-pressure anomaly and wind speed. The largest discrepancy develops during the event build-up while winds remain strong.

Figure 3.1 shows that DKSS captures the occurrence and general evolution of the event but anticipates and overestimates the peak. The residual develops over several hours during the storm build-up rather than appearing as one isolated outlier, so its evolution carries information about the trajectory of the forcing and the water level before the crest.

Figure 3.2 broadens this view with four registered station–events selected from the tide-gauge and hydrodynamic-model series alone, without consulting neural-model predictions. The cases represent different discrepancy modes: peak underestimation, peak overestimation, peak-timing mismatch and sustained signed error over the event window. The hydrodynamic series is adjusted by the station mean error over the complete quality-controlled record so that the lower panels isolate the time-varying component available for post-processing.

The examples are descriptive, not probabilistic. They estimate nothing about how often each error mode occurs, and they were not chosen by looking at what the neural network later did with them. They show only that the remaining discrepancy can develop coherently through an event, and that it fails in different ways at different gauges.

That much already rules out one explanation. A constant reference-level offset cannot generate a timing mismatch, change sign during a storm, or evolve alongside the event trajectory. Something time-varying is left, which is what the network-wide diagnostics below set out to characterize.

The raw DKSS error also changes systematically with observed sea level (Figure 3.3), but not in the same way at every coast. Drogden-Fyr shows the steepest

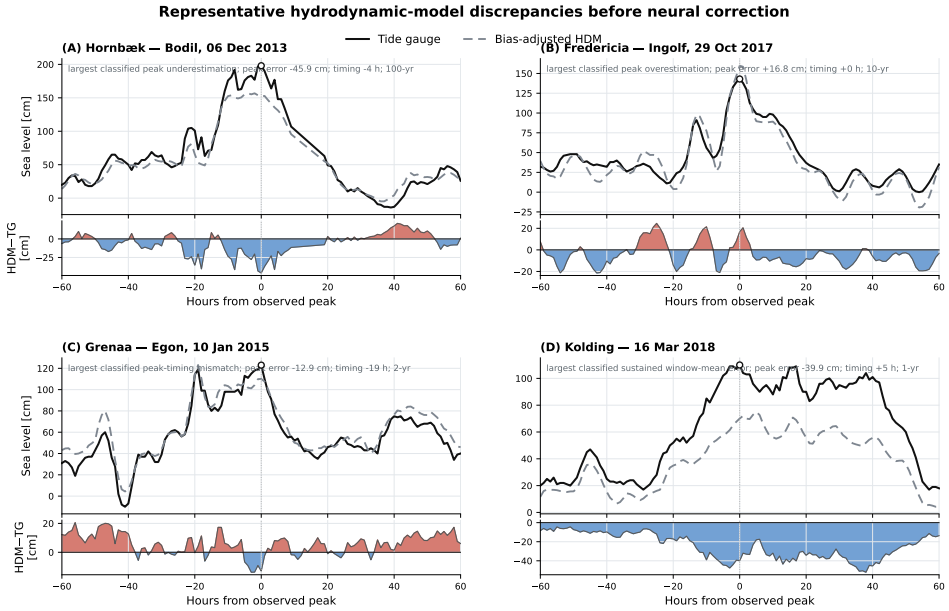


Figure 3.2: Representative high-water events illustrating four discrepancy modes before neural correction: peak underestimation, peak overestimation, timing mismatch and sustained signed error. Black lines show tide-gauge observations and grey dashed lines the bias-adjusted DKSS baseline. Lower panels show baseline-minus-gauge error, with negative values indicating underestimation and positive values overestimation.

dependence: its error is positive at low water, contracts as the level rises and changes sign near the top of the populated range. Grenaa retains a positive reference offset with a gentler slope, whereas Faaborg remains closer to zero through most populated bins. This comparison separates two features of the discrepancy: the station-dependent offset later removed by β_i, and the level-dependent spread that cannot be eliminated by subtracting one constant.

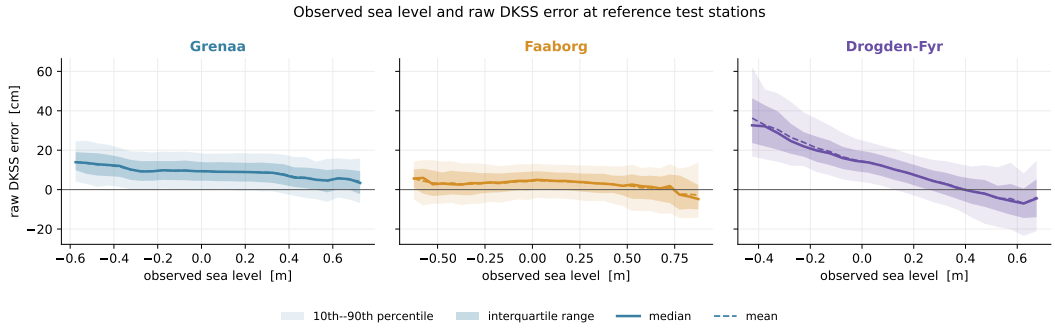


Figure 3.3: Raw DKSS-minus-gauge error conditional on observed sea level at Grenaa, Faaborg and Drogden-Fyr. Levels are grouped in 5-cm bins; only bins with at least 100 valid paired hours are shown. Solid and dashed lines give the median and mean, while the darker and lighter bands give the interquartile and 10th–90th-percentile ranges. Station colours follow the same blue, orange and purple key used in the distribution and frequency diagnostics.

3.3 Wind direction and coastal geometry

The relationship between wind forcing and coastal response is strongly anisotropic in the Danish waters. The same wind speed can produce different sea-level responses depending on the direction from which the wind blows, its persistence and the orientation of the local coastline. A given wind sector may drive water into one basin while favouring export or a weaker response at another station.

Wind direction is examined explicitly for that reason. The convention used here is the meteorological wind-from direction, namely the direction the wind originates from. From the DMI HIRLAM/HARMONIE wind components it is written as

$$\theta_{\text{from}} = \left[\text{atan2}(-u_{10}, -v_{10})\frac{180}{\pi} + 360\right] \bmod 360. \tag{3.2}$$

Thus, for example, a westerly wind denotes wind coming from the west rather than wind directed towards it.

Conditioning the residual on wind speed and wind-from direction provides a compact test of whether particular forcing sectors are associated with systematic model overestimation or underestimation. These summaries are diagnostic rather than predictive: they reveal directional structure in the residual but are not themselves used as correction rules.

Figure 3.4 shows that the dynamic residual is not independent of wind direction. Drogden-Fyr and Dragør exhibit the strongest directional contrasts, with DKSS tending to overestimate for westerly sectors and underestimate for easterly sectors. Grenaa, Faaborg and Bagenkop show smaller conditional means, while Fredericia is intermediate.

The contrast across stations is as important as the sign of any individual sector. A forcing direction associated with a large residual at one station need not produce the same response elsewhere, consistent with the different exposure and geometry sampled by the network. The panels do not identify the unique cause of an individual model failure; they establish that the error contains a reproducible directional dependence.

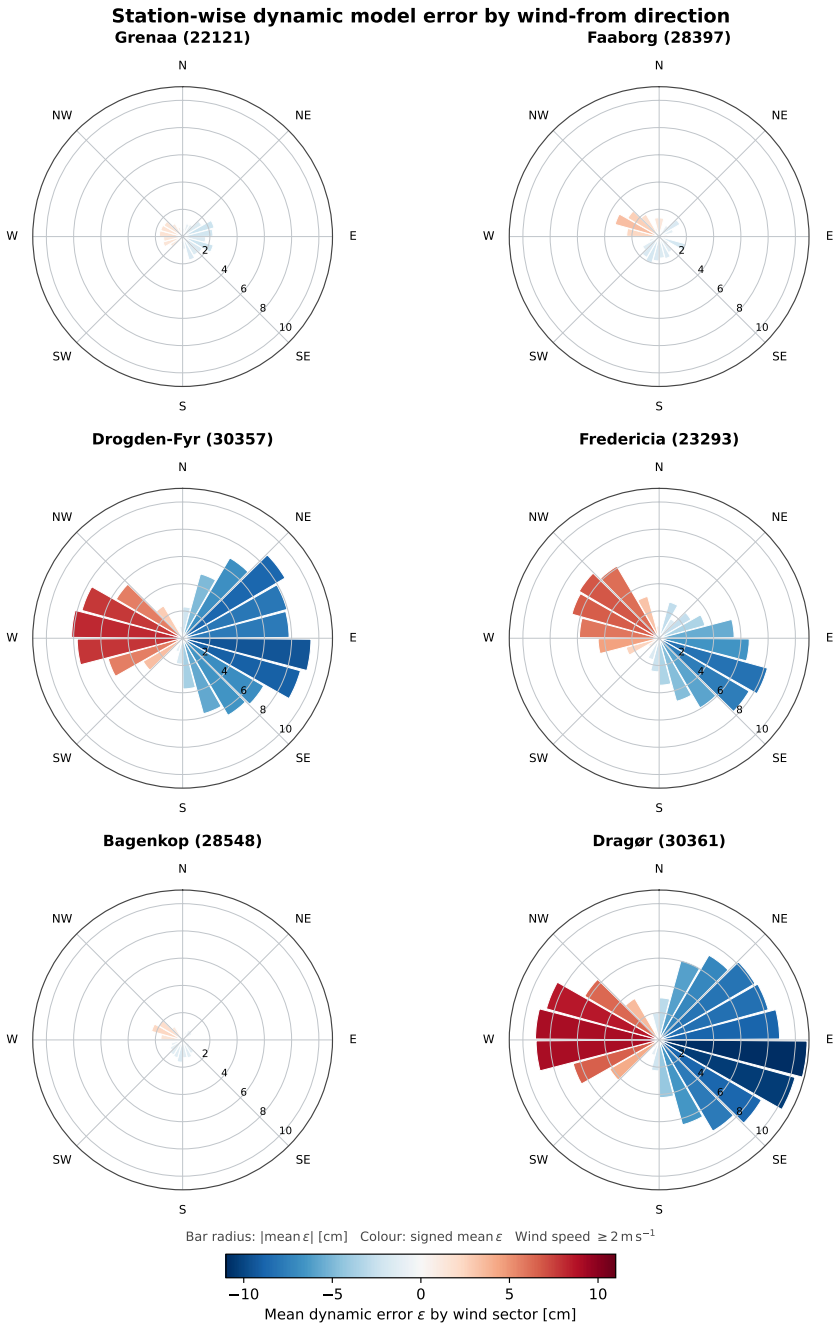


Figure 3.4: Dynamic residual ε_i conditioned on meteorological wind-from direction at Grenaa, Faaborg, Drogden-Fyr, Fredericia, Bagenkop and Dragør. Hours with wind speed of at least 2 $\mathrm{m\,s^{-1}}$ are assigned to 15° sectors. Bar radius gives the absolute sector mean and colour its sign: red denotes mean DKSS overestimation and blue underestimation. Common radial and colour scales allow direct comparison between stations.

3.4 Observed memory and frequency structure

If the dynamic residual evolves together with atmospheric and hydrodynamic episodes, its successive hourly values should not behave as independent samples. The autocorrelation function provides a direct test of that memory.

A useful approximation for the effective sample size is

$$N_{\text{eff}} \approx \frac{N}{1 + 2\sum_{k=1}^{K} \rho(k)}, \tag{3.3}$$

where N is the number of hourly samples and $\rho(k)$ is the autocorrelation at lag k. This expression is not used as an exact estimator in the thesis. It illustrates instead that persistent temporal dependence reduces the number of effectively independent realizations relative to the raw number of hourly timestamps.

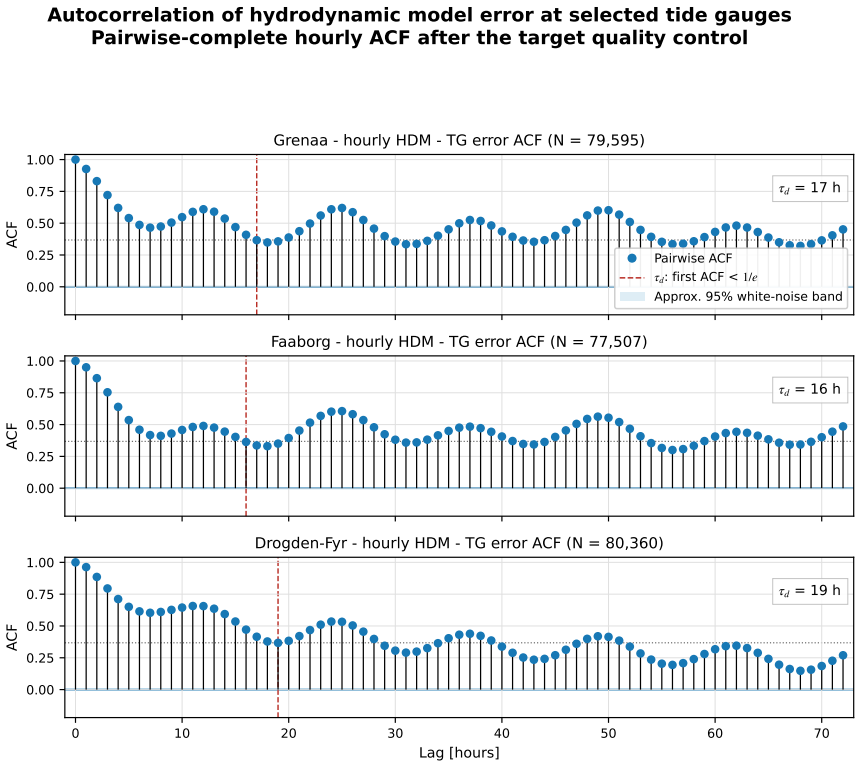


Figure 3.5: Autocorrelation of the hourly hydrodynamic-model error at the three reference stations, computed pairwise-complete after target quality control. The e-folding lag lies between 16 and 19 hours, while correlations remain above the white-noise band for several days.

Figure 3.5 confirms substantial temporal dependence at the three reference stations. The e-folding time is approximately 16–19 hours, while the correlation remains above the white-noise band for several days. Memory on that scale is measurable, and it is far longer than a single hourly step.

Two consequences follow for the experimental design. Predictor histories running over several days can carry information that no instantaneous mapping has access to; and neighbouring hours drawn from one storm are not independent realizations, so they cannot be treated as such when transfer is assessed. The protocol answers the first by supplying long causal windows and the second by splitting complete station records between model development and testing, with the KDI register providing a separate event-based evaluation.

Chapter 2 documented non-zero-frequency structure at three representative sta-

tions. Three gauges are not a network, though, and Figure 3.6 asks whether the same structure survives across all 28 and what physical time scales it occupies.

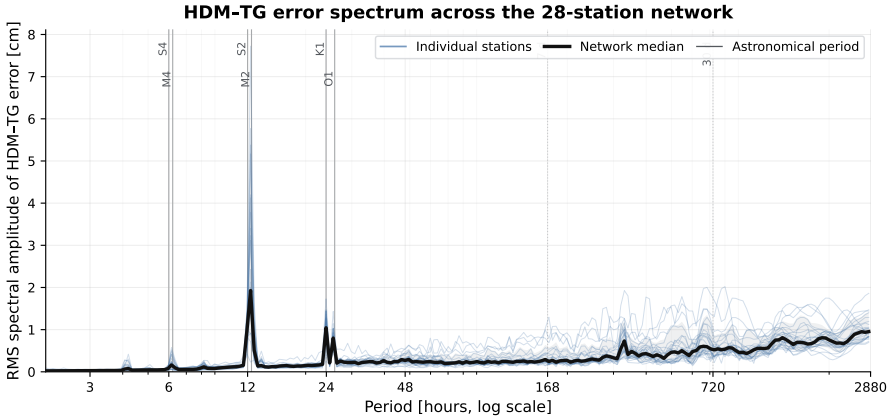


Figure 3.6: HDM–TG error spectra across the final 28-station network. Thin blue curves show individual stations and the thick black curve their pointwise median. Vertical guides mark the main semidiurnal, diurnal and shallow-water astronomical periods, while the 7- and 30-day guides provide non-tidal scale references.

The network-wide spectrum makes the narrow-band structure difficult to dismiss as a peculiarity of one gauge. The median spectrum reaches its largest short-period amplitude, approximately 1.93 cm RMS, near 12.32 hours, between the S2 and M2 guides and closest to M2. Secondary median peaks occur near 23.93 hours (1.05 cm) and 25.5 hours (0.80 cm), close to the K1 and O1 diurnal constituents.

A static station offset contributes at zero frequency and nowhere else, so these peaks belong to the dynamic residual and not to the mean that was removed. Their proximity to astronomical periods is consistent with residual phase or amplitude differences between modelled and observed tidal motion, perhaps compounded by local geometry and by grid-to-gauge collocation. The spectrum on its own cannot separate those candidates.

The narrow-band peaks are only one component of the target. A broad continuum extends over multi-day periods and is consistent with synoptic and storm-surge variability. Together with the autocorrelation analysis, this shows that the residual combines short-period periodic structure with longer, event-scale memory.

Taken together, the diagnostics point towards a temporal representation that works across several time scales at once. They do not fix the architecture or the receptive field numerically; those choices are specified and tested in Chapter 4.

3.5 KDI registered high-water events

The preceding diagnostics characterize the residual over the complete record. Storm-surge performance, however, also requires an external definition of which high-water episodes matter. For the Danish coast, this reference is provided by the Danish Coastal Authority (Kystdirektoratet, KDI), which publishes high-water statistics for the national tide-gauge network (Poulsen et al. 2024).

Each station sheet contains fitted tail parameters, a return-level curve, printed design levels and a dated register of measured extreme water levels ("Målte ekstremer"; Figure 3.7). Section 1.1.1 used those return-period classes to argue why a few centimetres matter. Here they do a different job: the register is an externally defined list of which high-water episodes occurred during 2013–2022, drawn up without reference to this thesis.

Event classes are reproduced from the published KDI station sheets rather than fitted to the ten-year thesis record. The reported Reference-90 event level is compared with return levels generated from the printed KDI parameters and with the published 20-, 50- and 100-year design levels. No new extreme-value distribution is estimated.

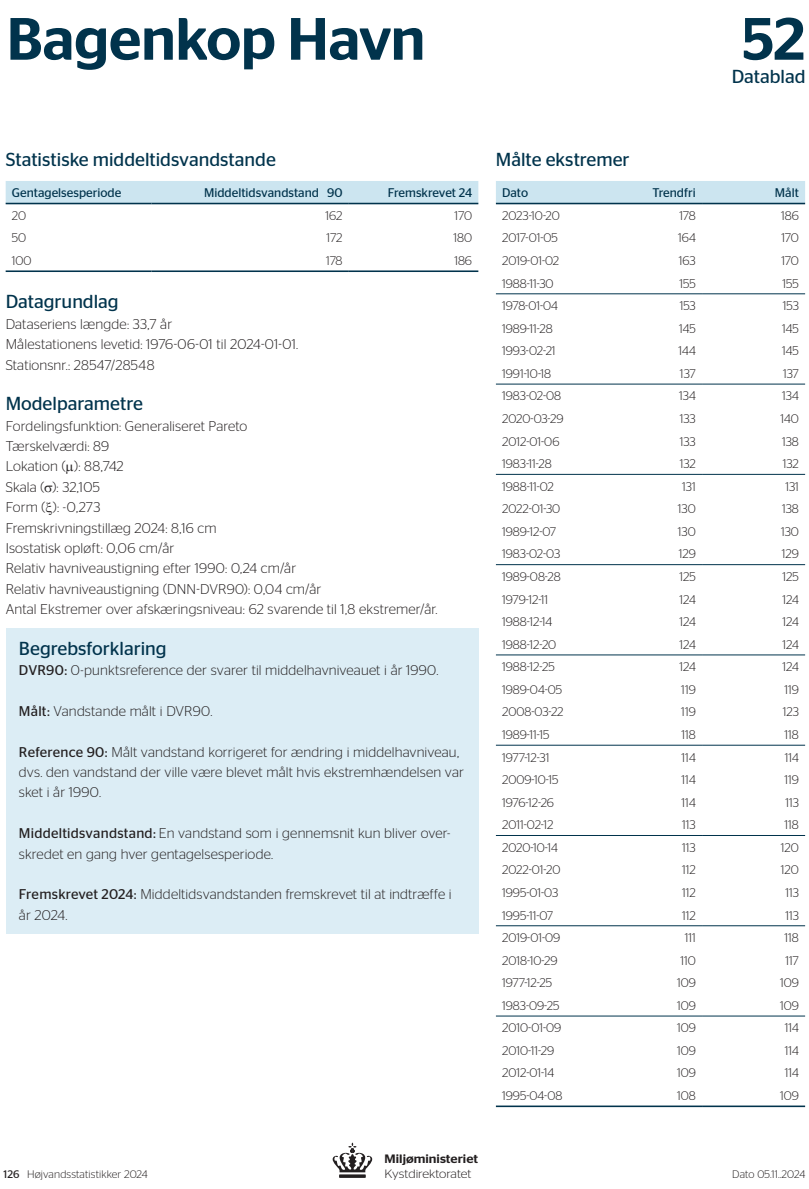
KDI derives its station statistics using a peaks-over-threshold procedure with station-specific declustering choices (Poulsen et al. 2024; Coles 2001). A return period represents the average recurrence interval of independent exceedances above a given level; it does not imply that an event of that magnitude can occur only once during the corresponding period. In this thesis, return-period terminology is used only to reproduce and interpret the KDI classifications.

The KDI report covers 24 of the 28 thesis stations. Hov, Vedbæk, Dragør and Bandholm have no corresponding station sheet and are therefore excluded only from the official-event evaluation. For the covered stations, the extraction preserves the reported date, measured level and Reference-90 level. Classes 1, 2, 5, 10, 20, 50 and 100 years are reconstructed from the published quantities. For 1–10 years, the KDI parameters are evaluated through the published return-level relationship; for 20 years and above, the values printed in the station tables are used.

Recomputing the printed 20-year level from the published parameters gives a transcription and consistency check. Agreement is within 1 cm at 23 of the 24 covered stations, and so also within the 1.5 cm tolerance used for the design-level statistics of Section 1.1.1. Kolding is the single exception: its printed parameters return a 20-year level 5 cm below the printed value, and the gap widens to 10 cm at 100 years. The printed parameters and design levels are internally inconsistent. Kolding is neither a test station nor part of any reported score, and it is excluded from the design-level statistics.

The register contains 192 station-events during 2013–2022 across the 24 covered stations, where the same named storm observed at two gauges counts twice. Of these, 145 reach at least the reproduced one-year class.

The subset that is eventually scored follows from the protocol rather than from inspection of the results, and it is worth stating here so that no selection step stays implicit. Of the 192 registered station-events, 46 fall at the seven test stations defined in Section 4.7; one of them, Bagenkop on 20 January 2022, does not reach the minimum number of valid hours required inside the peak-search window, which leaves 45 scoreable station-events. Thirty-nine of these carry a reproducible official class: ten at one year, thirteen at two, six at five, three at ten, six at twenty and one at one hundred. The remaining six are dated registered extremes that fall below the reproduced one-year class, kept because the register, and not a fitted threshold,



Bagenkop Havn

52
Datablad

Statistiske middeltidsvandstande

Gentagelsesperiode	Middeltidsvandstand	90	Fremskrevet 24
20		162	170
50		172	180
100		178	186

Datagrundlag

Dataseriens længde: 33,7 år
Målestationens levetid: 1976-06-01 til 2024-01-01.
Stationsnr.: 28547/28548

Modelparametre

Fordelingsfunktion: Generaliseret Pareto
Tærskelværdi: 89
Lokation (μ): 88,742
Skala (σ): 32,105
Form (ξ): -0,273
Fremskrivningstillæg 2024: 8,16 cm
Isostatisk opløft: 0,06 cm/år
Relativ havniveaustigning efter 1990: 0,24 cm/år
Relativ havniveaustigning (DNN-DVR90): 0,04 cm/år
Antal Ekstremer over afskæringsniveau: 62 svarende til 1,8 ekstremer/år.

Begrebsforklaring

DVR90: 0-punktsreference der svarer til middelhavniveauet i år 1990.

Målt: Vandstande målt i DVR90.

Reference 90: Målt vandstand korrigeret for ændring i middelhavniveau, dvs. den vandstand der ville være blevet målt hvis ekstremhændelsen var sket i år 1990.

Middeltidsvandstand: En vandstand som i gennemsnit kun bliver overskredet en gang hver gentagelsesperiode.

Fremskrevet 2024: Middeltidsvandstanden fremskrevet til at indtræffe i år 2024.

Målte ekstremer

Dato	Trendfri	Målt
2023-10-20	178	186
2017-01-05	164	170
2019-01-02	163	170
1988-11-30	155	155
1978-01-04	153	153
1989-11-28	145	145
1993-02-21	144	145
1991-10-18	137	137
1983-02-08	134	134
2020-03-29	133	140
2012-01-06	133	138
1983-11-28	132	132
1988-11-02	131	131
2022-01-30	130	138
1989-12-07	130	130
1983-02-03	129	129
1989-08-28	125	125
1979-12-11	124	124
1988-12-14	124	124
1988-12-20	124	124
1988-12-25	124	124
1989-04-05	119	119
2008-03-22	119	123
1989-11-15	118	118
1977-12-31	114	114
2009-10-15	114	119
1976-12-26	114	113
2011-02-12	113	118
2020-10-14	113	120
2022-01-20	112	120
1995-01-03	112	113
1995-11-07	112	113
2019-01-09	111	118
2018-10-29	110	117
1977-12-25	109	109
1983-09-25	109	109
2010-01-09	109	114
2010-11-29	109	114
2012-01-14	109	114
1995-04-08	108	109

Figure 3.7: Official KDI station sheet for Bagenkop (Poulsen et al. 2024), including the published peaks-over-threshold parameters, return-level curve, design levels and registered extreme events. The thesis reproduces the published event classes but does not fit a new extreme-value model.

defines the sample. The scoring procedure itself is specified in Section 4.8.

3.5.1 What the register shows about the events

Grouping registered events that occur within three days of one another identifies storm episodes affecting multiple parts of the network simultaneously. Table 3.1 shows that the three most widespread episodes (Malik, the January 2017 Baltic surge and Alfrida) were registered at 19–21 stations. Bodil ranks fourth with 11 stations.

Two of the four, Bodil and the January 2017 surge, carry a registered level in the highest reproduced class at some gauge.

Table 3.1: The twelve most widespread registered storm-surge episodes of 2013–2022, ranked by the number of network stations recording them. "Stations" counts distinct tide gauges registering an event within each episode; "Max class" is the highest published return-period class reached (Poulsen et al. 2024).

Episode	Named storm	Stations	Max class
30 Jan 2022	Malik	21	20-yr
05 Jan 2017	January 2017 Baltic surge	20	100-yr
02 Jan 2019	Alfrida	19	20-yr
06 Dec 2013	Bodil	11	100-yr
29 Mar 2020	–	10	2-yr
20 Jan 2022	–	9	2-yr
29 Oct 2017	Ingolf	8	20-yr
09 Jan 2019	–	8	5-yr
29 Oct 2018	–	8	1-yr
27 Dec 2016	–	7	20-yr
10 Feb 2020	Ciara	6	10-yr
14 Oct 2020	–	6	1-yr

Reaching many gauges is not the same as reaching them equally. An episode registered at twenty stations can arrive at very different levels at each, which is the spatial heterogeneity of the Danish waters showing through the event record. The main analysis keeps the network as one population for that reason, and tests geographical specialization separately in Chapter 5.

3.6 Domain-scale forcing modes

The wind-direction analysis shows that local forcing matters, but the atmosphere acting over the Danish waters is not purely local. Pressure and wind evolve as spatial patterns extending across the domain. A single grid cell at a station can describe the local state, but it cannot represent the full geometry of an approaching pressure system, a coherent wind pattern or the large-scale forcing acting across the basin.

A second representation of the atmospheric archive is therefore built from Empirical Orthogonal Functions (EOFs), equivalently Principal Component Analysis applied to the gridded fields. Neighbouring grid cells are highly correlated, so a small number of modes can describe much of the recurring spatial variability.

The four atmospheric fields are concatenated after each grid-point feature has been centred and scaled,

$$\widetilde{X}_{tj} = \frac{X_{tj} - \mu_j}{\sigma_j}, \tag{3.4}$$

so that pressure, temperature and wind enter the joint decomposition on comparable scales. Here j identifies one variable at one grid cell, while μ_j and σ_j are calculated over the complete common 2013–2022 record.

The resulting time-by-feature matrix contains 81,089 hourly fields and 68,252

standardized features. Denoting it by $\widetilde{\mathbf{X}}$, the singular-value decomposition is

$$\widetilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}, \tag{3.5}$$

where the columns of $\mathbf{V}$ define the spatial loading patterns and the columns of $\mathbf{U}\mathbf{\Sigma}$ their temporal scores. Retaining the first K modes gives

$$\widetilde{X}_{tj} \approx \sum_{k=1}^{K} a_k(t) v_{jk}, \tag{3.6}$$

where v_{jk} is loading j of the k-th EOF and $a_k(t)$ its temporal score.

In the physical units of feature j, the same truncated reconstruction is

$$X_{tj} \approx \mu_j + \sigma_j \sum_{k=1}^{K} a_k(t) v_{jk}. \tag{3.7}$$

The post-processing model does not reconstruct the atmospheric fields. It uses the score histories $a_k(t)$ directly as a compact representation of the evolving domain-scale state.

Figure 3.8 shows the first three modes of the joint four-variable DMI atmospheric decomposition. Sea-level-pressure loading is represented by colour, 2-m temperature loading by contours and 10-m wind loading by arrows. The three modes explain 29.6%, 23.3% and 18.0% of the standardized atmospheric variance, respectively.

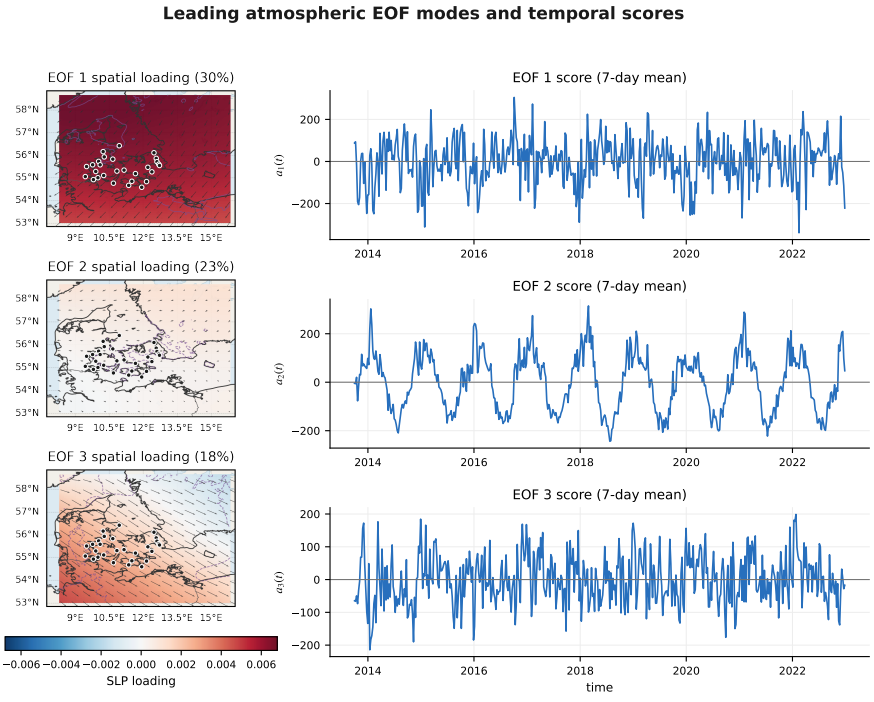


Figure 3.8: First three modes of the joint four-variable EOF decomposition of the complete 2013–2022 DMI HIRLAM/HARMONIE record. Sea-level-pressure loading is shown by colour, 2-m temperature loading by contours and 10-m wind loading by arrows. The corresponding temporal scores are shown as seven-day means for legibility.

The modes combine broad pressure structures with distinct wind orientations, while the second temporal score also shows a pronounced seasonal modulation associated with the temperature contribution. Their value is not that each mode corresponds to one named meteorological regime, but that a small number of scores captures coherent atmospheric variability that would otherwise require thousands of gridded predictors.

The domain-scale modes complement rather than replace the collocated atmospheric variables. The local predictors describe the forcing near a station, whereas the EOF scores describe the larger synoptic configuration in which that station is embedded. The distinction is relevant because coastal sea level can depend both on the local wind and pressure and on the larger-scale redistribution of water through the connected basins.

Two cautions belong with that reading. EOF modes form a mathematical basis, not a unique set of physical mechanisms: modes with neighbouring eigenvalues can mix, and the sign of any mode is arbitrary (North et al. 1982). And while the loading maps make the compression interpretable, they do not show that positive and negative residual events occupy cleanly separated atmospheric regimes. Only the narrow conclusion is carried forward: the EOF scores are a compact representation of the evolving domain-scale forcing. How many of them the model receives, and what it does with them, is settled in Chapter 4.

3.7 Related residual-correction studies

Residual post-processing preserves the hydrodynamic model as the physical backbone of the prediction while assigning a data-driven component only the remaining discrepancy. Several recent coastal sea-level studies have adopted related approaches, although their targets, station protocols and predictor availability differ from those considered here.

Tedesco et al. (2024) train station- and lead-specific neural-network corrections for Norwegian storm-surge forecasts and show that the achievable improvement decreases with forecast lead time. Their study establishes the value of residual correction in an operational forecasting context, but does not test transfer to stations outside the training population.

HIDRA2 predicts total sea level at Koper using atmospheric forecasts, astronomical tide and the previous 72 hours of observed sea level (Rus et al. 2023). Its use of recent observations is particularly relevant to event phase and seiche behaviour, but it differs from the present setting because the target-station gauge history is available during inference.

The northern Adriatic comparison of Campos-Caba et al. (2025) evaluates several statistical and neural architectures using gridded predictors compressed by principal components. It shows that architecture alone does not determine performance: the representation of extreme events in the loss and evaluation procedure can be equally important. Its protocol, however, does not evaluate transfer beyond the training

population.

The closest methodological precedent is the Baltic study of Jahanmard et al. (2023). Their WaveNet-inspired model uses dilated causal convolutions to learn dynamic hydrodynamic-model error separately from vertical-reference bias and evaluates transfer using distinct training, validation and test stations. It is directly relevant here because it combines dynamic residual learning with spatial transfer.

The present thesis narrows that broader Baltic setting to the Danish Straits and western Baltic storm-surge problem. It evaluates one shared residual model at stations whose observations do not enter model fitting or selection and places particular emphasis on independently registered high-water events.

Taken together, previous studies establish the value of neural and hybrid post-processing for coastal sea-level prediction (Tedesco et al. 2024; Giaremis et al. 2024; Rus et al. 2023; Campos-Caba et al. 2025; Jahanmard et al. 2023). The specific emphasis here is on the combination of spatial transfer, causal predictor histories and an evaluation protocol anchored to an external register of observed extremes.

3.8 From diagnosis to model design

The diagnostics in this chapter put empirical weight behind the design choices sketched in Chapter 1. Residual memory is not merely plausible on physical grounds: the observed autocorrelation stays significant over several days. The wind sectors show the discrepancy varying systematically with forcing direction, and the atmospheric EOFs show that the relevant forcing also has coherent domain-scale structure, which no single collocated grid cell can represent.

The frequency analysis adds a second temporal constraint. The residual contains both narrow short-period components and a broader multi-day continuum, so its structure spans more than one characteristic scale. At the same time, the KDI register shows that the largest high-water episodes affect multiple stations without producing a spatially uniform response.

What these results establish is which information a correction could plausibly exploit. They do not establish the architecture. Chapter 4 takes the observed temporal memory and the complementary local and domain-scale forcing representations as constraints on the input design, and from there specifies the receptive field, the network, the training objective and the validation procedure.

The move is from diagnosis to implementation. This chapter has shown that structured, physically interpretable information remains in $\varepsilon_i(t)$; the next asks how much of it can be learned without contemporaneous tide-gauge observations and without knowing which station is which.

Chapter 4

Methodology

This chapter sets out the framework used to correct the dynamic residual of DKSS and to test whether the correction transfers to tide gauges left out of model fitting. Stated plainly, the problem is supervised regression on ordered time series: a sample is a station i and an hourly prediction time t, and the model maps atmospheric forcing and recent hydrodynamic variability over $[t - R + 1, t]$ onto the dynamic DKSS residual at t. The network never replaces the hydrodynamic simulation and never predicts total sea level on its own. It estimates the time-dependent part of the model-minus-gauge residual once the station-dependent static bias has been taken out.

The final model is a Temporal Convolutional Network (TCN) whose weights are shared across stations. The domain-scale atmospheric representation and all station-local time series are projected into one hidden sequence before the temporal convolutional backbone. Its convolutions use left padding so that the estimate at t cannot depend on later inputs. The same learned weights are applied independently at every station, and no station identifier is supplied. Spatially shared information enters through the atmospheric EOF scores; local pressure, temperature, wind and recent DKSS variability describe conditions at each gauge.

The experimental design separates model development from final assessment. Training stations determine the network parameters, validation stations determine the checkpoint and architecture, and test stations are evaluated only after these choices have been frozen. The same 2013–2022 time axis is available in every station role, so the main experiment tests spatial rather than chronological transfer.

Four constraints carry the diagnosis of Chapter 3 into this experiment. The correction has to keep the hydrodynamic model as its physical backbone; it must use no contemporaneous gauge value at inference; every predictor history must end at the prediction time; and one learned mapping must serve stations whose residuals took no part in fitting or checkpoint selection. Those four requirements are what produce residual rather than total-level prediction, causal temporal windows, shared network weights and a station-disjoint evaluation.

4.1 Residual post-processing formulation

Chapter 2 defined the supervised target; it is restated here in the form the regression acts on. For station i and time t, with $\eta_i^{\mathrm{HDM}}(t)$ the collocated DKSS level and $\eta_i^{\mathrm{TG}}(t)$ the quality-controlled observation,

$$\varepsilon_i(t) = \left[\eta_i^{\mathrm{HDM}}(t) - \eta_i^{\mathrm{TG}}(t)\right] - \beta_i, \tag{4.1}$$

with β_i estimated as in Section 2.8. A positive value places the bias-corrected model above the observation, a negative value below it.

Given the predictor history $\mathbf{x}_i(t-R+1:t)$ ending at prediction time t, the network produces

$$\hat{\varepsilon}_i(t) = f_\theta(\mathbf{x}_i(t-R+1:t)), \tag{4.2}$$

where R is the temporal receptive field and θ holds the trainable parameters. Under a purely quadratic objective this mapping would target a conditional mean. The final model uses the asymmetric extreme-oriented objective of Section 4.5 instead, so at high water what it learns is an underestimation-averse conditional predictor and not an unrestricted least-squares mean.

The corrected sea level is reconstructed by subtracting the predicted dynamic residual from the bias-corrected DKSS level,

$$\hat{\eta}_i^{\mathrm{corr}}(t) = \eta_i^{\mathrm{HDM}}(t) - \beta_i - \hat{\varepsilon}_i(t). \tag{4.3}$$

The evaluation reference is therefore $\eta_i^{\mathrm{HDM}}(t) - \beta_i$ throughout, which keeps the reported skill attributable to the dynamic term alone. The same separation between reference-level bias and dynamic residual underlies previous Baltic residual-correction work (Jahanmard et al. 2023).

Equation (4.2) is the formal statement of the operating condition described in Section 1.3. No tide-gauge value at time t appears in the argument of f_θ, which means the event-scale variation that β_i cannot follow has to be estimated from forcing histories alone.

4.2 Time-ordered samples and predictor windows

Time-series observations are ordered realizations of a dynamic process, and that order has to be preserved when predictors and validation experiments are built (Madsen 2007). A sample is accordingly a pair (i, t) formed from a window ending at prediction time t, and every time-dependent input is restricted to

$$[t-R+1, \ldots, t], \tag{4.4}$$

so no value later than t can influence $\hat{\varepsilon}_i(t)$. The formulation is a zero-lead correction or nowcast because predictor values at time t are available; it is not a forecast at a positive lead time. The same ending-at-t convention is applied to every input

channel.

The input length is set equal to the exact structural receptive field,

$$\tau = R, \tag{4.5}$$

which avoids supplying an older part of the sequence that the output at time t cannot reach. The final model uses $R = 316$ h, or approximately 13.2 days of input history ending at t. This multi-day context is physically appropriate because Baltic and Danish sea level can depend on preceding wind and pressure conditions, basin preconditioning and the subsequent relaxation of accumulated water rather than only on instantaneous forcing (Leppäranta and Myrberg 2009).

For one prediction time t, the model receives a shared atmospheric tensor and a station-local tensor. Their unbatched dimensions are

$$\mathbf{A}_t \in \mathbb{R}^{R\times 10}, \qquad \mathbf{C}_{i,t} \in \mathbb{R}^{R\times 6}, \tag{4.6}$$

so a mini-batch of B station–time samples has dimensions $B \times R \times 10$ and $B \times R \times 6$, respectively. Windows are generated from the compact hourly arrays when requested rather than stored as duplicated tensors on disk.

A sample survives only if the target at time t is valid and the predictor histories meet the availability rules of Chapter 2. The target is never imputed. Predictors and labels are treated differently here on purpose: a sparse missing predictor entry is replaced after standardization by zero, which is the fitted channel mean, whereas a missing residual label removes that station–time sample from training and evaluation altogether.

4.3 Predictor representation

Every predictor enters the network as a time series, which makes the model purely temporal, though not purely local. Domain-scale atmospheric fields are compressed before training into EOF-score time series shared across the network, while collocated predictors describe the forcing and the recent model variability at each gauge. There is no spatial encoder anywhere in the design: the TCN convolves along time in one dimension and receives neither station identity nor any direct history of the target residual.

The complete input contains 16 channels divided into three physical groups:

1. ten domain-wide EOF scores obtained jointly from the DMI SLP, $T_{2\mathrm{m}}$, u_{10} and v_{10} atmospheric fields;

2. five station-local atmospheric histories: SLP, $T_{2\mathrm{m}}$, u_{10}, v_{10} and the backward one-hour SLP tendency;

3. one station-local measure of recent DKSS variability, denoted DKSSstdev24.

The first group forms the shared domain-scale tensor of ten channels. The remaining groups form the six station-local channels $(5 + 1)$; the two tensors are assembled separately below.

The domain-scale atmospheric history is

$$\mathbf{A}(t - R + 1 : t) = [a_1(t - R + 1 : t), \ldots, a_{10}(t - R + 1 : t)] \in \mathbb{R}^{R\times 10}, \tag{4.7}$$

where $a_k(t)$ is the score of atmospheric mode k. Before the decomposition, every grid-point variable is centred and scaled over the complete common 2013–2022 record as in Equation (3.4). The same score vector is then available to every station at time t because it represents the synoptic state of the full atmospheric crop. The ten scores come from one joint EOF of the stacked SLP, $T_{2\mathrm{m}}$, u_{10} and v_{10} fields, preserving their cross-variable covariance. They retain 93.2% of the standardized atmospheric variance while reducing the $113 \times 151 \times 4$ gridded state to ten values per hour. The network uses these domain-wide histories together with, rather than in place of, the local atmospheric sequence defined below.

The local atmospheric sequence at station i is

$$\mathbf{L}_i(t - R + 1 : t) = \left[p'_{\mathrm{SL},i}, T_{2\mathrm{m},i}, u_{10,i}, v_{10,i}, \Delta p'_{\mathrm{SL},i}\right]_{t-R+1:t} \in \mathbb{R}^{R\times 5}, \tag{4.8}$$

where the pressure tendency is computed from the present and the immediately preceding SLP value. It supplies a local pressure-change signal without reaching for a later pressure field.

The only predictor derived from DKSS sea level is DKSSstdev24. The absolute collocated DKSS level and its temporal history are deliberately excluded. Let

$$s_i(t) = \mathrm{sd}\Big(\eta_i^{\mathrm{HDM}}(t - 23 : t)\Big) \tag{4.9}$$

be the trailing 24-hour sample standard deviation. A station-dependent modifier is defined as

$$\chi_i = \frac{\overline{\sigma^2_{\mathrm{day},i}}}{\max\limits_{j\in\mathcal{S}_{\mathrm{train}}} \overline{\sigma^2_{\mathrm{day},j}}}, \tag{4.10}$$

where $\overline{\sigma^2_{\mathrm{day},i}}$ denotes the mean squared daily DKSS variability at station i over the frozen reference-statistics interval ending before 1 January 2021. This interval is used for the descriptor normalization and loss threshold only; it is distinct from the complete-record calibration of β_i and from the complete-record atmospheric EOF basis. The retained feature is

$$m_i(t) = \chi_i s_i(t). \tag{4.11}$$

The normalization denominator is determined from training stations and then frozen. This channel communicates whether the local hydrodynamic model has recently been dynamically active without exposing its absolute sea-level state.

What the final predictor set leaves out is as deliberate as what it keeps: the

absolute point-wise DKSS level, HDM EOF scores, the observed residual history, station identity, coordinates, bathymetry and fetch are all absent. Only the trailing variability descriptor DKSSstdev24 survives from the hydrodynamic side. Those exclusions keep the mapping free of an explicit station label and stop it from copying the level whose residual defines the target, while still letting recent local hydrodynamic activity inform the correction.

4.3.1 Standardization and calibration scope

Continuous predictors are standardized channel by channel,

$$\widetilde{x}(t) = \frac{x(t) - \mu_x}{\sigma_x + \delta}, \tag{4.12}$$

where μ_x and σ_x are frozen scaling constants and δ prevents division by zero. Standardization retains the sign of pressure and wind anomalies and makes a filled value of zero correspond to the fitted channel mean. The neural-network parameters are fitted only from training-station residuals, and validation and test residuals are reserved for their declared model-selection and final-evaluation roles.

The station-specific reference-level bias is calibrated from every quality-controlled residual available at each site over the complete common 2013–2022 record. Every test station consequently arrives with its vertical reference already fixed, which bounds what the spatial experiment can claim: transfer inside a calibrated network, and nothing about a coast that has never been gauged.

4.4 Shared-weight TCN architecture

One TCN, with causal left-padded convolutions and weights shared across all stations, is applied to each station's predictor sequence independently (Figure 4.1). Domain-scale and station-local channels are first mapped to a common hidden width H and then added. Fusing them this early lets atmospheric circulation and local forcing interact from the first temporal block, and it avoids replicating the same atmospheric EOF tensor as a separate learned branch at every station.

At each time step, the atmospheric scores are projected as

$$\mathbf{g}(t) = \mathbf{W}_g \mathbf{A}(t) + \mathbf{b}_g, \qquad \mathbf{g}(t) \in \mathbb{R}^H, \tag{4.13}$$

whereas the local context is

$$\mathbf{c}_i(t) = [\mathbf{L}_i(t), m_i(t)] \in \mathbb{R}^6 \tag{4.14}$$

and is projected through

$$\mathbf{l}_i(t) = \mathbf{W}_l \mathbf{c}_i(t), \qquad \mathbf{l}_i(t) \in \mathbb{R}^H. \tag{4.15}$$

Both projections act pointwise in time: the same weight matrix is applied indepen-

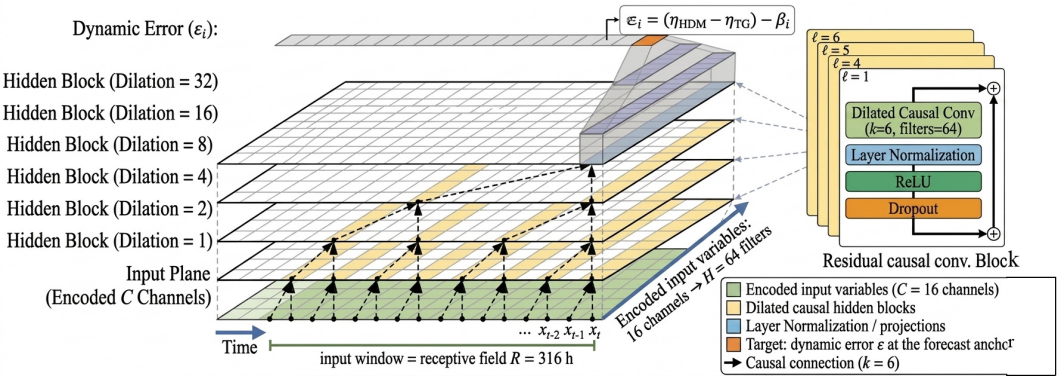


Figure 4.1: Architecture of the final shared-weight TCN. Ten domain-wide atmospheric EOF-score histories and six station-local histories are projected to hidden width $H = 64$, fused by addition and processed by six residual temporal blocks with dilations $1, 2, 4, 8, 16, 32$. A scalar head estimates the dynamic residual $\widehat{\varepsilon}_i(t)$ at prediction time t. The input history and exact receptive field are both 316 h.

dently at every time step and mixes only the channel dimension, leaving the temporal axis of length R intact. No flattening of the receptive field takes place here, so all temporal structure is handled by the causal convolutions of the shared backbone (Equation (4.17)) and the model remains a temporal convolutional network. Only the global projection contains an explicit bias term. The resulting hidden sequence is

$$\mathbf{h}_i^{(0)}(t) = \mathbf{g}(t) + \mathbf{l}_i(t), \tag{4.16}$$

and the shared backbone computes

$$\mathbf{h}_i^{(L)}(t) = \mathrm{TCN}_\theta\Big(\mathbf{h}_i^{(0)}(t - R + 1 : t)\Big). \tag{4.17}$$

The station axis is not a communication axis: changing a local predictor affects only that station sequence, while changing the shared atmospheric scores can affect all stations because they represent the common synoptic state.

4.4.1 Causal dilated convolutions and receptive field

For a one-dimensional sequence z_t, a causal convolution with kernel size K is

$$y_t = \sum_{k=0}^{K-1} w_k z_{t-k}. \tag{4.18}$$

Left padding preserves sequence length while keeping the output at t dependent on present and past inputs alone. Throughout this thesis *causal* refers to the temporal construction of each training sample and to left padding in the temporal convolutions. It does not mean causal inference, and it is not a label for the post-processing system as a whole. Dilation then widens the temporal support without a proportionally deeper network,

$$y_t = \sum_{k=0}^{K-1} w_k z_{t-dk}, \tag{4.19}$$

where d is the dilation factor. The blocks use the exponentially increasing sequence

$$d_\ell = 2^\ell, \qquad \ell = 0, \ldots, L-1, \tag{4.20}$$

which combines dense short-lag sampling with access to progressively older states, following the causal dilated-convolution principle introduced by WaveNet-type architectures (Oord et al. 2016).

With one causal convolution per residual block, the exact receptive field is

$$R = 1 + (K-1)\sum_{\ell=0}^{L-1} 2^\ell = 1 + (K-1)(2^L - 1). \tag{4.21}$$

For the final configuration, $K = 6$ and $L = 6$, hence

$$R = 1 + 5(1 + 2 + 4 + 8 + 16 + 32) = 316 \text{ h}. \tag{4.22}$$

Since the supplied input length equals the receptive field, every timestamp in the input can influence the estimate at time t. The 316-hour value describes the receptive field over derived hourly features, not the longest dependency on a raw source value. At the oldest feature time, the backward pressure difference reaches one additional hour and the trailing DKSSstdev24 reaches approximately $t - 338$.

4.4.2 Residual block

Each temporal block learns an increment rather than a replacement,

$$\mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{(\ell)} + F_\ell\Big(\mathbf{h}_i^{(\ell)}\Big), \tag{4.23}$$

where F_ℓ contains one dilated causal convolution, Layer Normalization across channels, a ReLU activation and dropout. Normalization acts independently at each time step and never aggregates along the temporal axis, so it cannot leak future information into an earlier representation, and it avoids any dependence on mini-batch statistics. The identity skip gives gradients a direct path through the stack and lets each block refine what the preceding blocks produced.

4.4.3 Prediction head and output

Only the hidden state at prediction time t is passed to the output head. The scalar dynamic-residual estimate is

$$\hat{\varepsilon}_i(t) = \mathbf{w}_2^\top \,\mathrm{SiLU}\Big(\mathbf{W}_1\mathbf{h}_i^{(L)}(t) + \mathbf{b}_1\Big) + b_2. \tag{4.24}$$

The final architecture has no auxiliary multi-hour output. Event-window scores are assembled during evaluation from consecutive hourly predictions around each registered event, not from a sequence-to-sequence forecast head.

4.5 Extreme-oriented training objective

A squared objective serves average accuracy well, but it is blind to the asymmetry set out in Section 1.1.1: at high water a level predicted too low and a level predicted too high do not carry the same consequence, and a squared loss treats them identically. Most hourly targets come from ordinary conditions, so an unmodified quadratic objective spends almost all of its attention there. The final model is trained instead with a point-output adaptation of the extreme-oriented loss (EoL) of Han et al. (2026). At ordinary water levels the contribution stays quadratic; once the observed level crosses a station-specific high-water threshold the loss changes shape and charges more for the direction that would leave the post-processed level too low. RMSE is not a second training objective. It is kept only as an evaluation metric in physical units.

The direction of this asymmetry is easiest to understand at the level of the supervised target. Define the target-prediction error as

$$\delta_i(t) = \hat{\varepsilon}_i(t) - \varepsilon_i(t), \tag{4.25}$$

where $\varepsilon_i(t)$ is the true dynamic residual and $\hat{\varepsilon}_i(t)$ is the residual predicted by the TCN. A positive δ_i therefore means that the network has *overestimated the residual target*; it does not mean that sea level has been overestimated. Because the predicted residual is subtracted from DKSS in Equation (4.3), the sign reverses. The signed post-processed sea-level error is

$$e_i^{\eta}(t) = \hat{\eta}_i^{\mathrm{corr}}(t) - \eta_i^{\mathrm{TG}}(t) = \varepsilon_i(t) - \hat{\varepsilon}_i(t) = -\delta_i(t). \tag{4.26}$$

Thus, overestimating the target ($\delta_i > 0$) produces a post-processed sea level that is too low ($e_i^{\eta} < 0$). Conversely, underestimating the target produces a sea level that is too high. Table 4.1 records this sign change explicitly.

Table 4.1: Relationship between error in the predicted residual target and error in the resulting post-processed sea level. "Overestimation" in the first column refers to the target ε, not to sea level.

Residual-target prediction	$\delta = \hat{\varepsilon} - \varepsilon$	Post-processed level	EoL branch
Target overestimated	$\delta > 0$	too low: $e^{\eta} < 0$	strong penalty
Target underestimated	$\delta < 0$	too high: $e^{\eta} > 0$	weaker penalty

For the loss, the sea-level error is nondimensionalized with a fixed scale $s_0 = 0.10$ m, written with a subscript to keep it clear of the trailing standard deviation $s_i(t)$ of Equation (4.9),

$$\widetilde{z}_i(t) = \frac{e_i^{\eta}(t)}{s_0} = -\frac{\delta_i(t)}{s_0}. \tag{4.27}$$

Let $u_{i,0.99}^{\mathrm{hist}}$ be the station-specific q99 observed sea-level threshold computed from quality-controlled hours in the same frozen reference-statistics interval ending before

1 January 2021, and define

$$\alpha_i(t) = \begin{cases} 1, & \eta_i^{\mathrm{TG}}(t) < u_{i,0.99}^{\mathrm{hist}}, \\ 0, & \eta_i^{\mathrm{TG}}(t) \geq u_{i,0.99}^{\mathrm{hist}}. \end{cases} \tag{4.28}$$

The pointwise contribution is

$$\ell_{\mathrm{EoL},i}(t) = [1 - \alpha_i(t)]\, \Phi(\tilde{z}_i(t)) + \alpha_i(t)\tilde{z}_i(t)^2, \tag{4.29}$$

with

$$\Phi(\tilde{z}) = \begin{cases} \exp(-\tilde{z}) - 1, & \tilde{z} < 0, \\ \exp(\tilde{z}/\lambda) - 1, & \tilde{z} \geq 0, \end{cases} \qquad \lambda = 25. \tag{4.30}$$

Because the negative branch grows as $\exp(-\tilde{z})$ whereas the positive branch grows as $\exp(\tilde{z}/25)$, the strong branch is activated when the residual target has been overestimated and the resulting sea level is therefore too low. This direction is operationally important near high water because it can suppress a peak in the post-processed series.

Figure 4.2 shows why this counts as an extreme-oriented loss and not a sample-weighting rule: below the station-specific switch the contribution is quadratic, above it the functional form itself changes and underestimation grows exponentially faster than overestimation. The q99 switch is a training device and nothing else. The observed level and its indicator enter the evaluation of the loss but are never supplied as predictors at inference, and the switch plays no part in defining the headline event sample, which comes independently from the KDI registered-event archive.

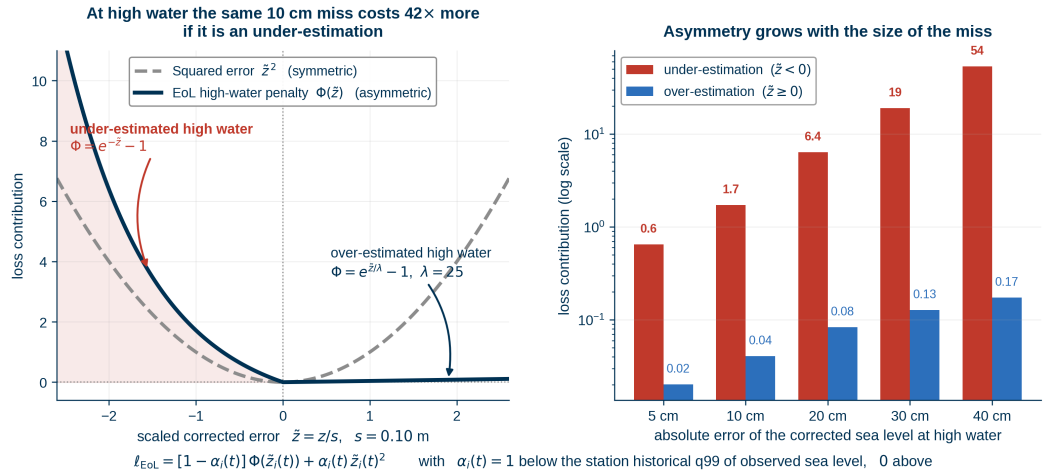


Figure 4.2: Extreme-oriented loss at high water, expressed on the post-processed sea-level error $\tilde{z} = e^{\eta}/s_0 = -\delta/s_0$. The negative branch corresponds to *overestimating the residual target*, which makes the post-processed sea level too low. With $s_0 = 0.10$ m and $\lambda = 25$, a 10-cm sea-level underestimation contributes approximately 42 times more loss than a sea-level overestimation of the same magnitude.

Training samples follow the full available distribution, with no event-only re-sampling anywhere in the final suite. Extremes are emphasized through the shape

of the loss, not through how often they are drawn, which keeps the optimization population aligned with the complete sea-level record while still raising the cost of underpredicting high water.

4.6 Optimization and checkpoint selection

The optimizer is AdamW with learning rate 3×10^{-4} and zero weight decay. The mini-batch size is 32. Gradients are clipped to a maximum norm of 1 to reduce the influence of exceptional batches, and mixed-precision arithmetic is used where supported. All runs in the final battery use random seed 42.

Training continues for at least 100 epochs before early stopping may terminate a run. At every evaluated epoch, the training and validation EoL components are logged and the retained checkpoint is the one with minimum validation EoL, which for the final model occurs at epoch 83. Validation uses the full distribution with a deterministic temporal stride of two; the test evaluation uses the full requested set of valid station–time samples without that stride. This distinction affects computational density, not the station roles, and is reported explicitly because the validation history does not contain every consecutive hour.

Checkpoint and architecture selection use validation information only, and the test set is evaluated once the selected checkpoint has been restored. One seed and one fixed station partition mean the final experiment quantifies neither optimization uncertainty nor the uncertainty attached to alternative splits.

4.7 Spatial experimental protocols

The main experiment evaluates the research question through a spatial station split. Training, validation and test stations are mutually disjoint, but all roles share the 2013–2022 temporal domain and the same domain-scale atmospheric forcing. Shared atmospheric covariates do not contain tide-gauge labels; the separation concerns the station residuals used to fit and select the model.

4.7.1 Primary station-split evaluation

The global experiment trains one network-wide model on 15 stations, selects its checkpoint on 6 validation stations and evaluates it on 7 test stations: Aarhus, Grenaa, Bagenkop, Faaborg, Køge, Drogden-Fyr and Korsør. None of the seven appears in a gradient update or in checkpoint selection. What the experiment tests, then, is whether a shared residual relationship learned across the wider Danish and western Baltic network carries to new station locations under the complete-record reference-level calibration defined above. Figure 4.3 shows the full spatial partition.

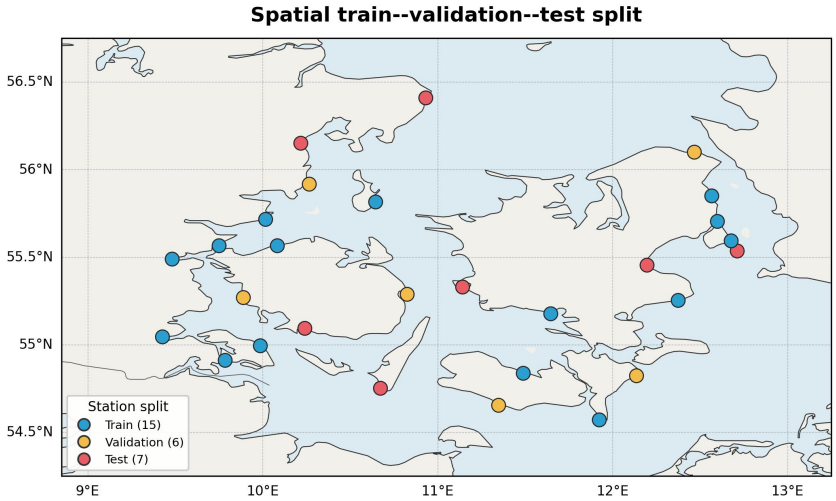


Figure 4.3: Spatial train–validation–test partition of the 28 tide gauges: 15 training stations, 6 validation stations and 7 test stations. Marker colour denotes only the station role; uniform circular markers carry no geographical subclassification.

4.7.2 Independent input-group ablation

The ablation experiment uses the global station split and retrains the complete architecture after removing exactly one of seven predeclared input groups: all atmospheric EOF scores, DKSSstdev24, local SLP, local SLP tendency, local u_{10}, local v_{10} or local $T_{2\mathrm{m}}$. The removals are independent and do not form a sequential backward-selection procedure. Their primary statistic is

$$\Delta\mathcal{L}_{\mathrm{val},g} = \mathcal{L}_{\mathrm{val}}^{(-g)} - \mathcal{L}_{\mathrm{val}}^{(\mathrm{full})}, \tag{4.31}$$

where $\mathcal{L}_{\mathrm{val}}^{(-g)}$ is the validation loss (the extreme-oriented loss, EoL) of the model retrained with input group g removed, and $\mathcal{L}_{\mathrm{val}}^{(\mathrm{full})}$ is the validation loss of the full model that keeps all groups. A positive $\Delta\mathcal{L}_{\mathrm{val},g}$ means validation EoL rises when group g is removed, meaning that performance worsens and that the group was carrying useful information. Test RMSE is kept as secondary descriptive evidence and never used to revise the input set after the fact.

4.8 Evaluation metrics

All physical scores compare the corrected level with the tide-gauge observation and use the bias-corrected DKSS series as baseline. For N valid station–time observations, the principal complete-record metric is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\hat{\eta}_n^{\mathrm{corr}} - \eta_n^{\mathrm{TG}}\right)^2}. \tag{4.32}$$

Mean absolute error, mean signed error and correlation are reported where they add diagnostic information. Percentage improvement over the paired baseline is

$$I = 100\frac{\mathrm{RMSE}_{\mathrm{base}} - \mathrm{RMSE}_{\mathrm{model}}}{\mathrm{RMSE}_{\mathrm{base}}}\%, \tag{4.33}$$

so $I > 0$ denotes improvement and $I < 0$ degradation.

To see whether skill changes towards the upper tail without defining events from the same record, conditional RMSE is also computed above selected observed-level quantiles. Let $Q_{i,q}$ be the observed-level q-quantile at station i, taken over the same frozen reference-statistics interval ending before 1 January 2021 that defines the training switch. The conditioned sample collects the hours at which each station exceeds its own threshold,

$$\mathcal{T}_q = \left\{ n : \eta_n^{\mathrm{TG}} \geq Q_{i(n),q} \right\}, \tag{4.34}$$

where $i(n)$ is the station of observation n, and

$$\mathrm{RMSE}_q = \sqrt{\frac{1}{|\mathcal{T}_q|} \sum_{n \in \mathcal{T}_q} (\hat{\eta}_n^{\mathrm{corr}} - \eta_n^{\mathrm{TG}})^2}. \tag{4.35}$$

These quantile-conditioned scores are descriptive evaluation slices. Because the thresholds are fixed before evaluation rather than recomputed from the test sample, a single network-wide quantile would place a different set of hours in the same slice. They share the reference interval with the training q99 switch of Equation (4.28) but not its role, and both are separate from the independent KDI event register.

The headline high-water evaluation uses only the dated events published by the Danish Coastal Authority and introduced in Chapter 3. No extreme-value distribution or event threshold is fitted from the 2013–2022 thesis series. For an event e with observed peak time t_e^*, the operational window is

$$\mathcal{W}_e = [t_e^* - 60\ \mathrm{h}, t_e^* + 60\ \mathrm{h}]. \tag{4.36}$$

The absolute error at the observed maximum is

$$E_{\mathrm{peak},e} = \left| \hat{\eta}^{\mathrm{corr}}(t_e^*) - \eta^{\mathrm{TG}}(t_e^*) \right|, \tag{4.37}$$

and the event-window RMSE is

$$E_{\mathrm{win},e} = \sqrt{\frac{1}{|\mathcal{W}_e|} \sum_{t \in \mathcal{W}_e} (\hat{\eta}^{\mathrm{corr}}(t) - \eta^{\mathrm{TG}}(t))^2}. \tag{4.38}$$

Errors are aggregated in physical units before percentage improvement is calculated, avoiding unstable event-wise percentages when the baseline error is close to zero. Equation (4.37) samples both series at the observed peak time t_e^* and therefore measures co-timed amplitude error, not peak timing. A shifted or overshooting

corrected maximum can consequently look farther from the observed crest than the reported value at t_e^*. The ±60-hour RMSE complements this measure by evaluating the complete build-up, maximum and relaxation trajectory; a dedicated timing metric remains outside the present assessment.

The protocol therefore yields three levels of evidence: complete-record accuracy, conditional upper-tail accuracy, and performance on independently registered events. Between them they separate general residual correction from the harder question the thesis actually asks: whether the model improves the amplitude of a high-water event at a station it was never fitted on.

Chapter 5

Results and discussion

This chapter answers the research question with the post-processing model and one declared station split. The model was fitted at 15 training stations, selected at six validation stations and evaluated at seven test stations. All reported test results come from seed 42. The regional models and input ablations are secondary analyses of the fixed final configuration; they do not replace the selected model after seeing the test set.

Two baselines run through this chapter, and they answer different questions. Raw DKSS is where an operational service starts, so a raw-to-final comparison measures the whole calibration-and-correction chain. Bias-corrected DKSS is what the network was actually asked to improve: β_i is removed before training and is never something the model learns. Every skill percentage reported below is measured against the bias-corrected series. A few figures show raw DKSS instead, to display the complete operational change, and say so in their captions.

The chapter is organised around three main findings: the dynamic residual transfers across stations; registered high-water events reveal both large gains and bounded failure cases; and the same correction remains available when the local gauge record does not. Two supporting analyses sit between the second and the third: a regional training experiment, reported as a robustness check rather than as part of the research question, and the input ablation and frequency diagnostics that document what the model uses and what it removes.

5.1 Finding 1: The dynamic residual is learnable and transferable

5.1.1 What the baseline already accounts for

Before reference-level correction, pooled DKSS RMSE at the seven test stations is 10.90 cm. Removing the station-specific constant lowers it to 8.39 cm; the network lowers it again to 5.37 cm. Put on a common denominator, the two steps contribute 23.1 and 27.6 percentage points of the 50.7% total.

Splitting it that way is not bookkeeping. For a signed error E with mean μ

and standard deviation σ, $\mathrm{RMSE}^2 = \mu^2 + \sigma^2$, so removing β_i moves the mean term and leaves everything else untouched: not the variance, not the correlation, not the timing or the shape of an event. The network works on what is left, which is why every percentage quoted for the network takes 8.39 cm as its denominator.

Køge and Drogden-Fyr show how differently the two steps can divide the same problem. Køge carries a large reference offset ($\beta_i = 9.90$ cm), and removing it alone accounts for 46.2% of the raw error; the 6.36 cm that survives is already close to what the network reaches anywhere, and it improves by a further 4.7%, worth 2.5 percentage points of the raw total. At Drogden-Fyr the constant accounts for only 27.4%, but the 11.83 cm left behind is the worst baseline in the test set and the network removes 54.1% of it, worth 39.3 points. The two stations finish within 0.7 cm of each other. What differs is the structure of the error, not the amount of skill available.

5.1.2 Pooled and station-level transfer

Across 555,343 valid test-station hours, the selected model reduces pooled RMSE from 8.389 to 5.374 cm, or 35.9% (Table 5.1). MAE falls from 6.43 to 4.12 cm and Nash–Sutcliffe efficiency rises from 0.843 to 0.936. Because none of the seven test-station residual records entered fitting or model selection, this gain is evidence that a shared part of the dynamic DKSS error transfers beyond the training population.

Table 5.1: Headline results for the final post-processing model. Error values are in centimetres and the baseline is bias-corrected DKSS. Full-record, quantile and event-hour rows pool the eligible hours; the macro row averages the 45 event-wise RMSE values, and peak rows average absolute errors. KDI events come from the official register, which is separate from the q99 training switch.

Evaluation register	Baseline	Post-processed	Reduction	Support
Complete-record RMSE	8.389	5.374	35.9%	555,343 h
RMSE above q95	10.786	6.939	35.7%	28,860 h
RMSE above q99	12.759	8.860	30.6%	5,899 h
RMSE above q99.9	14.002	11.178	20.2%	539 h
KDI event-hour RMSE (±60 h; pooled)	11.038	7.508	32.0%	5,306 h
KDI event RMSE (±60 h; macro mean)	10.068	7.135	29.1%	45 events
KDI observed-peak MAE	12.074	8.254	31.6%	45 peaks
KDI ≥ 20-year peak MAE	15.382	5.527	64.1%	7 peaks

All seven stations improve over the complete record, but the size of the gain varies: Drogden-Fyr 54.1%, Faaborg 42.9%, Aarhus 39.6%, Korsør 35.7%, Bagenkop 29.7%, Grenaa 22.7% and Køge 4.7%. The more revealing result is what happens to the spread. The sample standard deviation of station RMSE falls from 1.76 to 0.63 cm and its range from 5.47 to 1.98 cm. After correction, every station lies between 4.08 and 6.06 cm.

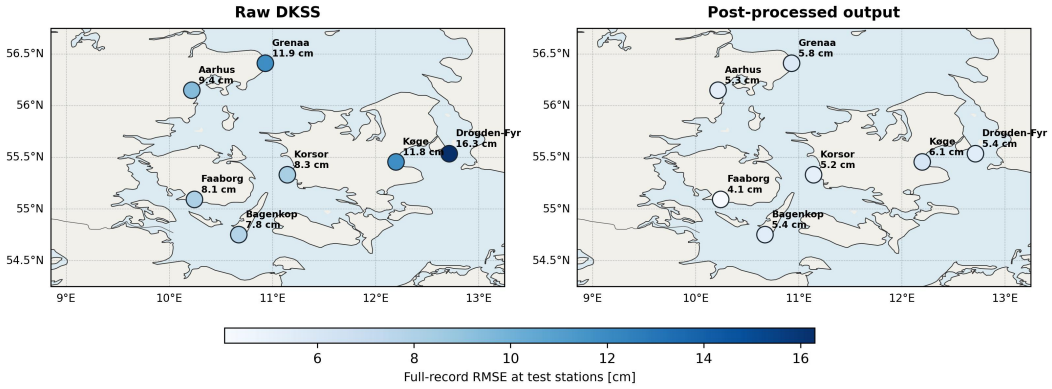


Figure 5.1: Full-record RMSE for raw DKSS and the final post-processed output at the seven test stations. Both panels share one colour scale and geographic rendering; labels report RMSE in centimetres. This is the complete endpoint comparison, whereas the 35.9% headline skill isolates the learned gain over bias-corrected DKSS. Test-station residuals entered neither fitting nor checkpoint selection.

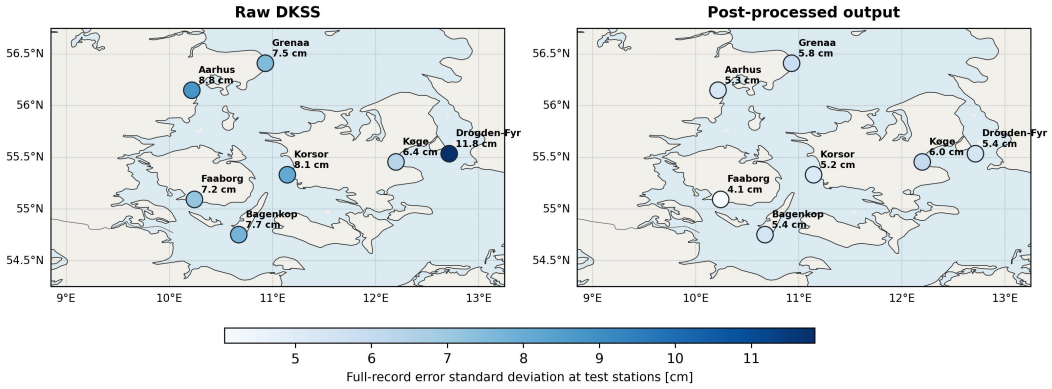


Figure 5.2: Sample standard deviation of the signed full-record error for raw DKSS and the final post-processed output. A station-wise constant offset does not change this quantity, so the left panel also represents the spread remaining after removal of β_i. The shared scale shows that post-processing reduces time-varying error at every test station.

The two maps make the same point at network scale. Pooled RMSE falls from 10.903 to 5.374 cm and pooled error standard deviation from 9.264 to 5.364 cm, but the more useful number is the range: station RMSE spans 7.81–16.29 cm before correction and 4.08–6.06 cm after it. An average improvement of a third conceals a convergence. Drogden-Fyr gains most because it starts furthest above the band the others end in; Køge gains least in the incremental comparison because it is already close to that band once β_i has gone.

5.2 Finding 2: High-water skill depends on the evaluation sample

The apparent behaviour at high water depends on how the evaluated hours are selected. Within the full hourly record, the gain decreases from 35.7% above q95 to 30.6% above q99 and 20.2% above q99.9. On the independent register of 45 KDI station-events, pooled RMSE over the observed peak ±60 hours falls from 11.04 to 7.51 cm (32.0%); the event-wise macro mean falls from 10.07 to 7.13 cm. Mean absolute error at the observed peak falls from 12.07 to 8.25 cm (31.6%).

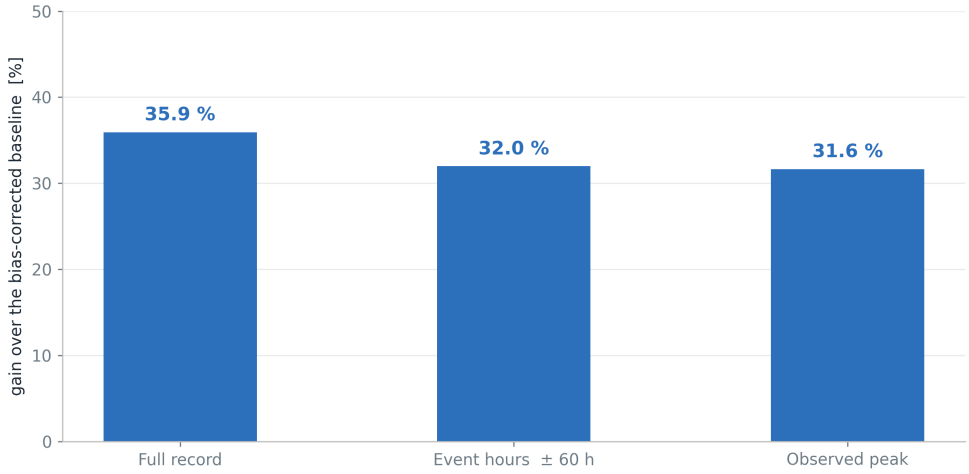


Figure 5.3: Performance of the selected model across the complete record, KDI event windows and observed event peaks. The event register is external to the q99 switch used during training.

The two evaluation samples do not represent the same population. Upper-tail exceedances select the hours at which the total observed level is highest, including hours when the remaining high-frequency residual is large. The KDI register instead follows named and dated high-water episodes through their build-up, crest and relaxation. The predictor set is well matched to the synoptic evolution of those episodes, so event skill need not follow the monotonic decline of the conditional quantiles.

The event improvement is not universal. Peak error improves in 31 of 45 events and window RMSE improves in 37 of 45, but only 27 improve on both measures; 41 improve on at least one. Mean peak error worsens in the five-year class, from

8.04 to 12.76 cm, and in the ten-year class, from 7.64 to 9.65 cm. In contrast, the 20-year class improves from 13.69 to 5.71 cm and the single 100-year case from 25.55 to 4.44 cm. The weak classes belong beside the strong ones, because together they say where the model earns its place: wherever the hydrodynamic baseline is badly wrong, and not simply wherever the water is high.

Figure 5.4 takes one of those events apart without summary overlays. During the January 2017 Baltic surge at Drogden-Fyr, raw DKSS misses the observed crest by 14.4 cm and carries a 14.8 cm window RMSE; the post-processed series brings those to 4.4 and 8.5 cm. The correction is not a vertical shift, though. It overshoots through part of the relaxation, and the lower panel shows where.

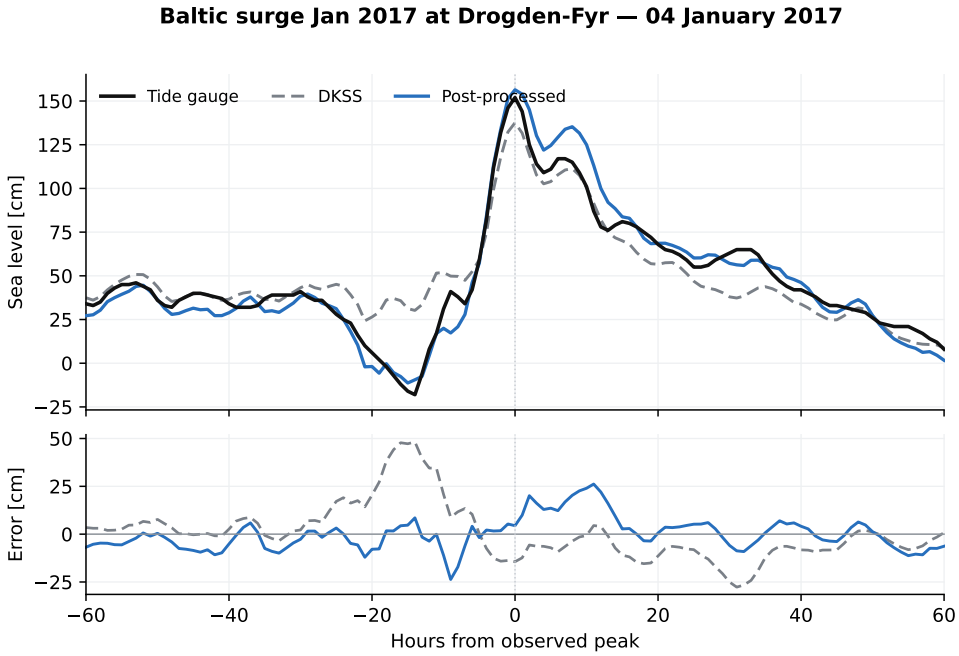


Figure 5.4: The first case in the frozen representative set: the January 2017 Baltic surge at Drogden-Fyr. The upper panel compares the tide gauge, raw DKSS and final post-processed output; the lower panel shows the corresponding signed raw and post-processed errors. Time zero is the observed peak. Missing gauge hours, where present, remain missing.

Figure 5.5 places that example beside five deliberately contrasting event–station cases. Alfrida at Bagenkop shows another strong recovery of an underestimated peak; Malik at Grenaa shows a moderate improvement; and Gorm at Grenaa shows removal of a larger raw-DKSS overestimation. Ingolf at Korsør is retained as a clear failure, while Ciara at Aarhus combines lower window RMSE with larger peak error. The selection therefore exposes both what the model learns and where a deterministic correction can over-correct.

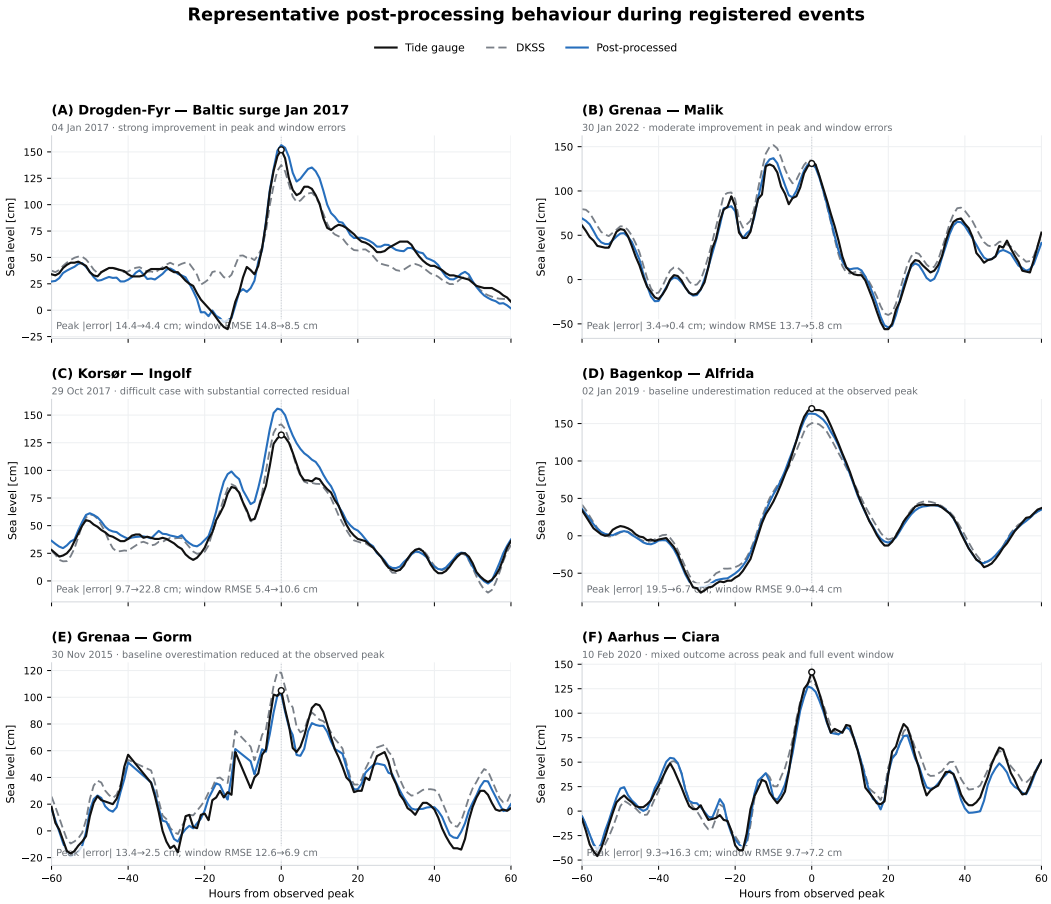


Figure 5.5: Six contrasting station–event examples illustrating the range of correction outcomes. Black is the tide gauge, dashed grey is raw DKSS and blue is the post-processed series. Each panel reports raw-to-final absolute peak error and ±60-hour RMSE. The set includes strong and moderate gains, corrected under- and overestimation, one difficult failure and one mixed outcome; it is descriptive rather than an estimate of event frequency. Missing gauge hours remain missing and are never interpolated.

5.2.1 Peak errors contract towards a common residual scale

The relationship becomes clearer when corrected peak error is plotted against baseline peak error. Among the 45 registered station-events, 32 have a baseline error of at least 8 cm. The correction improves 27 of them and lowers their mean absolute peak error from 15.26 to 8.94 cm. The remaining 13 events already have baseline error below 8 cm; only four improve, and their mean absolute peak error rises from 4.22 to 6.57 cm. The corrected errors therefore concentrate around a residual scale of several centimetres under the tested inputs and configuration.

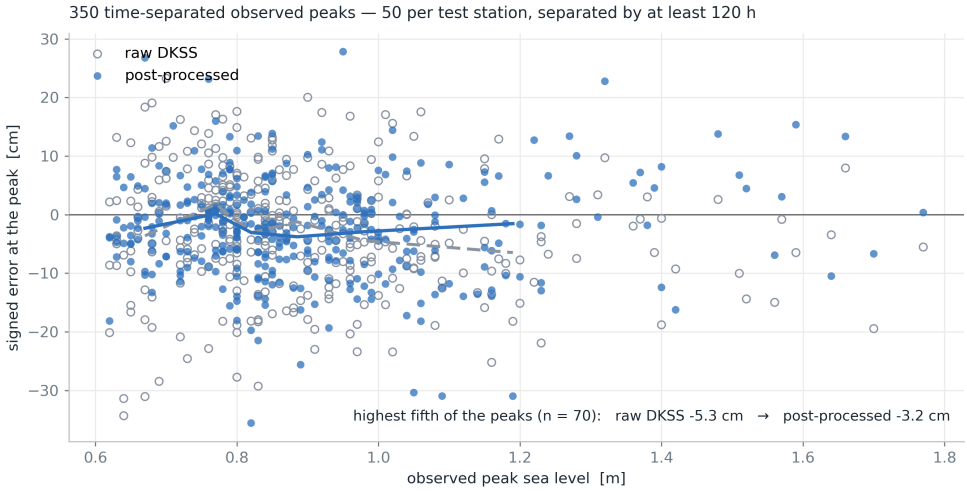


Figure 5.6: Signed error for raw DKSS and the final post-processed output at 350 observed peaks, 50 per test station and separated by at least 120 hours. Lines join medians in six equal-count observed-level bins. This endpoint view includes the fixed reference-level calibration; a residual spread and individual degradations remain.

The 350 peaks tell the same story without the event catalogue, and they show why the choice of baseline matters most at the crest. Measured from raw DKSS, the signed median error at a peak moves from −2.80 to −2.23 cm and RMSE from 10.53 to 8.86 cm; in the highest fifth of peaks the mean signed error moves from −5.3 to −3.2 cm. Those look like modest gains, and they are modest for a reason: β_i is positive at most test stations, so the raw series sits above the gauge on average and part of its underestimation at high water is cancelled by an offset that has nothing to do with the storm. Remove that offset and the learned step alone carries the signed median from −8.46 to −2.23 cm and the highest-fifth mean from −10.3 to −3.2 cm. The level-dependent underestimation is dynamic, and it is the dynamic correction that removes it, though what survives is still dispersed and does not go away.

Section 1.1.1 argued that this is the quantity with operational consequences. A median of only 18 cm separates the published 20- and 100-year design levels, so cutting more than six centimetres from a systematic peak underestimation narrows, without closing, the gap that puts an event in the wrong return-period class. These are retrospective errors at observed crests, not calibrated probabilities of future exceedance.

What separates a success from a failure is the starting error rather than the

severity of the event: the same correction that recovers a crest missed by more than 15 cm will disturb one already reproduced to within a few centimetres.

5.3 Secondary robustness example: regional specialisation

Chapter 3 showed that the 28 gauges do not sample one homogeneous coastline: strait geometry, wind exposure and open-boundary influence vary systematically along the domain. A reasonable objection to a single shared model is therefore that it averages over regimes that would be better learned apart. This section tests that objection, and it is the only place in the thesis where the network is treated as anything other than one population.

The four groups partition the same 28 stations: eastern Jutland, Samsø and northern Funen (R1); the Little Belt and southern Funen (R2); Øresund and the western Baltic (R3); and the Great Belt and Smålandsfarvandet (R4). Architecture, seed and optimization protocol are held fixed while one model is fitted inside each group and evaluated only on that group's test station or stations, so the comparison changes the training population and nothing else.

The event context of those four groups is summarized in Table 5.2. It is descriptive and does not define either the training loss or the scored event sample.

Table 5.2: Measured level and official class of the largest event, together with the total number of registered station-events in each region (Poulsen et al. 2024). Levels are the measured values reported in the station sheets.

Region	Highest-level station	Date	Level [cm]	Class	Events
R1 (Kattegat / E Jutland)	Grenaa	06 Dec 2013	165	20-yr	71
R2 (Little Belt / S Funen)	Aabenraa	05 Jan 2017	177	20-yr	60
R3 (Øresund / W Baltic)	Hornbæk	06 Dec 2013	196	100-yr	39
R4 (Great Belt)	Slipshavn	06 Dec 2013	146	20-yr	22

The regional experiment asks whether restricting training to a geographical subset improves on the single network-wide model. It does not do so systematically. On identical rows, regional minus global complete-record RMSE is -0.002 cm in R1, $+1.074$ cm in R2, $+0.192$ cm in R3 and $+1.805$ cm in R4. R1 is effectively tied; the other three regions are worse.

The pooled comparison in Table 5.3 shows a narrower trade-off. Regional training is worse over the complete record (6.00 versus 5.37 cm) and slightly worse across complete event windows (7.61 versus 7.51 cm), but its mean peak error is lower (7.87 versus 8.25 cm). The secondary experiment therefore suggests a possible peak-specific benefit, but not a consistently stronger post-processing model.

R4 is the clearest boundary case. Its validation set contains only Slipshavn and selects epoch 20. The resulting regional model reaches 7.01 cm compared with 5.21 cm for the network-wide model at Korsør. This is not evidence that geographical structure is irrelevant; the pooled peak result points in the opposite direction. It

Table 5.3: Errors in centimetres on identical test observations. Complete-record and event-window RMSE values are pooled over the eligible hours; peak values are mean absolute errors. "Regional" combines the four regional models, and the bias-corrected DKSS baseline is common to both strategies. The regional protocol is the one defined at the start of this section.

Evaluation register	Baseline	Network-wide	Regional
Complete-record RMSE	8.389	5.374	5.996
KDI event-hour RMSE (±60 h; pooled)	11.038	7.508	7.606
KDI observed-peak MAE	12.074	8.254	7.867

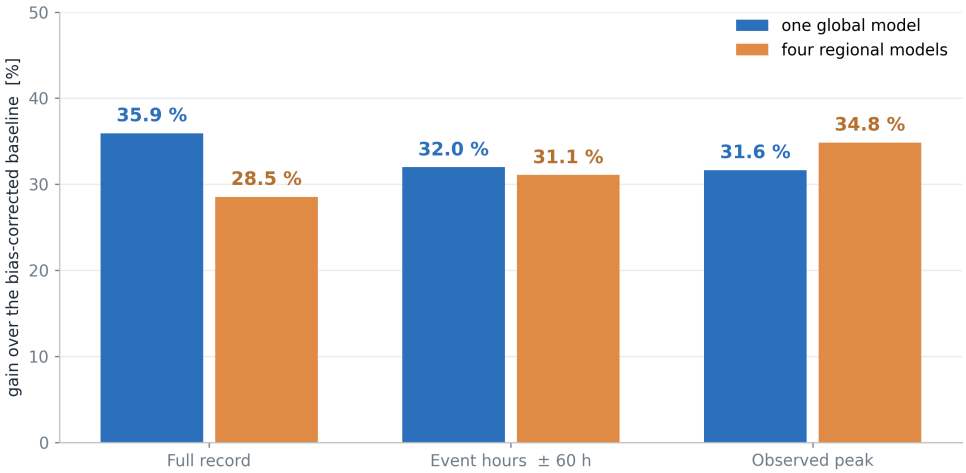


Figure 5.7: Comparison of the network-wide model with the four regionally trained models. Regional training changes only the training population; architecture, seed and test observations remain fixed.

shows that the larger and more diverse training population is the more robust default for complete-record and event-window error under the declared split.

5.4 What the model uses and what it removes

5.4.1 Which inputs carry the transferable signal?

Seven leave-one-group-out models were retrained independently with the same architecture, split, seed and full-record bias definition as the selected model. Figure 5.8 and Table 5.4 order them by the change in their minimum validation extreme-oriented loss relative to the full model. Test RMSE and tail metrics appear only as descriptive diagnostics; neither the ranking nor the frozen predictor set is revised after seeing them.

Removing DKSSstdev24 produces the largest validation deterioration (ΔEoL = +0.0270), followed by the atmospheric EOFs (+0.0237) and local SLP (+0.0148). Their training EoL also rises, especially without the EOFs, which is consistent with the network using both recent hydrodynamic variability and domain-scale meteorological state rather than relying on one of them alone. Local u_{10} has a smaller positive validation effect (+0.0066), while removing T_{2m} is nearly neutral (+0.0004).

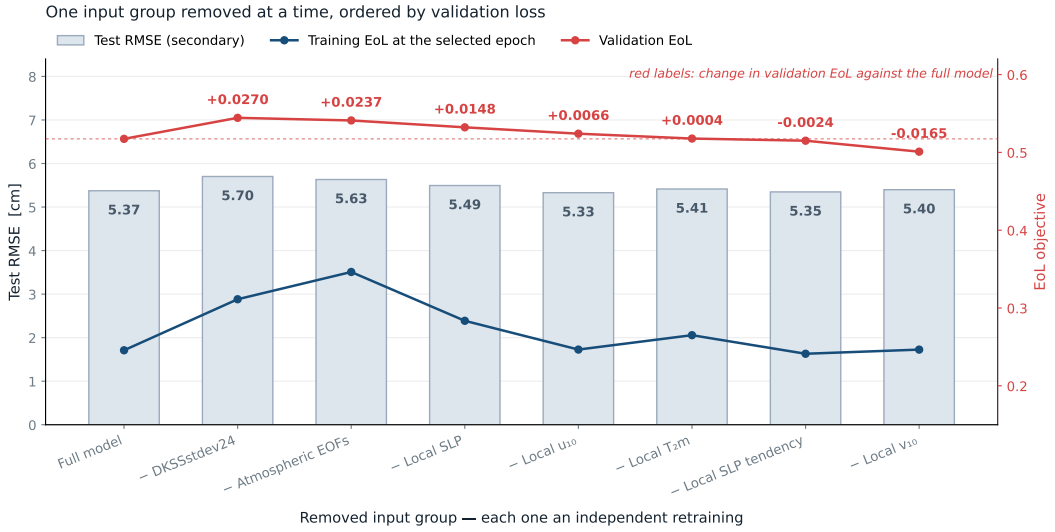


Figure 5.8: Independent retraining after removing one input group. The full model is the fixed reference; subsequent groups are ordered from the largest to the smallest increase in minimum validation EoL. Red labels report that validation change, the navy line shows training EoL at the selected epoch, and bars show test RMSE as a secondary diagnostic. Positive validation changes indicate worse model selection performance after removal.

The two negative changes caution against treating every retained input as independently necessary: removing local SLP tendency and local v_{10} lowers validation EoL by 0.0024 and 0.0165, respectively. Test RMSE spans 5.33–5.70 cm and does not follow the validation ranking; its lowest value occurs without u_{10}, but this is not a basis for post-hoc reselection. With one seed and one fixed station split, the battery identifies dependence and possible redundancy under the declared experiment, not

Table 5.4: Leave-one-group-out results for seed 42. Error values are in centimetres. The full model is followed by ablations ordered by validation EoL, not by held-out test performance. q99 and q99.9 use the fixed station-specific thresholds defined from the pre-2021 reference interval.

Configuration	Δ val. EoL	Test RMSE	q99	q99.9	Best epoch
Full model	0	5.374	8.860	11.178	83
− DKSSstdev24	+0.0270	5.702	8.749	10.413	58
− Atmospheric EOFs	+0.0237	5.633	8.627	11.031	59
− Local SLP	+0.0148	5.494	8.893	12.370	55
− Local u_{10}	+0.0066	5.329	8.527	10.851	87
− Local T_{2m}	+0.0004	5.414	8.631	10.581	47
− Local SLP tendency	−0.0024	5.348	8.593	11.230	108
− Local v_{10}	−0.0165	5.398	8.954	10.851	105

universal feature importance or a statistically resolved optimal subset.

5.4.2 Remaining frequency structure

The frequency domain shows what the aggregate metrics cannot. At Grenaa, Faaborg and Drogden-Fyr, the darker post-processing curve lies below the lighter raw-DKSS-error curve through most of the multi-day continuum. The reduction is especially visible at the longest resolved periods, while narrow diurnal and semidiurnal lines remain. The network therefore removes a substantial part of the slowly varying, meteorologically organised residual without resolving all short-period structure.

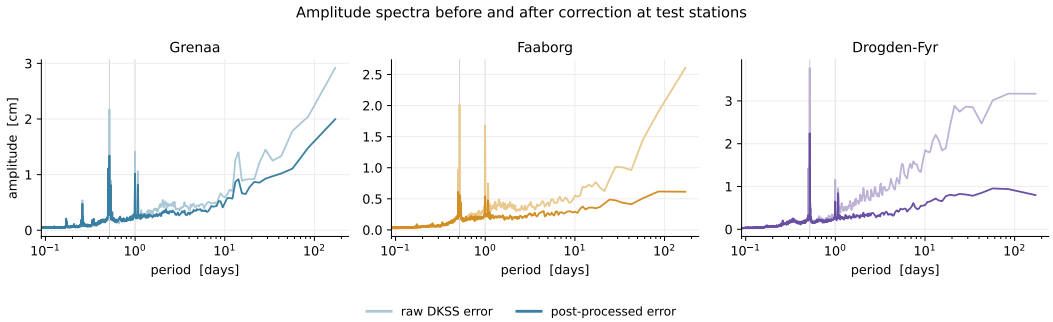


Figure 5.9: Complete-record residual spectra at Grenaa, Faaborg and Drogden-Fyr, shown in blue, orange and purple, respectively. Lighter lines denote raw DKSS error and darker lines the final post-processed error. A constant reference-level shift disappears under demeaning, so the raw and bias-corrected baseline spectra coincide outside zero frequency. The broad synoptic band contracts strongly, while narrow short-period lines remain visible.

This frequency-domain result closes the loop with the diagnosis in Chapter 3. The autocorrelation and broad multi-day continuum motivated a long input history ending at prediction time; the contraction of that continuum shows that the TCN learned part of the temporally organised error. The surviving narrow lines also explain why a non-zero local error remains after post-processing.

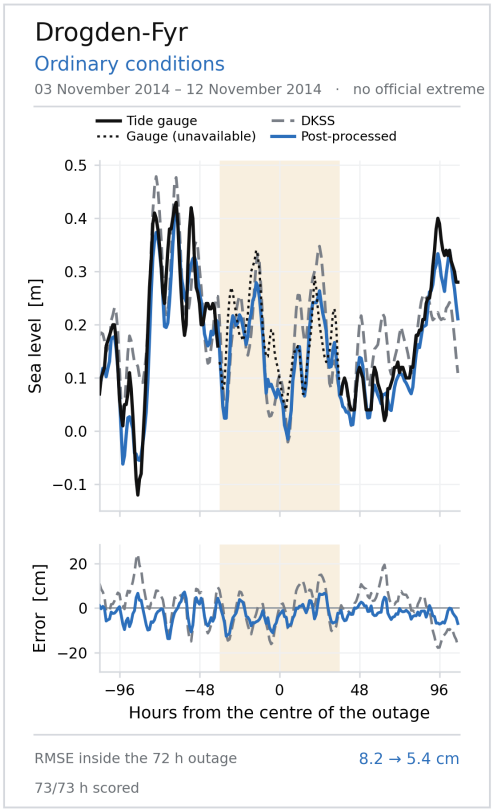
5.5 Finding 3: Post-processing remains available during gauge outages

Because inference does not use the contemporaneous gauge observation, the same correction remains available during an outage. Across 52 masked windows (45 registered events and seven ordinary periods), macro-mean RMSE falls from 10.87 to 7.50 cm (31.1%), with improvement in 40 cases. Event windows improve from 11.33 to 7.95 cm in 34 of 45 cases, and ordinary windows from 7.94 to 4.58 cm in six of seven cases.

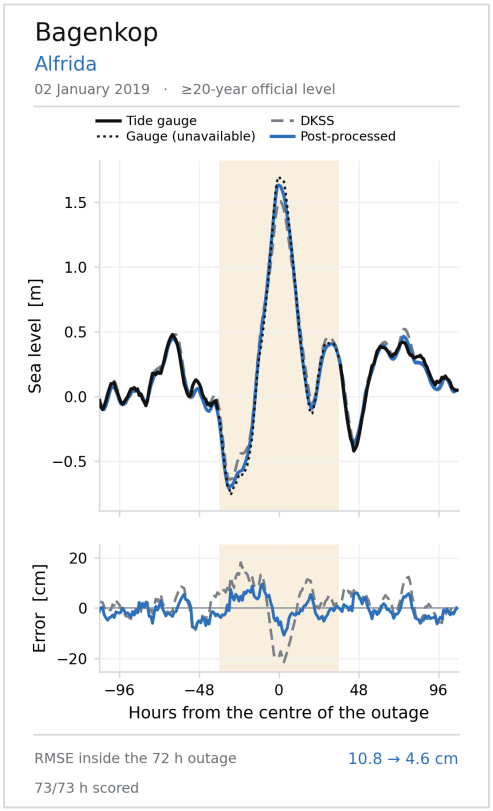
Table 5.5: RMSE in centimetres inside the 72-hour masked intervals. "Macro mean" averages the RMSE calculated separately in each window; "pooled hours" calculates one RMSE over every valid masked hour in the stated register. The observations are hidden from the post-processing model and retained only for scoring.

Register	Aggregation	Baseline	Post-processed	Reduction	Support
All windows	Macro mean	10.87	7.50	31.1%	40/52 improved
All windows	Pooled hours	12.35	8.13	34.2%	3,699 h
Registered events	Macro mean	11.33	7.95	29.8%	34/45 improved
Registered events	Pooled hours	12.83	8.54	33.4%	3,188 h
Ordinary conditions	Macro mean	7.94	4.58	42.3%	6/7 improved
Ordinary conditions	Pooled hours	8.78	4.79	45.4%	511 h

Figure 5.10 uses the same card format as the storm examples so that the operating condition is explicit. The observed series is masked inside the shaded 72-hour outage; its dotted continuation is displayed only to score the reconstruction. The TCN receives no contemporaneous residual or gauge value from that interval.



(a) Ordinary conditions at Drogden-Fyr.



(b) Alfrida at Bagenkop.

Figure 5.10: Masked-gap reconstructions with the final post-processing model. Black is the available observed record, dotted black is the observation hidden from the model, dashed grey is raw DKSS and blue is the reconstructed post-processed series. The printed score is raw-to-final RMSE inside the 72-hour outage; every masked hour with a valid hidden observation is scored. Table 5.5 separately reports the incremental comparison with bias-corrected DKSS.

It is worth being precise about how narrow this is. Once a station has an established vertical reference, the dynamic correction keeps working when the live gauge record stops, which is a statement about the predictor set, not about deployment at an unobserved coast.

5.6 Synthesis of the findings

The evaluation settings differ in how much skill they indicate; they agree about its shape. The network removes several structured components of DKSS error, including synoptic-band variance, level-dependent peak underestimation and station-to-station spread. What it leaves behind is strikingly consistent across evaluation settings: about 5.4 cm over the complete record, 8.25 cm at registered event peaks and 11.18 cm for hours above the station-specific q99.9 thresholds.

That scale accounts for the successes and the failures under a single rule. Stations and events whose baseline sits well above it improve strongly; those that begin below it can be made worse. The regional experiment fits the same reading, since smaller training populations did not estimate the shared structure any more reliably. The answer to the research question is consequently yes, with its limits attached: the dynamic residual is learnable and it transfers between stations, but the predictors available here neither remove the full local error nor say anything about how confident the correction is.

Chapter 6

Conclusions

This thesis investigated how much of the dynamic DKSS residual can be learned by a single post-processing model and whether that correction transfers to stations excluded from model development, including during registered high-water events. Under the declared station split, a substantial part of the residual is transferable. The largest benefits occur where the hydrodynamic model exhibits a strong and systematic error, while smaller gains are obtained where the original prediction is already accurate.

6.1 Main findings

The dynamic residual transfers across stations

One set of network weights provides a useful correction at stations excluded from training and validation. The complete endpoint comparison reduces pooled RMSE from 10.90 cm for raw DKSS to 5.37 cm after post-processing, corresponding to a 50.7% reduction. Within this complete correction chain, the neural component reduces RMSE from 8.39 to 5.37 cm relative to DKSS after removal of the static station bias.

Because the residual records of the seven test stations entered neither parameter fitting nor checkpoint selection, the experiment demonstrates spatial transfer of a shared residual relationship rather than independent fitting at individual stations.

High-water underestimation is substantially reduced

The hydrodynamic error becomes increasingly negative toward high observed sea levels, showing a systematic tendency to underestimate the largest peaks. The post-processing model substantially reduces this behaviour.

Across 350 selected peaks, with peaks at each station separated by at least 120 hours, the signed median error changes from -8.46 cm for bias-corrected DKSS to -2.23 cm after post-processing. For the 45 KDI registered station events, mean absolute error at the observed peak decreases from 12.07 to 8.25 cm. For the seven events classified in the 20- and 100-year categories, it decreases from 15.38 to 5.53 cm.

The correction is therefore particularly effective when the original hydrodynamic model substantially underestimates the event. It is less beneficial when the original peak is already reproduced within a few centimetres.

Performance becomes more uniform across stations

Post-processing also reduces the spatial variability of the error. Across the seven test stations, the standard deviation of station RMSE decreases from 1.76 to 0.63 cm, while the range decreases from 5.47 to 1.98 cm. After correction, station RMSE lies between 4.08 and 6.06 cm.

The correction therefore does more than improve the pooled mean. It also produces a more consistent level of performance across geographically separated stations.

A residual error remains

The correction does not drive the error to zero. The final model reaches approximately 5.4 cm RMSE over the complete record and about 8.3 cm mean absolute error at the crest of registered events. These values should not be interpreted as universal limits. They describe the residual variability that remains unexplained by the predictors and model configuration used in this thesis.

6.2 Scientific and methodological contribution

The main contribution is a residual correction framework designed around spatial transfer rather than a new neural architecture. The approach separates the persistent station reference bias from the dynamic error, uses only atmospheric and hydrodynamic information available at the prediction time, and evaluates one shared model at stations whose residuals were excluded from model development.

The final model uses 16 predictor channels. Ten describe the domain-scale atmospheric state through EOF scores, while six describe local atmospheric conditions and recent DKSS variability. The TCN contains 151,809 trainable parameters, six residual blocks and a receptive field of 316 hours.

Removing recent hydrodynamic variability or atmospheric EOFs caused the largest increases in validation loss under the tested configuration. The result is consistent with the physical interpretation developed in Chapter 3: local coastal error depends both on recent local model behaviour and on the larger atmospheric state acting across the connected Danish and western Baltic waters.

The extreme event evaluation is independent of the threshold used during training. The high-water branch of the loss is activated using a station-specific observed quantile, whereas the final storm evaluation uses the external event register published by the Danish Coastal Authority.

6.3 Operational interpretation

The model is best interpreted as a post-processing layer added directly downstream of an existing hydrodynamic forecasting system. In an operational setting, DKSS would first produce its forecast in the usual way. The post-processing model would then receive the available atmospheric and hydrodynamic predictor histories, estimate the dynamic residual for the forecast location, and immediately return a corrected sea level. The final product delivered to the user would therefore remain physically anchored to the hydrodynamic forecast, but with both the persistent station bias and the predictable dynamic error removed.

This is the operational role envisaged for the framework: not replacing the forecasting model, but refining its output before it is used for warning, threshold assessment or other coastal decisions. The intended chain is therefore straightforward: hydrodynamic forecast, post-processing correction, corrected forecast.

This distinction is particularly relevant at high water. The results show that the original model tends to underestimate some of the largest observed peaks and that the post-processing model substantially reduces this tendency. Since the differences between published Danish return levels can themselves be of the order of only a few tens of centimetres, reducing a systematic peak error by several centimetres can materially change the interpretation of an event relative to an operational threshold.

The correction can also be evaluated when the local tide gauge is temporarily unavailable because the contemporaneous observation is not an input to the model. Across the masked gauge experiments, this produces lower reconstruction errors than the hydrodynamic model alone. This should nevertheless be interpreted as reconstruction at historically calibrated stations, rather than as evidence that the method already works at locations with no observational history.

The present experiment uses zero-lead historical inputs and therefore demonstrates the potential of this operational post-processing concept rather than a complete deployed forecasting system. A future implementation would require archived atmospheric and hydrodynamic forecasts at explicit lead times, together with latency-aware preprocessing and evaluation of how the correction changes with forecast horizon.

6.4 Limitations

The first limitation is that stations are processed independently by the shared temporal model. Although every station receives the same domain-scale atmospheric EOFs, no information is exchanged directly between neighbouring or dynamically connected stations. The Danish Straits form a connected hydrodynamic system, and an error observed or inferred at one location may contain information relevant to another. A graph neural network could represent stations as nodes connected according to geographical or hydrodynamic relationships and allow information to propagate through the network. The present TCN does not exploit this type of connectivity.

The second limitation is the deterministic nature of the correction. The model returns one residual estimate and provides no measure of predictive uncertainty. This is particularly restrictive for high-water applications, where the distinction between a prediction close to a warning threshold with low uncertainty and the same prediction with high uncertainty is operationally important. The reported results quantify average and event-specific errors but do not provide calibrated probabilities or confidence ranges.

The third limitation is the absence of predictors that describe only spatial properties of a station. The model uses atmospheric histories and recent DKSS variability but does not receive bathymetry, coastline geometry, fetch, water depth, coordinates or explicit information about the connectivity of straits and basins. Such quantities do not change through time, but they control how the same atmospheric forcing is transformed into a local sea-level response. Their absence may contribute to the remaining differences in performance between stations.

These three limitations are closely related. The current model learns a common temporal relationship and relies on the input histories to contain enough information for spatial transfer. It does not explicitly describe where a station is, how it is connected to the rest of the basin, or how uncertain its correction is.

6.5 Future work

The first priority is to test whether the correction remains valid outside the tide-gauge network used in this thesis. The present experiment demonstrates transfer between observed stations, but it does not provide an independent validation between gauges or offshore. Satellite altimetry offers the most direct route to such a test. A future experiment could compare post-processed DKSS sea level with quality-controlled altimetric observations at locations not represented by the tide-gauge network. This would require careful treatment of the common vertical reference and of the reduced reliability of satellite altimetry close to the coast, but it would provide a genuinely independent test of spatial generalization.

A second development is to represent the station network explicitly as a connected system. A graph neural network could combine the temporal representation developed here with links between stations based on distance, basin connectivity, straits or estimated hydrodynamic relationships. Such a model could test whether information from neighbouring locations improves the correction of local errors, particularly during events that propagate across the Danish waters.

Static spatial predictors should be investigated at the same time. Bathymetry, coastline orientation, local depth, fetch and other geographical descriptors could provide information about why identical atmospheric conditions generate different errors at different stations. Their contribution should be tested carefully to ensure that they encode transferable physical structure rather than simply acting as station identifiers.

The deterministic correction should also be extended toward probabilistic predic-

tion. Predicting quantiles or a full conditional distribution would provide uncertainty information around the corrected sea level and allow evaluation through reliability, sharpness and threshold exceedance probabilities in addition to conventional error metrics.

Finally, the framework should be evaluated with archived atmospheric and hydrodynamic forecasts at explicit lead times. This would move the analysis from the present zero-lead experiment toward a forecasting system and would allow separate assessment of amplitude, peak timing, threshold detection and forecast degradation with increasing lead time.

References

Andrée, Elin, Jian Su, Morten Andreas Dahl Larsen, Martin Drews, Martin Stendel, and Kristine Skovgaard Madsen (2023). "The role of preconditioning for extreme storm surges in the western Baltic Sea". In: *Natural Hazards and Earth System Sciences* 23.5, pp. 1817–1834. DOI: 10.5194/nhess-23-1817-2023 (cit. on pp. 2, 25).

Campos-Caba, Rodrigo, Paula Camus, Andrea Mazzino, Michalis Vousdoukas, Massimo Tondello, Ivan Federico, Salvatore Causio, and Lorenzo Mentaschi (2025). "Storm surge dynamics in the northern Adriatic Sea: comparing AI emulators with high-resolution numerical simulations". In: *EGUsphere [preprint]*. DOI: 10.5194/egusphere-2025-5313 (cit. on pp. 36, 37).

Coles, Stuart (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. London: Springer. DOI: 10.1007/978-1-4471-3675-0 (cit. on p. 32).

Frishfelds, Vilnis, Jun She, and Jens Murawski (2025). "On-demand coastal modelling with two-way nesting". In: *Ocean Dynamics* 75, p. 23. DOI: 10.1007/s10236-025-01670-x (cit. on p. 25).

Giaremis, Stefanos, Noujoud Nader, Clint Dawson, Hartmut Kaiser, Carola Kaiser, and Efstratios Nikidis (2024). *Storm Surge Modeling in the AI Era: Using LSTM-based Machine Learning for Enhancing Forecasting Accuracy*. arXiv preprint arXiv:2403.04818 (cit. on p. 37).

Han, Yuchen, Yiyang Wang, Lei Wang, Changze Zhou, and Shijin Yuan (2026). "Extreme-oriented loss: powering a novel framework for improving spatiotemporal met-ocean forecasting". In: *Expert Systems with Applications* 309, p. 131115. DOI: 10.1016/j.eswa.2026.131115 (cit. on p. 45).

Jahanmard, Vahidreza, Robinson Hordoir, Nicole Delpeche-Ellmann, and Artu Ellmann (2023). "Quantification of hydrodynamic model sea level bias utilizing deep learning and synergistic integration of data sources". In: *Ocean Modelling* 186, p. 102286. DOI: 10.1016/j.ocemod.2023.102286 (cit. on pp. 3, 6, 17, 37, 39).

Leppäranta, Matti and Kai Myrberg (2009). *Physical Oceanography of the Baltic Sea*. Berlin, Heidelberg: Springer (cit. on pp. 25, 40).

Madsen, Henrik (2007). *Time Series Analysis*. Chapman and Hall/CRC (cit. on pp. 3, 39).

North, Gerald R., Thomas L. Bell, Robert F. Cahalan, and Fanthune J. Moeng (1982). "Sampling Errors in the Estimation of Empirical Orthogonal Functions". In: *Monthly Weather Review* 110.7, pp. 699–706 (cit. on p. 36).

Oord, Aäron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv preprint arXiv:1609.03499 (cit. on p. 44).

Poulsen, Bastian, Christoffer Grupe, Simon Frey Frank, Thorsten Piontkowitz, and Lene Bonde (2024). *Højvandsstatistikker 2024*. Tech. rep. Revised 5 November 2024. ISBN 978-87-7179-004-7. Available at `https://kyst.dk/klimatilpasning/vaerktoejer/hoejvandsstatistikker`. Lemvig, Denmark: Kystdirektoratet, Miljøministeriet (cit. on pp. 2, 10, 31–34, 58).

Rus, Marko, Anja Fettich, Matej Kristan, and Matjaž Ličer (2023). "HIDRA2: deep-learning ensemble sea level and storm tide forecasting in the presence of seiches – the case of the northern Adriatic". In: *Geoscientific Model Development* 16, pp. 271–288. DOI: `10.5194/gmd-16-271-2023` (cit. on pp. 36, 37).

She, Jun and Jacob Woge Nielsen (2019). "'Silent' storm surge extremes in the western Baltic Sea on 4 January 2017". In: *Journal of Operational Oceanography* 12.sup1, S111–S117. DOI: `10.1080/1755876X.2019.1633075` (cit. on pp. 2, 25).

Tedesco, Paulina, Jean Rabault, Martin Lilleeng Sætra, Nils Melsom Kristensen, Ole Johan Aarnes, Øyvind Breivik, Cecilie Mauritzen, and Øyvind Sætra (2024). "Bias correction of operational storm surge forecasts using Neural Networks". In: *Ocean Modelling* 188, p. 102334. DOI: `10.1016/j.ocemod.2024.102334` (cit. on pp. 6, 36, 37).

Declaration of tools and computational resources

Software and development environment

Visual Studio Code was used as the primary local development environment for writing, reviewing and debugging the analysis code. The computational workflow was implemented primarily in Python and PyTorch, with Git and GitHub used for version control and traceability. The thesis was prepared in LaTeX and synchronized with Overleaf.

High-performance computing

Model development, training and evaluation used the DTU High-Performance Computing facility and its LSF batch system. Development experiments were submitted to both the gpuv100 and gpua100 queues, which provided NVIDIA V100 and A100 accelerators. The final 16-input model and its associated final evaluation runs were executed through the gpuv100 queue. Each final training job requested one GPU in exclusive-process mode; the documented device was an NVIDIA Tesla V100 PCIe GPU with 32 GB of memory. CPU resources were also used for data preparation, quality control, post-processing and figure generation.

AI-assisted and language tools

ChatGPT by OpenAI and Claude by Anthropic were used as assisted-development tools. They supported code review, debugging, refactoring, test design, documentation and suggestions for improving computational efficiency. Grammarly was used for spelling, grammar and language-style correction.

All suggestions produced by these tools were reviewed before acceptance. The author retained responsibility for the research design, implementation, verification of code and results, interpretation of the evidence and final wording of the thesis. No numerical result reported in the thesis was accepted from a generative tool without verification against the executed analysis and its stored outputs.