JGR Oceans

RESEARCH ARTICLE

10.1029/2020JC016402

Key Points:

- Generative Adversarial Networks successfully reconstruct basin-wide sea level in the North Sea using data from tidal gauges
- Machine learning appeared successful when learning from different data sources
- The proposed method is skillful at learning and replicating processes with multiple time scales

Supporting Information:

- Supporting Information S1

Correspondence to:

E. V. Stanev,
emil.stanev@hzg.de

Citation:

Zhang, Z., Stanev, E. V., & Grayek, S. (2020). Reconstruction of the basin-wide sea-level variability in the North Sea using coastal data and generative adversarial networks. *Journal of Geophysical Research: Oceans*, *125*, e2020JC016402. https://doi.org/10.1029/2020JC016402

Received 14 MAY 2020
Accepted 9 NOV 2020

Reconstruction of the Basin-Wide Sea-Level Variability in the North Sea Using Coastal Data and Generative Adversarial Networks

Zeguo Zhang[1], Emil V. Stanev[1], and Sebastian Grayek[1]

[1]Institute of Coastal Research, Helmholtz-Zentrum Geesthacht, Geesthacht, Germany

Abstract We present an application of generative adversarial networks (GANs) to reconstruct the sea level of the North Sea using a limited amount of data from tidal gauges (TGs). The application of this technique, which learns how to generate datasets with the same statistics as the training set, is explained in detail to ensure that interested scientists can implement it in similar or different oceanographic cases. Training is performed for all of 2016, and the model is validated on data from 3 months in 2017 and compared against reconstructions using the Kalman filter approach. Tests with datasets generated by an operational model ("true data") demonstrated that using data from only 19 locations where TGs permanently operate is sufficient to generate an adequate reconstruction of the sea surface height (SSH) in the entire North Sea. The machine learning approach appeared successful when learning from different sources, which enabled us to feed the network with real observations from TGs and produce high-quality reconstructions of the basin-wide SSH. Individual reconstruction experiments using different combinations of training and target data during the training and validation process demonstrated similarities with data assimilation when errors in the data and model were not handled appropriately. The proposed method demonstrated good skill when analyzing both the full signal and the low-frequency variability only. It was demonstrated that GANs are also skillful at learning and replicating processes with multiple time scales. The different skills in different areas of the North Sea are explained by the different signal-to-noise ratios associated with differences in regional dynamics.

Plain Language Summary The variability of sea level is one of the most important elements of the ocean dynamics. Basin-wide observations are due to satellite altimeters, observations in coastal stations are provided by tidal gauges. The first are not very accurate in the coastal areas, the second do not provide basin-wide coverage. The task in the present work is to use machine learning to reconstruct the sea-level variability in the North Sea, which is an almost enclosed ocean region, using observations only. Using data from 19 coastal stations and data from numerical models as a representation of the true ocean (synthetic observations), we demonstrated that the generative adversarial networks reconstruct almost perfectly the sea level of the North Sea. The application of this technique, which learns how to generate datasets with the same statistics as the training set, is explained in detail to ensure that interested scientists can implement it in similar or different oceanographic cases.

1. Introduction

In the first part of the 20th century, Proudman and Doodson (1924) demonstrated how the fundamental dynamical equations of the tides may be used to obtain knowledge of the distribution of the surface elevation over the entire North Sea from observational data. They showed that the cotidal and corange lines can be easily determined, provided the elevation in some coastal stations and limited amount of open sea locations is known, and some local current observations exist. Additionally, they used some hypotheses about the frictional forces. In the present study, approximately 100 years later, we encounter the same issue from a different perspective.

Many important developments in the physical oceanography of the North Sea have followed the study of Proudman and Doodson (1924). Numerical modeling of tides and storm surges initiated by Hansen (1956) and Heaps (1969) has become a fundamental tool in surge prediction (Flather & Proctor, 1983; Peeck et al., 1983; Soetje & Brockmann, 1983). A dense network of tidal stations has been developed around

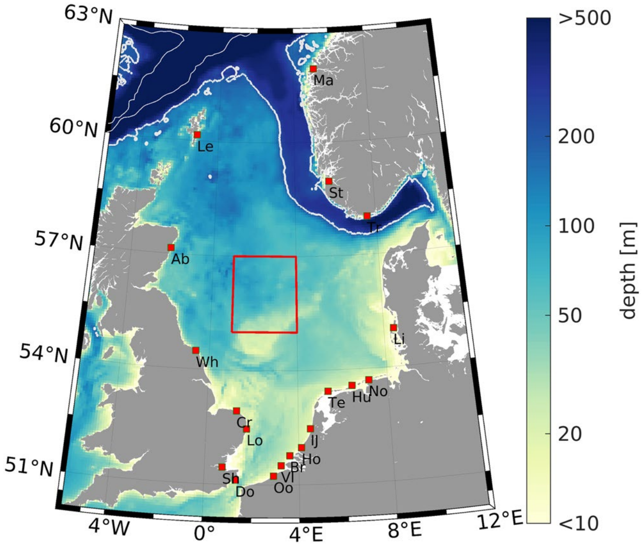

Figure 1. Topography of the North Sea, the positions of the TG stations (red squares), and the subsampled region (grid of 32 × 32 points) shown by the large red rectangle. Ab, Aberdeen; Br, Brouwershavensegat 8; Cr, Cromer; Do, Dover; Hu, Huibertgat; Ij, IJgeulstroompaal 1; Le, Lerwick; Li, List; Lo, Lowestoft; Ma, Maloy; No, Hoek van Holland; No, Norderne; Oo, Oostende; Sh, Sheerness; St, Stavanger; Te, Terschelling Noordzee; Tr, Tregde; VL, Vlakte van de Raan; Wh, Whitby.

the North Sea coasts, which operates over long periods and provides high-quality records (Wahl et al., 2013). A comparison between numerical simulations and satellite observations (Andersen, 1999) revealed good agreement between the two data sources. The low-frequency variability in altimeter and tide gauge data over the North Sea shows a reasonably good correlation (Cipollini et al., 2017). A recent important development in predicting the sea level in the North Sea is achieved within the framework of the North-West European Shelf forecasting system (Tonani et al., 2019).

The North Sea (Figure 1) is a shallow sea located at the European continental shelf with an average depth of ~90 m (Becker et al., 1992; Huthnance, 1991; Otto et al., 1990). The sea-level dynamics in this basin can be considered as a response to different forcings, such as barotropic and baroclinic tides (Haigh et al., 2019), wind and atmospheric pressure, air-sea heat and water exchanges, as well as forcing from the open boundaries and rivers. The processes that dominate the dynamics are, in most cases, coupled; that is, one cannot easily consider the response to individual drivers in isolation. Thus, there is a need to use analysis methods tailored to detect and reproduce nonlinear dynamics. Deep learning, which has much in common with neural networks, is well suited to resolve such processes. Our first objective in the present study is to explore the performance of deep learning techniques when reconstructing the basin-wide sea level in the North Sea using data from coastal stations. In our specific application, we will use generative adversarial networks (GANs; Goodfellow et al., 2014). This technique learns how to generate datasets with the same statistics as the training set. Unlike some previous studies (e.g., Cipollini et al., 2017), we will focus both on the shorter- and longer-term variability ranging from intratidal to monthly time scales. Under "exploring the performance of deep learning techniques," we also identify the application limits. This will be illustrated by setting up several experiments with different reconstruction potentials.

Our second objective is to compare the goodness of reconstructions based on adversarial networks against other known reconstruction methods. One such method, which uses a Kalman filter approach, was proposed by Schulz-Stellenfleth and Stanev (2010) as an instrument to reconstruct sea level in the German Bight using a small number of observations (tide gauges, satellite altimeters, and high-frequency radar). The same method was applied by Grayek et al. (2011) to extrapolate one-dimensional FerryBox data acquired along the ferry routes to larger two-dimensional areas. This (second) objective of the present research is in line with the recent study of Barth et al. (2020), who compared the performance of convolutional neural networks to reconstruct sea surface temperature satellite observations with the method known as Data INterpolating Empirical Orthogonal Functions (Alvera-Azcárate et al., 2005; Beckers & Rixen, 2003).

We are not aware of any applications of deep learning to sea-level analysis and reconstruction, particularly in the region of the North Sea. This justifies our third objective, which is to present our results in a way that they ensure reproducibility by interested scientists and motivate potential oceanographic applications using similar or different data sets. Therefore, we will analyze many individual steps that led to the final application of the method to the entire North Sea area. The major focus is on what GANs can reconstruct successfully and what they cannot. The analysis of the results demonstrates the power of reconstructions based on GANs. The study is structured as follows. In Section 2, we present the methods used. Section 3 presents the experiments; Section 4 presents the results, followed by the discussion in Section 5 and the conclusions.

2. Methods

2.1. Data

2.1.1. Data From the Operational Model for the North-West European Shelf

The reconstruction of the basin-wide sea level using data from coastal stations necessitates high-quality data from observations over the entire North Sea. Data with such coverage are available only from satellites. However, they cannot perfectly resolve the spatial and temporal variability, particularly at scales shorter than the time of the repeat cycle. Furthermore, close to the coast, these data are not quite accurate (Cipollini et al., 2017). Data from numerical models, although not absolutely correct, provide spatial and temporal coverage over the entire basin. Therefore, when developing and testing our method, we will use data from numerical simulations to represent the "true" sea level. In this research, the data set was obtained from the operational numerical Forecasting Ocean Assimilation Model with 7 km horizontal resolution, known as Atlantic Margin Model-7 (AMM7) (O'Dea et al., 2012), for 2016 and 2017. For brevity, we will refer to these data as to the AMM7 data below.

For the objective of the present research, sea-level data over 1 year are sufficient to cover some of the most important periodic variations. Therefore, we chose 1-year sea surface height (SSH) data, which are from approximately 8,640 hourly SSH maps, as the training data set. In addition, it is important to note that the training SSH map (in 2016) is independent from the validation data set. Our validation data set (from 2017) covers a total of 3 months (2,158 hourly SSH maps). This choice limits the analyses to processes with periods ranging from intratidal to monthly. In our study area and for the time ranges defined above, there are two basic processes that explain most of the variability. These are the short-periodic tides (daily and shorter periods) and atmospherically induced motions (e.g., due to synoptic variability in the atmosphere). As shown by Jacob and Stanev (2017), both types of motions are nonlinearly coupled, and their separation is not a trivial problem. To quantify the potential of deep learning techniques when analyzing and reconstructing SSH, we filtered signals with periods less than 48 h by using the Butterworth filter; thus, we will process two data sets: one data set containing all frequencies (briefly called AF) and the low-passed filtered data set (briefly called LF).

Example variability patterns of the AMM7 data are shown in Figure 2. The first two panels show the phase lines of the semidiurnal principal lunar (M2) tide (Figure 2a) and its amplitude (Figure 2b). They describe the known pattern of the dominant tidal oscillations consisting of three amphidromic areas; the Kelvin wave propagates counterclockwise (Proudman & Doodson, 1924). The standard deviation $s = \sqrt{1/(n-1)\sum_{i-1}^{n}(x_i - \overline{x})^2}$ between the current sea-level height x_i from the LF data set computed from AMM7 and its mean $\overline{x}$ for the period from January 01, 2016 to December 31, 2016 is shown in Figure 2c. In the above equation, n is the number of data maps. This panel quantifies the magnitude of low-frequency variability, which is largest in the coastal area, particularly in the German Bight. Notably, the spatial distribution of amplitudes caused by tides and wind is different. In the German Bight, the magnitude of the low-frequency signal is approximately two times lower than that of the signal associated with the M2 tide. Along the coasts of the British Isles, this ratio is larger than 5.

2.1.2. Data From the Geesthacht COAstal model SysTem for the North-West European Shelf

For the experiments discussed later in this study, we will need the output of another (independent) model. To this aim, we chose the numerical simulations performed in the Helmholtz-Zentrum Geesthacht based on the Nucleus for European Modelling of the Ocean (NEMO v3.6; Madec, 2016) with 3.5 km horizontal resolution, which is two times finer than in the AMM7. The respective model setup is part of the Geesthacht COAstal model SysTem (GCOAST), which is a coupled modeling framework that includes atmospheric, oceanic, wind wave, biogeochemical, and hydrological parts (Ho-Hagemann et al., 2020). For the purposes of the present study, we use only the ocean circulation part. The model area covers the Baltic Sea, the Danish Straits, the North Sea, and part of the Northeast Atlantic. The data used in the present study cover only the region shown in Figure 1. The vertical discretization uses 50 hybrid s-z* levels with partial cells. The model forcing for the momentum and heat fluxes is computed using bulk aerodynamic formulas and hourly data from atmospheric reanalyzes of the European Centre for Medium Range Weather Forecasts (ERA5

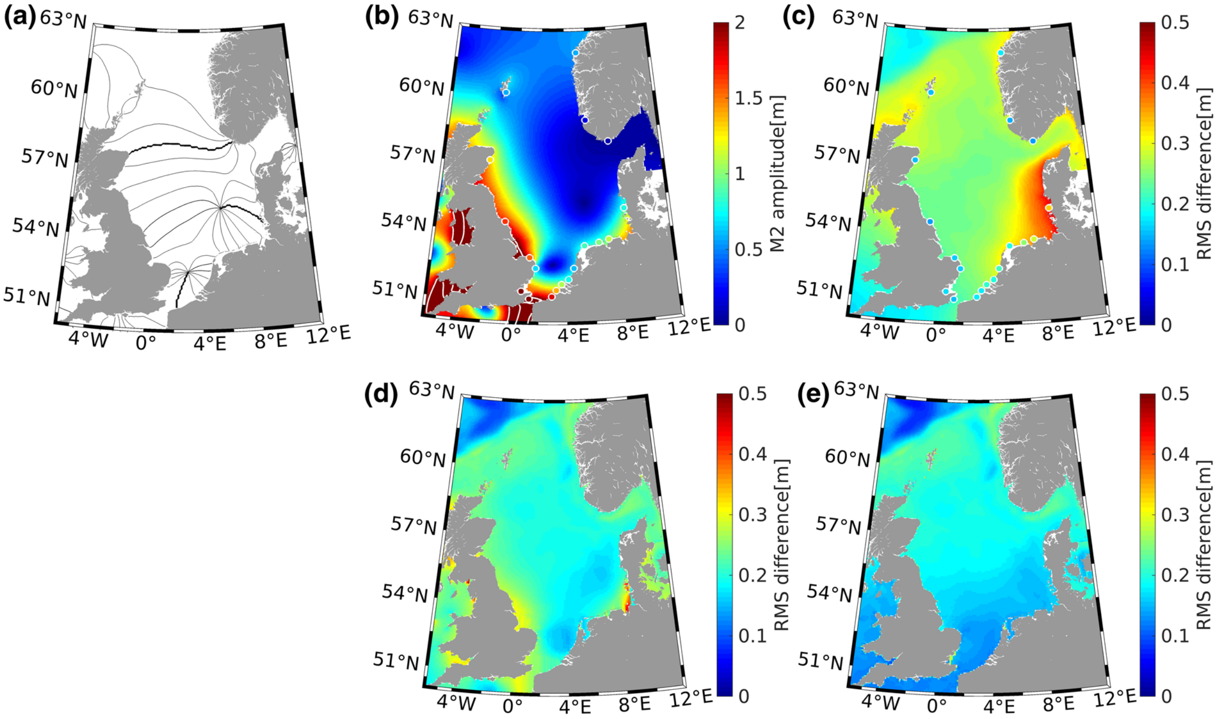

Figure 2. Phase lines of the M2 tide for the period computed from the AMM7 data using UTide (a). (b) is the amplitude corresponding to (a). The white isolines in (b) are lines of equal amplitude of 3, 4, and 5 m, respectively. (c) is the RMS of the LF variability of SSH. The respective values of the magnitudes of the M2 tide and the RMS variability from the TG sea level are superimposed with circles in (b) and (c). (d) and (e) are the RMS differences between the AMM7 and GCOAST models (AF, [d] and LF, [e]) shown in the AMM7 grids. AF, all frequencies (full data set); AMM7, Atlantic Margin Model-7; GCOAST, Geesthacht COAstal model SysTem; LF, low frequencies (low-pass filtered data set); RMS, root mean square; SSH, sea surface height; TG, tidal gauge.

ECMWF with a horizontal resolution of 0.25°). The tidal potential is also included in the model forcing (Egbert & Erofeeva, 2002). The daily climatology for the river runoff is based on river discharge datasets from the German Federal Maritime and Hydrographic Agency (Bundesamt für Seeschifffahrt und Hydrographie), the Swedish Meteorological and Hydrological Institute, and the United Kingdom Meteorological Office. The boundary conditions at the open boundaries use input from the AMM7 (O'Dea et al., 2012) distributed by the Copernicus Marine Environment and Monitoring Service. The output is stored hourly for 2016 and 2017. Data assimilation is not used.

Figures 2d and 2e show the root mean square (RMS) differences between the simulations produced by the AMM7 and GCOAST models over 1 year for the AF and LF datasets, respectively. Regarding the tidal signal, the differences between the two models are far below the level of variability (compare Figure 2d with Figure 2b). The largest deviations between the two models are located in the English Channel, in front of the mouth of the Elbe and around the Wash. Over most of the model area, the difference between the LF sea level in the two models is approximately two times lower than the standard deviation of the signal in each of them. This quantitative similarity between the GCOAST and AMM7 data is explained by the similar model setups, forcing, and boundary conditions. The major difference between the two models, which is that the horizontal resolution in GCOAST is two times finer than in AMM7, explains most of the differences between the two data sets.

2.1.3. Tidal Gauge Data

Observational data along the North Sea coast have been obtained from the Copernicus Marine Environment Monitoring Service (http://marine.copernicus.eu/). Altogether, 19 gauge stations with hourly resolution are used. Their positions are shown in Figure 1. The magnitudes of the M2 tide and the respective RMS variability of the observed sea level are superimposed with circular symbols in Figures 2b and 2c to illustrate the differences between the model and observational data. Obviously, these differences, which are quantified in Table 1, are one order of magnitude smaller than the magnitude of the respective signals (Figure 2).

Table 1
Quantification of Differences and Agreements Between Datasets in Coastal Stations

Tidal station	RMS (TG, AMM7) (m)	RMS (TG, GCOAST) (m)	RMS (AMM7, GCOAST) (m)
Lerwick	0.16	0.24	0.23
Aberdeen	0.17	0.20	0.23
Whitby	0.23	0.23	0.29
Cromer	0.33	0.24	0.26
Lowestoft	0.22	0.20	0.19
Sheerness	0.49	0.45	0.48
Dover	0.48	0.33	0.33
Oostende	0.22	0.22	0.31
Vlakte van de Raan	0.25	0.22	0.22
Brouwershavensegat 8	0.18	0.17	0.18
Hoek van Holland	0.15	0.15	0.14
IJgeulstroompaal 1	0.31	0.29	0.16
Terschelling Noordzee	0.25	0.22	0.17
Huibertgat	0.28	0.20	0.19
Norderney	0.22	0.26	0.23
List	0.25	0.27	0.26
Tregde	0.17	0.13	0.23
Stavanger	0.15	0.13	0.20
Maloy	0.18	0.12	0.24

Abbreviations: AMM7, Atlantic Margin Model-7; GCOAST, Geesthacht COAstal model SysTem; RMS, root mean square; TG, tidal gauge.

In this table, we show the RMS deviation $\mathrm{RMS}(P,Q) = \sqrt{\sum_{i}^{n}(P_i - O_i)^2 / n}$, where P_i and O_i are the SSHs from the two datasets (the observed and the modeled SSH, respectively, or the model-1 and model-2 SSHs, respectively) at the positions of tidal gauges (TGs). In the above equation, n is the number of observations (the index is i).

The time versus the along-coast distance diagrams (Figures 3a and 3c) give a clear illustration of the propagation characteristics of the tidal waves. Starting from Lerwick and traveling up to Whitby, the coastal wave propagates with the coast on its right (Figure 2a), and the slope of the contours gives a measure of the wave propagation speed, ranging between several to several tenths of ms^{-1} depending on the local conditions (the average depth of the North Sea of ∼90 m would result in a propagation speed of ∼30 ms^{-1}). At around the Wash, the propagation pattern changes dramatically because, to the south, the small amphydrome in the Southern Bight (Figure 2a) wedges into the large amphydrome in the southern North Sea. This is the reason the contours undergo a rapid transition until the Terschelling Noordzee station. If we omit the data between Cromer and Terschelling Noordzee (and just linearly interpolate the data between them), the contours would present much smoother patterns. The tidal amplitudes decrease strongly around the Tregde station (see Figure 1 for its position) when passing from the southern to the northern amphidromic area. Figure 3c is the same as Figure 3a; however, the data come from AMM7. Visually, the model and observations agree quite well, and the quantitative comparison between them can be better estimated from Table 1.

The LF signal (the panels on the right, Figures 3b and 3d) shows temporal variability dominated by synoptic time scales (in the atmosphere). Because of the much longer time axis compared with the panels on the left-hand side of the figure, the slope of the contours looks rather small. Along the eastern and western coasts, the propagation direction is from north to south; the change in the slope of contours occurs in the Southern Bight. Again, the consistency between the data from the TGs and the AMM7 seems quite good; all major low and high sea-level events in the observation data set have their counterparts in the numerical simulations. The simulated amplitudes are slightly lower than the observed amplitudes, which is explained by the quality of the atmospheric forcing. The above comparison between observations and simulations (Table 1 and Figure 3) shows that both datasets are similar but far from identical. The difference between the two model datasets (AMM7 and GCOAST) is comparable to the difference between each of them and the observations (see Table 1). As we will show in the next sections, these comparisons are important to understand the results from the experiments using machine learning (ML).

2.2. Generative Adversarial Network

2.2.1. Brief Introduction

LeCun et al. (2015) defined deep learning as a method allowing "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction." In many applications, deep learning uses feedforward neural networks, which learn to map an input data set (in many examples, an image is used as input) to an output (e.g., more abstract information such as the probability of belonging to a certain category). Artificial neural networks are inspired by biological neural networks, which learn by considering examples. Their structure consists of connected units (nodes) called artificial neurons, which receive and transmit a signal to other neurons. In deep learning, multiple levels of

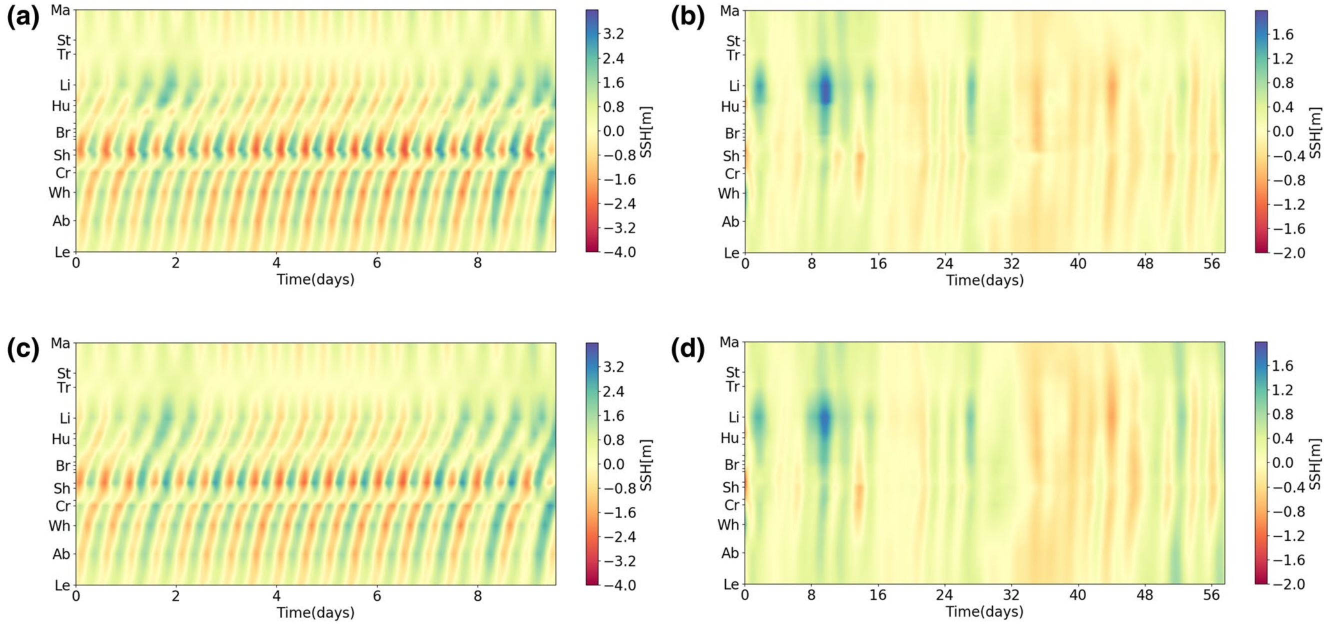

Figure 3. Time versus the along-coast distance (starting from the Lerwick station) diagram of the sea level from TGs (a) and AMM7 (c). The panels on the right, (b) and (d), show the same as (a) and (c) but for the LF signal and for longer periods. AMM7, Atlantic Margin Model-7; LF, low frequencies (low-pass filtered data set); TGs, tidal gauges.

information transformation from the previous layer to a higher layer (more abstract information) are used. Filters are applied to the input images to create feature maps that summarize the presence of those features in the input. The filter (e.g., a 3 × 3 matrix) is moved across the image. This movement, which is usually symmetrical in the x and y directions, is referred to as the stride. The default stride is (1, 1). A stride of (2, 2) would mean moving the filter two pixels in the horizontal and vertical directions. Thus, the neurons combine the input in such a way that the output is presented as a nonlinear combination of its inputs. A series of weights determine how the inputs are fed to the outputs. In many applications, the weight vectors are adjusted following the stochastic gradient descent algorithm.

Goodfellow et al. (2014) introduce a framework for estimating generative models via an adversarial process by training two models. The first model is a generative model. This model captures the data distribution. The second model is a discriminative model, and its role is to estimate the probability that a sample comes from the training data rather than the generative model. As a result, the generative model recovers the training data distribution.

Normally, the structure of a convolutional neural network consists of convolutional layers followed by pooling layers. The role of the latter is to reduce the amount of redundant information. The most commonly used method is the max-pooling method, which keeps only the most active neurons (out of every 2 × 2 square of neurons in the convolutional layers, the "max"). Experience shows that this pooling step does not reduce the performance of the network. In the U-Net architecture (Ronneberger et al., 2015), the pooling operations are replaced by upsampling operators. An expansive path is developed, which is more or less symmetric to the contracting part and yields a U-shaped network architecture (Figure 4). In the expansive path, in every other layer, the resolution of the output is increased. Thus, in the upsampling part of the network, information is propagated to higher-resolution layers. Two distinct models, a generator and a discriminator, constitute the GAN. The generator is trained adversarially by optimizing a minimax objective together with a discriminator. In the following, the specific application of the U-Net architecture to analyzing sea-level maps is described.

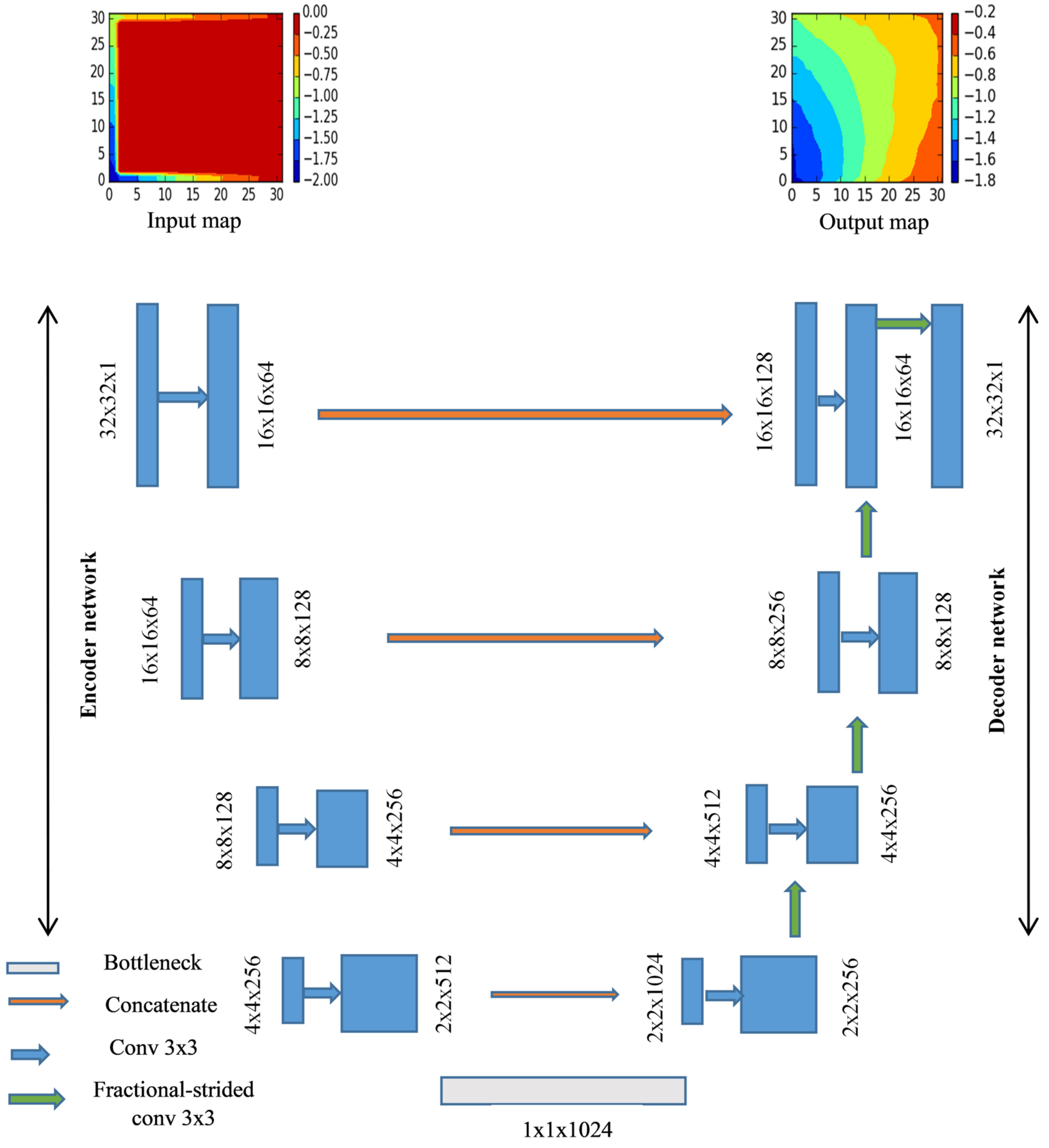

Figure 4. Schematic presentation of the generator part of the deep neural network model for sea surface height map reconstruction.

2.2.2. GAN for Tidal Reconstructions

2.2.2.1. Generator

The U-Net structure of the generator part of our deep neural network model for tidal reconstruction (Figure 4) is illustrated in the following using the 32 × 32 SSH hourly maps (rectangle in Figure 1) for 2016 from the AMM7. In the example considered here to train the model, we use only the SSH records along the sides of the rectangle as an input data set. This data set is named in Figure 4 as the "Input map." The target data set is the full AMM7 data set (see, e.g., "Output map" in Figure 4). The task of the generator is to provide a model of high-quality reconstructed SSH maps (as close as possible to the AMM7 maps) by using only the information at the boundary.

The U-Net convolutional neural network (Figure 4) consists of two parts: an encoder (on the left) and a decoder (on the right). The encoder transforms an image (map) into a compact latent feature representation.

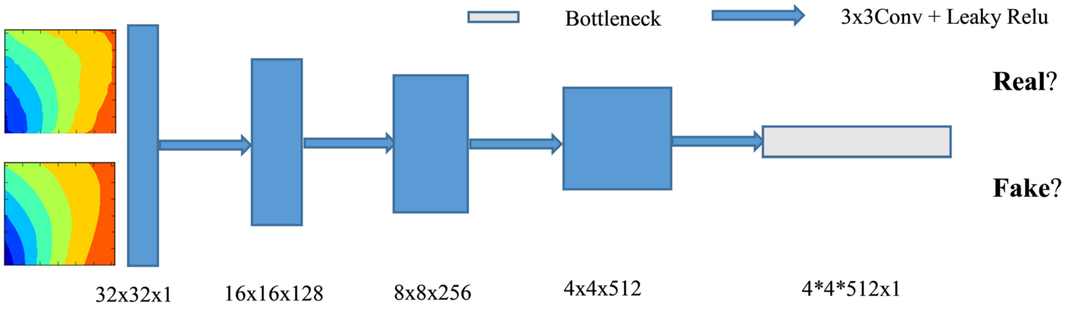

Figure 5. The discriminator part of the reconstruction model.

The decoder uses that representation to produce the missing image content. Thus, the encoding-decoding process learns the image features and generates full maps.

The encoder-decoder pipeline works as follows. The encoder takes an input image with missing data and produces a latent feature representation of that image. The decoder takes this feature representation and produces the missing image content. The encoder process consists of the repeated application of 3 × 3 convolutions with a stride of 2 for downsampling, each followed by a batch normalization layer (Ioffe & Szegedy, 2015) and Leaky logarithmic rectified linear unit (Leaky-L_ReLU, Maas et al., 2013) activation.

In the example shown in Figure 4, the first convolution layer is a 32 × 32 × 1 (width × length × depth) map. In each subsequent convolution calculation, we obtain more latent feature maps with a larger depth index and narrower width and length. The feature map represents higher-dimensional data distribution characteristics from the image. The bottleneck layer (Figure 4) represents the image fully compressed into a feature map with a depth of 1,024.

Decoding is the opposite of encoding; we call this process deconvolution. It consists of repeated applications of 3 × 3 convolutions with a stride of 2 using a transposed operator (also called a transposed convolution or fractionally strided convolution); that is, it performs a deconvolution. The upsampling layers in the original U-Net structure (Zador, 2019) are replaced with fractionally strided convolutional layers in our U-Net-like structure.

In image completion problems, corrupted images and output images share a certain amount of low-level features, such as prominent information from the noncorrupted regions, luminance, and resolution. However, deep network-based methods with bottleneck layers may lose details of images when propagating feature maps in the training stage. Moreover, these methods may suffer from the vanishing gradient problem as the network deepens. To shuttle the image information through the networks and reduce the training burden, we apply the skip connections strategy (Mao et al., 2016).

2.2.2.2. Discriminator

The discriminator (Figure 5) is used to determine the possibility that the prediction map comes from the training set (i.e., whether it is a real training image) or the prediction set (i.e., whether it is a fake image from the generator). During training, better fake images are generated, and the role of the discriminator is to correctly classify the real and fake images. When the generated prediction map is consistent with the ground truth of the image content (we will call this the target for short) and the GAN discriminator cannot determine whether the prediction map is from the training set or the prediction set, then the network model parameters are considered to have reached the optimal state.

The discriminator can be understood as the inverse of the generator with five 3 × 3 convolutions with a stride of two layers, where the last convolutional layer is fed into a single sigmoid activation function. The L_ReLU activation function is used for all the layers in the discriminator except for the output. Following GAN technology, the generator is trained adversarially against a discriminator, which is simultaneously trained with the generator.

2.2.2.3. Technical Details

We use 8,640 SSH maps with a size of 32 × 32 to train the GAN model. This data set is too large to be passed to the computer at once; therefore, we divide the data into smaller sizes. Two hyperparameters are defined: the number of training epochs (how many times we train the model) and the batch size (the number of samples used to train the model in one epoch). During the training stage, mini-batch learning is introduced (Cotter et al., 2011). This method divides the data into several small batches and updates the parameters in batches. Thus, a set of data in a batch determines the direction of the gradient, which is more stable and converges faster. In the present tidal application, 12 is selected as the batch size, which corresponds to the period of the M2 tide in the studied region (see also Riley, 2019). The generator model uses only the observations in tidal gauge locations to generate the entire 2D SSH maps. This generated SSH map is fed into the discriminator together with the SSH map from the numerical model (true data set) to check whether the generated SSH resembles the true SSH (see Annex 1).

The model epoch parameter, which is a number optimizing the gradient decent (to avoid overfitting or underfitting), is set to 60. The initial learning rate of the Adam optimizer of the GAN model is set to 0.00003. The learning rate determines the step size of gradient descent (i.e., how fast the model converges). Too large a rate may cause the parameters to move back and forth on both sides of the optimal value. Too small a rate will greatly reduce the optimization speed. To solve this problem, we introduce an exponential decay method from the TensorFlow framework (Loshchilov & Hutter, 2019). The learning rate was gradually reduced to make the model more stable in the later stages of training. The number of convolution levels is set to 9 (later in the text, we explain why this number must be changed in other experiments).

In this model, the discriminator loss is the same as the basic deep convolutional GAN (DCGAN) model (Radford et al., 2015), while for the generator loss, we introduce (on the basis of original generative loss) the least square errors (known as the L2 loss function) as the consistency loss into the generative loss function. The basic loss function of the GAN model meets the Nash equilibrium condition as much as possible (Osborne & Rubinstein, 1994). The principle behind this equilibrium is based on game theory and aims at continuously optimizing the generator and discriminator so that the generated data approach the real data (Dong & Yang, 2019). In this way, the GAN makes the samples generated by the generator approach the real sample in terms of both authenticity and diversity. To adapt the technology to our specific study and obtain more accurate results, we also established a new loss function combination for our pixel-wise GAN model by adding the pixel-wise reconstruction loss generation part based on the basic loss function (Zhao et al., 2017). The equations describing the loss functions of the generator and discriminator are given in Annex 1.

After 60 training epochs, we obtain a suitable generator structure that remembers the SSH high-dimensional features of the selected ocean region. To validate the generator model, the discriminator part is dismissed since the generator part is the main structure for reconstructing the completed SSH maps. Now, the validation data set (2,158 hourly, incomplete SSH maps) is fed into the generator part of the neural network model to generate feasible SSH maps.

2.2.3. GAN for LF SSH Reconstruction

As shown in Section 2.1.1, the amplitude of the remaining signal is lower than that of the dominant partial tides. Furthermore, the variability is less regular because meteorological drivers such as wind, storms, or atmospheric pressure have a certain level of randomness. Therefore, we introduce some changes in the GAN model described in Section 2.2.2 for the sake of obtaining more accurate reconstruction results (see Zador, 2019).

Our new model for LF SSH reconstruction consists of two steps: coarse and fine reconstruction (Figure 6). This architecture helps to stabilize training and enlarge the receptive fields, as mentioned by Yu et al. (2018). Figure 6 represents a 32 × 32 × 1 map with real data only at the boundary, and O_c represents the coarse-reconstructed SSH map. The refinement step takes the O_c and *In* maps together as input pairs to output the final result *Out*. Thus, O_c conditioned on *In* is selected as the input of the refined network that reconstructs the complete SSH map called *Out*. This type of input stacks information of the known areas to urge the network to capture valid features faster (Liu et al., 2019), which is critical for rebuilding the content of missing regions. Our refined structure also consists of an encoder and decoder, where a skip

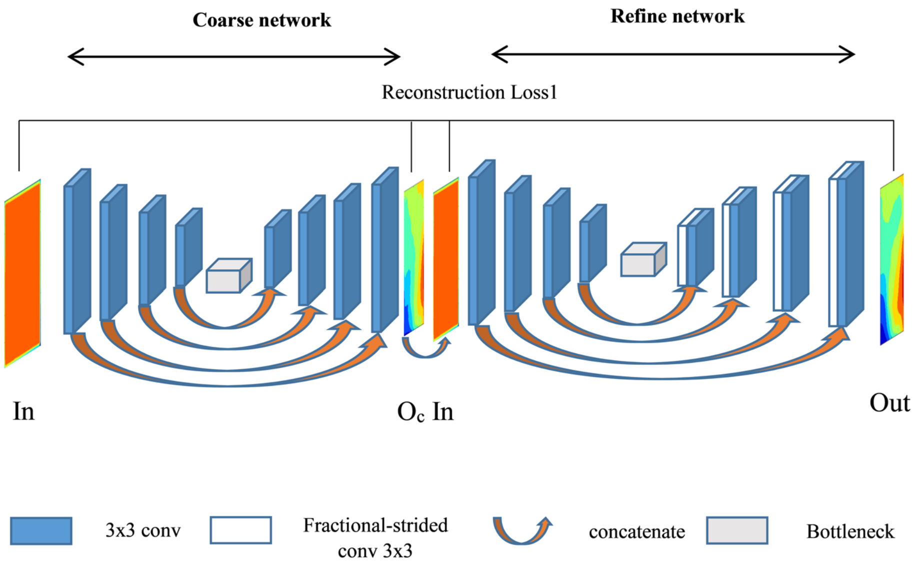

Figure 6. Generator structure for residual sea surface height map reconstruction.

connection is adopted, similar to a coarse network. In the encoder, each of the layers is composed of a 3 × 3 convolution, while in the decoder, a fractional stride convolutional layer with stride of 2 is adopted together with a 3 × 3 convolution. Finally, the discriminator has the same structure as the discriminator in our tidal reconstruction model.

The training step and reconstruction process are the same as in the tidal reconstruction model. However, the training sample batch size is set to 336 (2 weeks of hourly data). The number of training epochs is set to 5,000 since more uncertainty and greater magnitudes of the variations lead to training difficulties, which will require more training epochs to obtain a stable and desirable model. The initial learning rate of the Adam optimizer in this GAN model is set to 0.00007 to adapt to these model parameter changes.

2.3. Kalman Filter Approach

Schulz-Stellenfleth and Stanev (2010) proposed an optimal linear estimator to reconstruct ocean state parameters from observations knowing the prior distribution of the state and measurement errors. The method is similar to the approach of Frolov et al. (2008) and uses standard concepts of estimation theory. It will be very briefly presented below; for more detail, the interested reader is referred to Schulz-Stellenfleth and Stanev (2010) and Grayek et al. (2011). The method uses the background covariance matrix derived from the AMM7 data as a priori information. We will denote the global state vector of dimension m by $\boldsymbol{x}$. The data from 19 tidal gauges represent the measurement vector $\boldsymbol{y}$ of dimension n. The global state vector $\boldsymbol{x}$ contains SSH from AMM7 data at the individual position of the model area. The task is to find a reconstruction matrix A such that

$$J(A) = \sum_{j=1}^{q} \left\| x(t_j) - Ay(t_j) \right\|^2$$

is minimum, where q is the number of SSH maps (hourly maps in 1 year). This would ensure that the reconstruction error is as small as possible. Assume that the observations can be derived from the global states according to

$$\boldsymbol{y} = H\boldsymbol{x}$$

where H is the linear measurement operator. Schulz-Stellenfleth and Stanev (2010) showed that $J(A)$ is minimum if A is the Kalman gain matrix

Table 2
List of Experiments in a Small Domain

Name	Type of experiment (AF)	Type of experiment (LF)	Training (input data-target data)	Validation (input data-validation data)	Comment
AF1	X	-	AMM7-AMM7	AMM7-AMM7	Randomly distributed no-data locations.
AF2	X	-	-	-	At the validation step, there are missing data only in the interior.
AF3	X	-	-	-	At the validation step, data are available only at the boundary.
AFK	X	-	-	-	-
LF	-	X	-	-	-
LFK	-	X	-	-	-

Note. Index "K" in the experimental nomenclature stays for the "Kalman filter approach."
Abbreviations: AF, all frequencies (full data set); AMM7, Atlantic Margin Model-7; LF, low frequencies (low-pass filtered data set).

$$A = PH^T\left(HPH^T + R^{-1}\right)$$

where P is the background covariance matrix for the state $\boldsymbol{x}$ and R is the observation error.

It was demonstrated in the same study that if the dynamics of the state variables can be described by only a few empirical othogonal functions (EOFs), the dimension of the reconstruction problem can be significantly reduced. For the AMM7 SSH data set, only three EOFs describe more than 95% of the variance. Therefore, in the analyses addressed in the following, we used three EOFs only. R is taken as a diagonal matrix, with a constant error value of 1 cm.

3. Experiments

3.1. Experiments in Reduced Area

The first group of experiments presented below aims to analyze how appropriate the GAN is to reconstruct the sea level in a relatively small area (only 32 × 32 grids) in the interior of the North Sea. In the experiments' nomenclature (Table 2), we use the abbreviations AF and LF. In the AF experiments, we use the 1-h data, as they are produced by the AMM7 model. In the LF experiments, we also use 1-h data, but the variability with periods higher than 2 days is removed by low-pass filtering as explained above. Therefore, in the LF experiments, the sea level can be considered mainly driven by the atmosphere and by low-frequency tides (e.g., spring-neap variability). The training phase uses 8,642 hourly maps, which corresponds to 1 year. In the first type of experiment, for which there is a column called "Training" in Table 2, we use data from the same source in the training and validation steps. In this way, the data at the two steps are consistent with each other.

In AF1–3, we determine the quality of reconstruction if some input data are missing. In AF1, we randomly generate locations in the 32 × 32 matrix, where we assume that there are no available data (there are no "observations" in half of the locations in the original grid). Thus, we feed the GAN model with data only from the locations where there are "observations." These locations differ in the different experiments presented in Table 1. The input from the remaining grid points belonging to the model area is specified as zero. We use the data from all the locations as a target data set. At the validation step, we use the "observations" in only half of the locations of the original grid to reconstruct the 32 × 32 field (all locations) over a period of 3 months. A comparison between the AMM7 data and the reconstructed data will be analyzed in Section 4.

In AF2, we select a square area in the middle of the 32 × 32 matrix ($i = 5,..., 25$ and $j = 5,..., 25$), which is considered a no-data area. The basic difference from AF1 is that this no-data area is compact. In AF3, we extend the no-data area up to the boundary. This exercise can thus be interpreted as a reconstruction of the full data

Table 3
List of Experiments in the Entire North Sea

Name	Type of experiment (AF)	Type of experiment (LF)	Area	Training (input data-target data)	Validation (input data-validation data)	Comment
BAF1	X	-	Entire North Sea	-	-	Similar to the comment for AF (see Table 2).
BLF1	-	X	-	-	-	Similar to the comment for LF (see Table 2).
BAF2	X	-	-	-	TG-AMM7	Similar to the comment for BAF1 for the observations used at the validation step.
BLF2	-	X	-	-	-	Similar to the comment for BLF1 observations used at the validation step.
BAF2-G	X	-	-	-	GCOAST-AMM7	Similar to the comment for BAF2 GCOAST data used at the observation locations at the validation step.
BLF2-G	-	X	-	-	-	Similar to the comment for BLF2 GCOAST data used at the observation locations at the validation step.
BAF3	X	-	-	TG-AMM7	TG-AMM7	The data at the boundary are the same in the training and validation steps.
BLF3	-	X	-	-	-	-
BAF3-G	X	-	-	GCOAST-AMM7	GCOAST-AMM7	-
BLF3-G	-	X	-	-	-	-

Notes. Notice that all names of experiments in in this table start with "B." The fourth column makes explicit the difference from Table 2.
Abbreviations: AF, all frequencies (full data set); AMM7, Atlantic Margin Model-7; B, basin-wide; GCOAST, Geesthacht COAstal model SysTem; LF, low frequencies (low-pass filtered data set); TG, tidal gauge.

set using data only at the boundaries (all the boundary locations). In AFK ("K" stays for "Kalman"), we use the Kalman filter approach described in Section 2.3 to reconstruct the SSH following the scenario of AF3.

Experiment LF is essentially the same as AF3; that is, only data at the boundary are used to train the model. The difference is that LF analyses the capability of using a GAN to reconstruct the data set from which the high-frequency tides have been removed. LFK is the same as LF; however, the reconstruction method uses the Kalman filter approach. In all the experiments described above, the computational resources were relatively low. On one GPU node, which is an Nvidia Tesla V100 with 32 GB memory, it takes ~30 min for GAN to complete the training in the AF experiments and ~45 min to complete the training in the LF experiments. The latter takes a longer time than the former because the more stochastic signals associated with the atmospheric forcing compared to the periodic tidal signals make the convergence slower.

3.2. Experiments in the Entire North Sea

The second group of experiments is for the entire North Sea basin (index "B" in Table 3). Experiment BAF1 is essentially the same as AF3. The difference is that in the training phase, we use only data from 19 locations where TGs operate. These data in BAF1 are taken from AMM7 from the nearest to the observation location model grid points. As in AF3, the input from the remaining grid points belonging to the model area is specified as zero. The training and validation periods are the same as those in the experiments with reduced area. BLF1 is essentially the same as BAF1; the difference is that we analyze the quality of the reconstruction of the low-frequency North Sea data set, that is, the data used in this experiment are the low-pass filtered data used in BAF1. The idea to carry out this experiment was two-fold. It was assumed that removing the high-frequency oscillations would result in a better model when reconstructing the low-frequency variability of basin-wide SSH using only coastal data. The second consideration was that the high-frequency oscillations are not included in some altimeter products; thus, it is worth trying to test whether ML can well resolve only the low-frequency variability.

BAF2 is the same as BAF1; however, real observations from TGs are used in the validation step along with the same model developed in BAF1. Obviously, this experiment uses data of different origins (in the "Validation" column of Table 3, the data sources are different). Thus, these data are not fully consistent with each other. In the following, we will refer to this type of experiment as experiments with "inconsistent data." BLF2 is the same as BAF2, but BLF2 addresses the quality of the reconstruction of low-frequency variability. One important difference between the BAF and BLF experiments is that we use different ML models (see Section 2) because of the different spatiotemporal characteristics of the tidally and atmospherically driven sea level. Practically, in the BAF2 and BLF2 experiments, we assign the TG observations to the nearest model grid neighbors.

In BAF-G and BLF2-G, we do not use real observations as in BAF2 and BLF2 but rather data from the GCOAST model in the nearest to the observation locations model grid points.

The next two experiments, BAF3 and BLF3, use partially inconsistent data for training and validation. By "partially inconsistent," we mean the following. At the training step, at the positions of the TGs, we use data from the TGs. The target data set is the basin-wide SSH, which is produced using AMM7 (TG-AMM7 in the "Training" column of Table 3). It is expected that the ML model learns the consistency between the forcing data and the target. Therefore, the product is partially consistent with the data from the TGs and the AMM7. At the validation step, we use tidal gauge data and the ML model to reconstruct the SSHs and compare them to the AMM7 data. BLF3 is the same as BAF3, but BLF3 addresses the quality of the low-frequency reconstructions.

In the final two experiments (BAF3-G and BLF3-G), we used the same approach as in BAF3 and BLF3. In this case, in the observation locations (Figure 1), the data are sampled from the GCOAST model, which has a horizontal resolution two times better than that of the AMM7. These data are considered pseudo-observations with "different quality" than the quality of the coarser AMM7.

4. Results

4.1. Sea-Level Reconstruction in Idealized (Reduced) Areas

The results of all the reduced area experiments are presented in Figure 7. As representative characteristics measuring the agreement between the reconstruction data and the observations, we use the index of agreement (Willmott, 1981):

$$D(P,Q) = 1 - \sum_{i=1}^{n}\left(P_i - O_i\right)^2 / \sum_{i=1}^{n}\left(\mid P_i - \bar{O} \mid + \mid O_i - \bar{O} \mid\right)^2 \tag{1}$$

In the above equation, the overbar indicates the temporal mean, and the other notations are explained in Section 2.1.3. The above equation provides a statistical approach to compare model predictions (P) with observations (O). The numerator measures the average error magnitude, and the denominator gives a basis of comparison. The index of agreement measures the model performance as the degree to which P matches O, where 1 indicates perfect agreement and 0 indicates complete disagreement. Other possible indexes for model-data comparison are defined by Nash and Sutcliffe (1970), Legates and McCabe (1999); see also Willmott et al. (2012).

The results of the three AF experiments (AF1–AF3) are shown in Figures 7a–7c. They illustrate how the reconstruction results deteriorate if some input data are missing. However, "deterioration" is not an adequate word in this case because all three reconstructions are characterized by an index of agreement greater than 0.99. The pattern of D reflects some characteristics of the data distribution and dynamics. In AF1, the no-data locations are randomly distributed. However, the results in Figure 7a show that the lowest values of the index of agreement appear predominantly at the boundary. This finding is explained by the fact that,

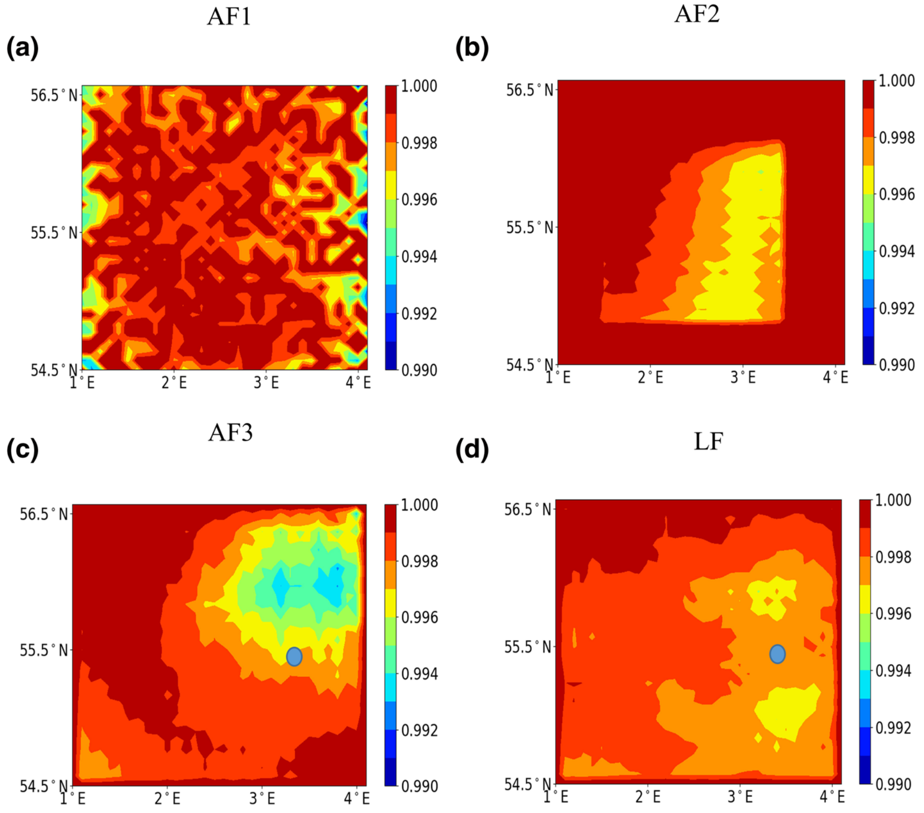

Figure 7. Index of agreement between the "true" and reconstructed SSHs in the experiments carried out in the reduced areas. The index was computed at the validation step. The numbers on the axes are the longitude and latitude (see Figure 1 for the position of this area). The blue dots are locations where time series are analyzed for the validation period. AF, all frequencies (full data set); LF, low frequencies (low-pass filtered data set); SSH, sea surface height.

in the basin interior, the no-data locations are uniformly surrounded by locations where observations are available. However, at the periphery of the studied area, the no-data locations are surrounded by fewer observations (because no observations exist outside of the area).

In AF2, where we prescribe a wide coastal area with data, the index of agreement is higher than 0.995. In the no-data area (in the middle of Figure 7b), the index of agreement shows a propagation pattern, which is in agreement with the propagation direction of the Kelvin wave (see the rectangle in Figure 1 and the phase lines in Figure 2a). The situation in AF3 (Figure 7c) is qualitatively similar to that in AF2 (a better agreement with the validation data set in the western part). However, in this experiment, only the data along the boundary are used; therefore, the index of agreement is slightly lower, and its pattern is less regular than that in AF2. The lowest D~0.994 in AF3 appears in the area closest to the amphydromic point (Figure 2a), where the amplitude of the signal is lower; therefore, the signal-to-noise level is also lower.

Experiment LF (Figure 7d), which quantifies the capability of GAN to reconstruct the SSH using the LF data set, shows a comparable skill as AF1–AF3. In all four cases, the index of agreement is above 0.99, and its ranges are comparable. The fundamental difference between the AF and LF experiments is the pattern of the index of agreement, which is no longer tidally dominant in the LF case.

The comparison between the performance of GAN and Kalman filter approach is presented in Figure 8 for one location shown in Figure 7 where the reconstruction quality of GAN is relatively low (D~0.997). Obviously, the two methods perform very similarly. For this specific location, the RMS differences between the AF reconstructions and AMM7 data are ~3 cm. The RMS differences between the LF experiments and low-pass filtered SSH are ~1 cm. These values, as seen in Figure 8, are negligibly smaller than the amplitude of the respective signals. It is clearly seen from these illustrations, in particular in the plots on the bottom,

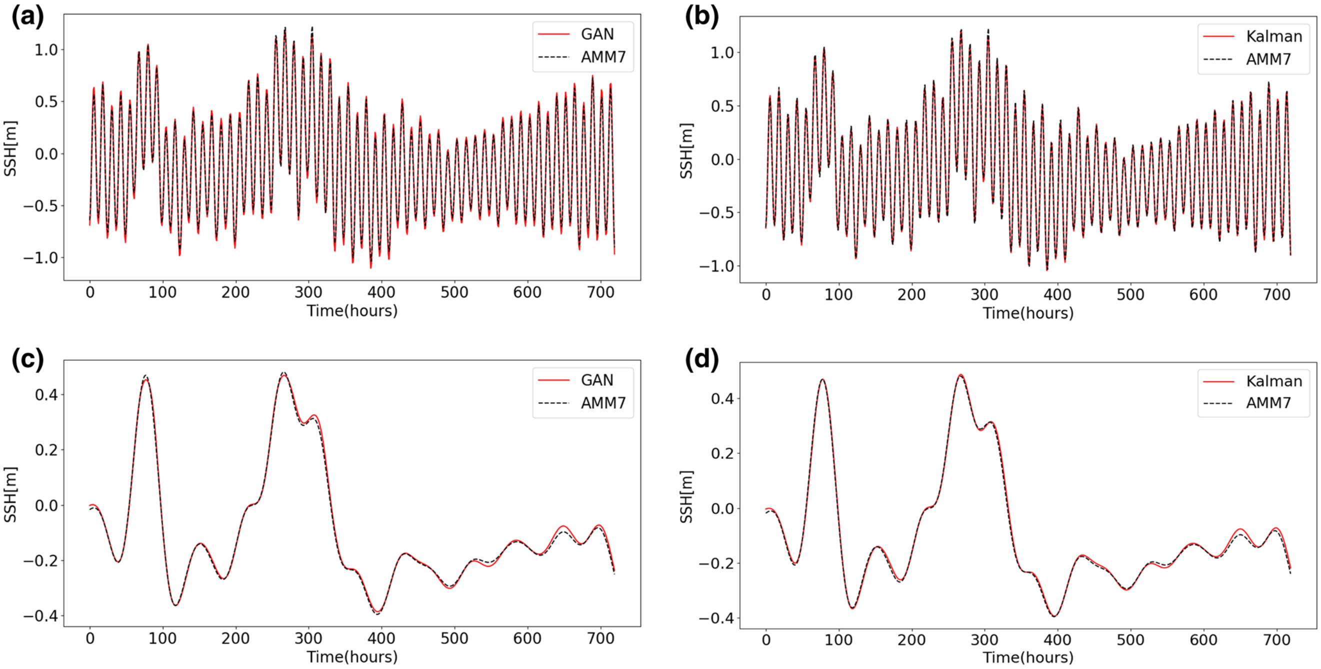

Figure 8. Sea level in the locations shown in Figure 7 for the 2-month validation period. (a and b) are AF experiments; (c and d) are LF experiments. Panels on the left are from GAN reconstructions; those on the right are from the reconstruction using the Kalman filter approach. AF, all frequencies (full data set); AMM7, Atlantic Margin Model-7; GAN, generative adversarial network; LF, low frequencies (low-pass filtered data set).

that the largest differences between the reconstructions using GAN and Kalman filter methods, from one side, and data, from the other, occur almost at the same times.

The above experiments are relatively easy, at least in terms of the volume of data used. In real-world applications, data sets are usually much larger, and it was not clear a priori whether the used method would have the same performance if larger datasets were used. By increasing the data volume by ~25 times, one reaches the volume of the data set generated from AMM7 over the entire North Sea. Therefore, we performed a preparatory experiment in which we interpolated the 32 × 32 matrices with a resolution five times better than in the AF experiments and repeated AF3 experiments using the new data set. Because of the increase in the data size, we increase the number of convolutional layers to 13; the number of epochs is set to 600. The initial learning rate is the same as in the case of the 32 × 32 data set. The computational time for training increased up to ~6 h, which is approximately 12 times longer compared to the 32 × 32 cases. The index of agreement in this additional experiment (not shown here) is higher than 0.993, and its pattern is close to that in AF3.

4.2. Sea-Level Reconstruction Over the Entire North Sea

Here, we discuss the skill of experiments introduced in Section 3.2. As a measure of the skill of each of them, we will show maps of the index of agreement (Figure 9). The RMS difference between the reconstructed and "true" data is shown in the supporting material (Figure S1). The results from experiment BAF1 (Figure 9a) demonstrate that using the data from only 19 locations where the TGs operate is sufficient for adequate sea-level reconstruction over most of the analyzed domain. Only in the northwestern part of the study area and in Kattegat, where TGs are not available, and in the area between the two amphidromic points in the eastern North Sea, the index of agreement drops to ~0.8. The smaller D in the area between the two amphidromic points is explained by the low-amplitude tides in this zone (small denominator in Equation 1).

The reconstruction of the LF variability (experiment BLF1, see Figure 9b) is approximately as good as the reconstruction of the full signal. The lower index of agreement in the interior of the North Sea is explained

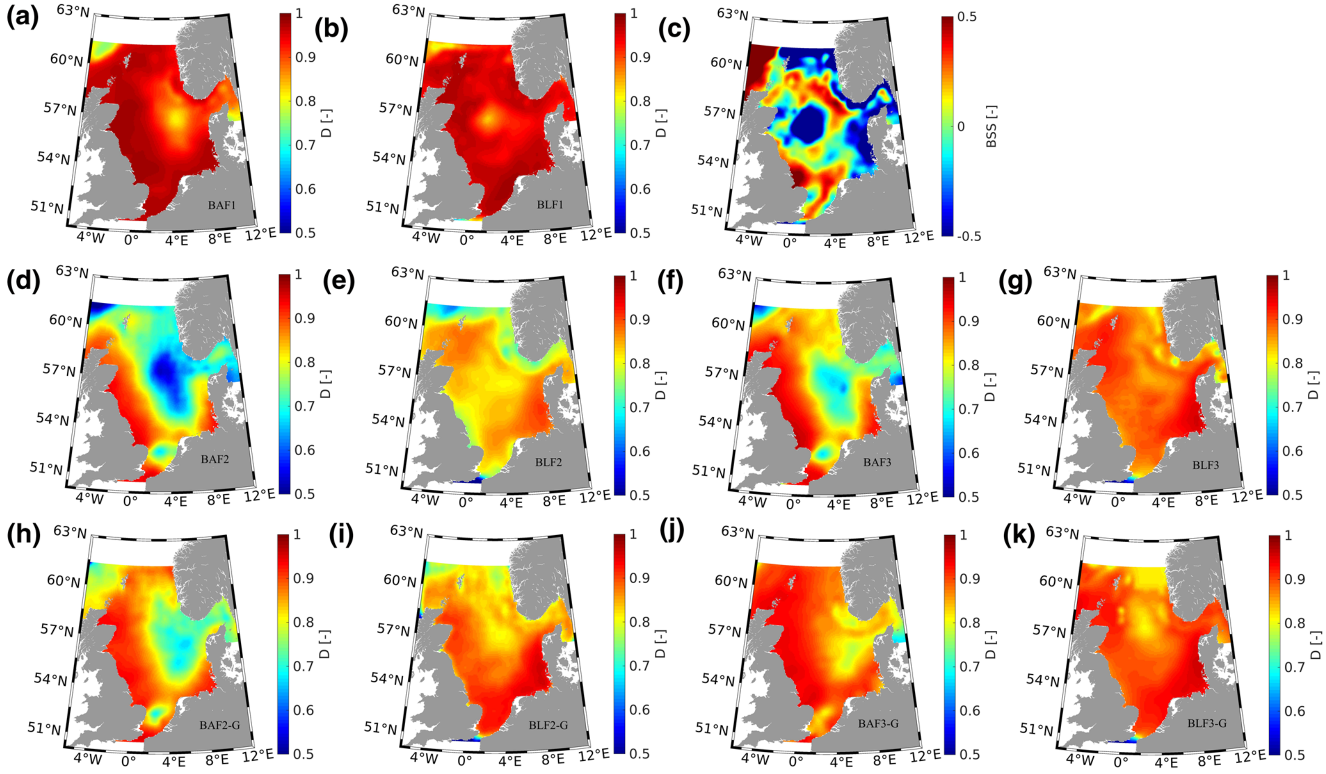

Figure 9. Index of agreement (see Equation 1) between the "true" and reconstructed SSHs in the experiments carried out over the entire North Sea (see Table 3). The training data set is from January 01, 2016 to January 01, 2017, and the validation data set is from January 01, 2017 to March 01, 2017. The names of the individual experiments are shown in each panel. The ML model BAF1 is used in all the experiments shown in the first column (a, d, and h). The panels in the second column (b, e, and i) use the BLF1 model. Panels f and j and panels g and k use the BAF3 and BLF3 models, respectively. (c) shows the Brier skill score (see Equation 2) of BLF1 against the low-pass output of BAF1. ML, machine learning.

by the relatively low amplitude of the signal there (compare with Figure 2c). Because of this phenomenon, the signal-to-noise ratio reduces the reconstruction skill. Overall, BAF1 and BLF1 demonstrate that if *consistent datasets* are used in the locations where TGs are located, the GAN model adequately reconstructs the basin-wide SSH.

We remind the reader here that the SSH reconstructed in BAF1 contains low- and high-frequency signals, while the output of the BLF1 experiment reproduces only the low-frequency variability. The initial expectation was that the GAN could better learn less complicated temporal and spatial variability, which is in the case when a high-frequency signal was removed from the data. Figure 9c shows the Brier skill score

$$\mathrm{BSS} = 1 - \mathrm{BS}_{\mathrm{BLF1}} \,/\, \mathrm{BS}_{\mathrm{BAF1}} \tag{2}$$

where $\mathrm{BS}(P,Q) = 1/n\sum_{i=1}^{n}\left(P_i - O_i\right)^2$ is the mean-squared error in each experiment and BAF1 is taken as the reference experiment. The reconstruction is perfect when the BSS is equal to 1. BSS = 0 means that there is no improvement in BLF1 compared to the results in BAF1. If BSS<0, the quality of BLF1 reconstruction is poorer than that in BAF1. Obviously, there are areas where BAF1 shows better agreement with the observations than BLF1. This result suggests that processing a much more complex data set (BAF1) is superior in many areas than processing low-pass filtered data, which demonstrates that the GAN can learn about processes with multiple time scales. As demonstrated by Jacob and Stanev (2017), in the North Sea, processes with multiple time scales in the ranges studied here are nonlinearly coupled. The fact that the reconstruction of the full signal is superior in many areas compared with the reconstruction of the LF signal provides indirect proof that the nonlinear interactions between processes with different time scales are well captured by ML. These patterns would not occur if nonlinear interactions between processes with different time scales did not exist. Figure S2 gives an illustration how M4 tides, which are due to nonlinear advection, are replicated. This result emphasizes the performance of ML method in the coastal regions.

In the BAF2 and BLF2 experiments, the reconstruction models are the same as in BAF1 and BLF1, respectively. However, unlike the BAF1 and BLF1 experiments, where the training and validation steps use consistent data, the BAF2 and BLF2 experiments belong to the class of experiments using inconsistent datasets at the validation step; that is, instead of using AMM7 data at the coast (consistent with the training data), we feed the model with real observations (which are inconsistent with the model). The level of inconsistency is quantified in Section 3 (see Figure 3 and Table 1). This substitution of data at the validation step resulted in a reduction in the reconstruction quality. What the GAN model can only adequately capture (index of agreement above 0.85) is the sea-level variability in the coastal areas of the western and southern North Sea (Figure 9d). The areas of low sea-level variability show a very low reconstruction skill (compare with Figure 2b, particularly the region of the small amphydrome in the Southern Bight). The reconstruction of the LF-signal is slightly better, particularly in the coastal zone of the German Bight, where the variability range of the LF-signal is the strongest (compare Figure 9e with Figure 2c). Obviously, the GAN model is not very flexible in using arbitrary types of data at the reconstruction step. The reduction in the reconstruction skill reminds us of the problems in data assimilation when errors in the data and model are not treated appropriately.

The BAF2-G and BLF2-G experiments, similar to the BAF2 and BLF2 experiments, belong to the group of experiments using inconsistent data sets at the validation step. In this case, the ML models are the same as in BAF1 and BLF1; however, the GCOAST data in the locations of TGs are used at the validation step. As shown in Section 2, the GCOAST data are slightly more consistent with the AMM7 data than with the TG data (Figure 3, Table 1). Therefore, the reconstruction skill improved in comparison to that in BAF2 and BLF2 (compare Figures 9h and 9i with Figures 9d and 9e, respectively). However, D is much lower than in the BAF1 and BLF1 experiments.

The "partial inconsistency" of the data for training and validation in BAF3 and BLF3 implies that, at the training step, the processing of the GCOAST data (at the coast) and the AMM7 data (as the target) tends to decrease the inconsistency between the two datasets in the GAN model. Thus, the results of new models, which are different from BAF1 and BLF1, respectively, are documented by the comparison between Figures 9f and 9g and Figures 9d and 9e, respectively. Obviously, the GAN model learns to adapt the solution to the data from different origins. The model skill is also dependent on the magnitude of the sea-level variability (compare with Figure 2b), which also explains the relatively good skill of BLF3 in the coastal zone of the German Bight. Notable is the more uniform and better reconstruction skill in the LF experiment than in the AF experiment over most of the area.

The final two experiments, also belonging to the group of experiments with "partially inconsistent" data (BAF3-G and BLF3-G), illustrate the improvement of the reconstruction skill if synthetic observations (the GCOAST data in the TG locations) are used in the training (compare Figures 9j and 9k with Figures 9h and 9i, respectively). The better agreement between the GCOAST and AMM7 data than the agreement of each of them with the real observations (see Table 1) explains why the index of agreement in Figures 9j and 9k are better than those in Figures 9f and 9g, respectively.

5. Discussion

One year of training data appeared sufficient for a model using a generative adversarial network to learn the structure of the SSH data and to adequately reconstruct the basin-wide SSH during the validation period using data from only 19 locations along the coast. The quality of the reconstructions was almost equally good for the full signal (AF) and low frequency one (LF). The high values of the index of agreement (area mean values of 0.937 and 0.945 for the BAF1 and BLF1 experiments, respectively) were possible provided data from the same source was used (in this case, the AMM7). Another reason for the reconstruction success is the relative smoothness of the SSH maps.

The use of the same model fed from either the observations (TGs) or independent numerical simulations (GCOAST data) reduced the quality of the reconstructions: 0.761 and 0.822 in BAF2 and BLF2, respectively, and 0.823 and 0.860 in BAF2-G and BLF2-G, respectively. The comparable numbers are explained by the comparable differences between the three data sets fed to the model. In the BAF experiments, the agreement between the reconstructions and validation data depends strongly on the patterns of the tidal amplitude: the lower the amplitude (in amphidromic points), the lower the agreement is. The BLF experiments show a much more uniform distribution of the index of agreement. In both the BAF and BLF experiments, the reconstruction model performs better if it is fed with the GCOAST data than when it is fed with actual observations.

Some drawbacks in the reconstruction could be avoided if data from the coastal stations are used in the training. In BAF3, BLF3, BAF3-G, and BLF3-G, the basin mean indexes of agreement are 0.823, 0.879, 0.882, and 0.889, respectively. Obviously, there are good perspectives by developing an optimal learning process to improve the reconstruction quality, which should include the study of the individual imprint of stations for the reconstruction of the basin-wide sea level. TGs are sometimes placed in locations that are not representative of the large-scale dynamics; therefore, the observed signal is not fully consistent with the basin-wide dynamics. Such stations would have low imprints but could also contaminate the learning process.

A further adjustment of the loss function or other parameters would improve the reconstruction quality, which is another technical task to solve in future research. This issue has not been addressed in the present study because our aim was to demonstrate the power of the GAN model in reconstructing SSH maps by using different types of inputs and targets.

Another issue that has not been discussed in the present study is the length of the data series that we use. As is well known, neural networks can reconstruct situations similar to those they encounter from the past. Therefore, another way to improve the reconstruction quality would be to extend the duration of the learning process and perhaps to set a clearer aim to the reconstruction exercise with respect to the time scales addressed.

One important question to discuss here is what we learn about physics. One zero-order answer would be “nothing” because what we see in the validation step is a synthesis of situations from the past. However, the basic message from Figures 9a and 9b is that a decent reconstruction capability is realized using a relatively short time series, which is an illustration of a substantial recurrence of patterns. This would imply that the spatial-temporal patterns repeat (quasi)periodically, and a relatively short-time record contains the most representative characteristics of SSH dynamics. While this was clear for the tides, it was not so obvious about changes in sea level caused by the atmosphere. Furthermore, the GAN has a good skill to learn and reconstruct dynamics with multiple time scales. The results presented in Figure 7a illustrate that the reconstruction capabilities of the model decrease when approaching the boundaries of the area addressed. This finding justifies that having data at the boundaries is an important prerequisite for optimal reconstructions. In other words, much information on the dynamics of the entire basin is encapsulated in the boundary data, which enables the good reconstruction skill of the specific GAN application.

The patterns in Figures 7 and 9 demonstrate that the errors in the reconstructions are closely linked to specific physical patterns; that is, one can also make useful analyses of the model errors to study the physical properties of the sea level. This analysis would be important when developing concepts to specify error covariance matrices in data assimilation models. Another aspect concerns the role of coasts, which constrain the circulation features in different ways. One example is the low index of agreement in the area of the Norwegian trench, which is a known challenge for numerical models. Evaluation of different types of nonlinearities and identification and quantification of the responsible processes with the help of ML is another issue of future development. Our preliminary analyses of other types of data (e.g., sea surface temperature and sea surface salinity in the German Bight) show that in some cases Kalman filter approach performs better, in some other cases, for example, reconstruction of sea surface salinity, it is the ML approach, which performs better. The studied here SSH maps is just one type of data with their respective temporal and spatial scales. They cannot be considered as a comprehensive set of different types of data with different spatial and temporal characteristics, as well as different level of stochasticity. Therefore, the experiments presented here do not allow to fully analyze the advantages and disadvantages of two methods. A deeper analysis of the performance of the ML and Kalman filter approaches when using different and more challenging data sets will be presented in a forthcoming study.

Longer periods are beyond the scope of the present research. Reconstructing basin-wide SSH over long times would require a different design of deep learning. One can expect that reducing the resolution of basin-wide data used in the training, both in time and space, would allow more efficient computations and extension of the addressed time scales to decadal and beyond. One fundamental issue to address is whether coarser resolution in space and time ML would ensure adequate decadal reconstructions. Such an exercise will be analyzed in a separate study using different data sampling and processing technologies.

Another natural extension of the present research would be the application of a GAN to data-only cases. One candidate is the amalgamation between satellite altimetry and TGs, which would open up the perspectives to improve and optimally use the observational networks in the North Sea.

6. Conclusions

The method proposed here to reconstruct the basin-wide SSH using TG data from a few coastal stations builds on the capability of GANs to detect and reproduce nonlinear dynamics, as well as learning the dominant relationships of different spatial and temporal signals. We presented the method in detail, motivating interested scientists to apply it to similar natural settings or other oceanographic datasets. In the case when the coastal and open ocean data are consistent (e.g., they are from the same source), as was the case in experiments BAF1 and BLF1, only 19 stations in the locations of the permanently operating TGs are enough for the GAN to ensure an adequate reconstruction. The relatively short time series, which is only 1 year, provides an illustration of a substantial recurrence of events. It was demonstrated that, in this case, the skills of the models used to reconstruct tidally and synoptically driven temporal and spatial variability were almost equally good and comparable to the skill when using the Kalman filter approach. However, differently from the case of optimal linear estimator (e.g., the Kalman filter approach), of particular value is the capability of the GAN to learn and replicate processes with multiple time scales and the associated nonlinear interactions between them.

Using data from different sources (real observations or data from another numerical model) resulted in a decrease in the skill, and the patterns of disagreement with the test data were constrained by the model dynamics, generally reflecting the signal-to-noise ratio. Thus, the index of agreement between the reconstructions and validation data depends strongly on the patterns of the tidal amplitude. Including real coastal observations in the learning process increased the skill of the model. Obviously, GANs optimally learn from data from different sources. The lower skill of the experiments, in which real coastal observations are not used in the training process, reveals a similarity with the problems in data assimilation when errors in the data and model are not treated appropriately. Using other independent observations when training the GAN has the potential to further increase the power of the proposed method in real applications. This method can be attempted in other oceanographic settings.

Annex 1: Loss Functions of the Generator and Discriminator

The Unet-like GAN is established to train two defined network models: Generator ($G(Z)$) and Discriminator (D). $G(Z)$ is established to reconstruct SSH maps with full information, where Z denotes the input SSH maps with only several tidal gauge point information available. Generated SSH maps from $G(Z)$ together with real SSH maps from the numerical model are fed into the discriminator (D), and the outputs of D are a dense scalar that represents the probability of discriminating whether the input maps are from the real input map (which is from the numerical model). The stopping criterion of model training is that the discriminator cannot distinguish whether the generated SSH maps are from real maps.

The GAN loss functions are defined as in Goodfellow et al. (2014):

$$L_G^{GAN} = E\left[\log\left(D\left(G\left(z\right)\right)\right)\right] \tag{3}$$

$$L_D^{GAN} = E\left[\log\left(D\left(x\right)\right)\right] + E\left[\log\left(1 - D\left(G\left(z\right)\right)\right)\right] \tag{4}$$

where L_G is the generator loss and L_D is the discriminator loss. The GAN model parameters are trained and updated based on the following minimization and maximization method:

$$\min \max V(G,D) = E_{x \sim P_{\text{data}}(x)}\left[\log\left(D(x)\right)\right] + E_{z \sim P_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (5)$$

here, $P_{\text{data}}(x)$ is the real SSH map data distribution, taking real data sample x from $P_{\text{data}}(x)$. $P_z(z)$ is the input(z) (as in Figure 10) SSH data distribution, sampling z from $P_z(z)$. E is the expectation operator. Based on the original DCGAN loss function, we also introduce $L1$ as pixel-wise reconstruction loss and $L2$ as content loss into the generator for accomplishing our experiments. The discriminator loss is the same as the original GAN. The overall loss for the generator is:

$$L_1 = \left\| G(z) - x \right\|$$

$$L_2 = \left\| G(z) - x \right\|_2^2$$

$$L_G^{GAN} = E\left[\log\left(D\left(G(z)\right)\right)\right] + L1 + L2$$

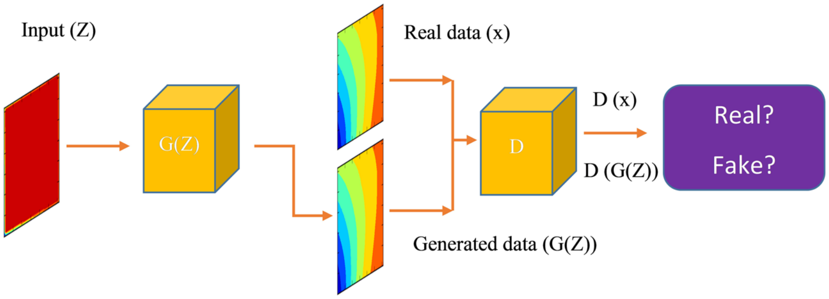

Figure 10. Schematic representation of the sea-level variability reconstruction GAN system. Input (Z) is the SSH input sample with only tidal gauge point data available, $G(Z)$ denotes the generator module for mapping input (Z) to generated data samples $G(Z)$. Real data (x) is the real numerical model SSH data sample, D represents the discriminator module for discriminating the real data (x) and the generated data ($G(Z)$) as real or fake samples. GAN, generative adversarial network; SSH, sea surface height.

In the LF reconstruction model, an additional pixel-wise reconstruction loss L_a is introduced into the generator part:

$$L_a = \left\| G_{OC}(z) - x \right\|$$

$$L_G^{GAN} = E[\log(D(G(z)))] + L1 + L2 + L_a$$

here, $G_{OC}(z)$ represents the generated coarse sea surface LF map data sample. x is the same as above, denoting the real numerical sample data.

Acknowledgments
We thank Joanna Staneva for the support and advice in developing the GCOAST model and discussions of the presented results. Thanks are also due to the anonymous reviewers for their useful comments. Reviewer #1 proposed to compare ML against Kalman filter approach. Zeguo Zhang was funded by the China Scholarship Council. Emil V. Stanev and Sebastian Grayek acknowledge funding from the Initiative and Networking Fund of the Helmholtz Association through the Advanced Earth System Modeling Capacity (ESM) project.

Data Availability Statement

This study was conducted using EU Copernicus Marine Service Information, which is available at https://marine.copernicus.eu/.

References

Alvera-Azcárate, A., Barth, A., Rixen, M., & Beckers, J.-M. (2005). Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions. Application to the Adriatic Sea surface temperature. *Ocean Modelling*, *9*, 325–346. https://doi.org/10.1016/j.ocemod.2004.08.001

Andersen, O. B. (1999). Shallow water tides in the northwest European shelf region from TOPEX/POSEIDON altimetry. *Journal of Geophysical Research*, *104*, 7729–7741. https://doi.org/10.1029/1998JC900112

Barth, A., Alvera-Azcárate, A., Licer, M., & Beckers, J.-M. (2020). DINCAE 1.0: A convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations. *Geoscientific Model Development*, *13*, 1609–1622. https://doi.org/10.5194/gmd-13-1609-2020

Becker, G. A., Dick, S., & Dippner, J. W. (1992). Hydrography of the German Bight. *Marine Ecology Progress Series*, *91*, 9–18. http://doi.org/10.3354/meps091009

Beckers, J.-M., & Rixen, M. (2003). EOF calculation and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology*, *20*, 1839–1856. https://doi.org/10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2

Cipollini, P., Calafat, F. M., Jevrejeva, S., Melet, A., & Prandi, P. (2017). Monitoring sea level in the coastal zone with satellite altimetry and tide gauges. *Surveys in Geophysics*, *38*, 35–57. https://doi.org/10.1007/s10712-016-9392-0

Cotter, A., Shamir, O., Srebro, N., & Sridharan, K. (2011). Better mini-batch algorithms via accelerated gradient methods. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24 (NIPS 2011)* (pp. 1647–1655). Granada, Spain: Curran Associates Inc., 57 Morehouse Lane Red Hook NY United States. Retrieved from http://papers.nips.cc/paper/4432-better-mini-batch-algorithms-via-accelerated-gradient-methods.pdf

Dong, H. W., & Yang, Y. H. (2019). *Towards a deeper understanding of adversarial losses*. eprint arXiv, Cornell University. Retrieved from https://arxiv.org/abs/1901.08753

Egbert, G. D., & Erofeeva, S. Y. (2002). Efficient inverse modeling of barotropic ocean tides. *Journal of Atmospheric and Oceanic Technology*, *19*(2), 183–204. https://doi.org/10.1175/1520-0426(2002)019<0183:EIMOBO>2.0.CO;2

Flather, R. A., & Proctor, R. (1983). Prediction of North Sea storm surges using numerical models: Recent developments in the U.K. In J. Sundermann, & W. Lenz (Eds.), *North Sea dynamics* (pp. 299–317). Berlin, Heidelberg: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-68838-6_21

Frolov, S., Baptista, A., & Wilkin, M. (2008). Optimizing fixed observational assets in a coastal observatory. *Continental Shelf Research*, *28*, 2644–2658.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27 (NIPS 2014)* (pp. 2672–2680). Montreal, Quebec, Canada: Curran Associates, Inc. Retrieved from https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Grayek, S., J.Staneva, J., Schulz-Stellenfleth, W., & Stanev, E. V. (2011). Use of FerryBox surface temperature and salinity measurements to improve model based state estimates for the German Bight. *Journal of Marine Systems*, *88*(1), 45–59.

Haigh, I. D., Pickering, M. D., Green, J. A. M., Arbic, B. K., Arns, A., Dangendorf, S., et al. (2019). The tides they are A-changin': A comprehensive review of past and future nonastronomical changes in tides, their driving mechanisms and future implications. *Annual Reviews of Geophysics*, *57*. https://doi.org/10.1029/2018RG000636

Hansen, W. (1956). Theorie zur Errechnung des Wasserstandes und der Stromungen in Randmeeren nebst Anwendungen. *Tellus*, *8*, 287–300. https://doi.org/10.3402/tellusa.v8i3.9023

Heaps, N. S. (1969). A two-dimensional sea model. *Philosophical Transactions of the Royal Society A*, *265*, 93.

Ho-Hagemann, H. T. M., Hagemann, S., Grayek, S., Petrik, R., Rockel, B., Staneva, J., et al. (2020). Internal model variability of the regional coupled system model GCOAST-AHOI. *Atmosphere*, *11*(3), 227. https://doi.org/10.3390/atmos11030227

Huthnance, J. (1991). Physical oceanography of the North Sea. *Ocean and Shoreline Management*, *16*(3–4), 199–231. https://doi.org/10.1016/0951-8312(91)90005-M

Ioffe, S., & Szegedy, C. (2015). *Batch normalization: Accelerating deep network training by reducing internal covariate shift* (pp. 448–456). Proceedings of the 32nd International Conference on Machine Learning, PMLR. Retrieved from http://proceedings.mlr.press/v37/ioffe15.pdf

Jacob, B., & Stanev, E. V. (2017). Interactions between wind and tidally induced currents in coastal and shelf basins. *Ocean Dynamics*, *67*, 1263–1281. https://doi.org/10.1007/s10236-017-1093-9

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Legates, D. R., & McCabe, G. J., Jr. (1999). Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241.

Liu, H. Y., Jiang, B., Xiao, Y., & Yang, C. (2019). *Coherent semantic attention for image inpainting*. Paper presented at the IEEE International Conference on Computer Vision (ICCV 2019), pp. 4170–4179, Seoul, Korea, IEEE. Retrieved from http://openaccess.thecvf.com/content_ICCV_2019/html/Liu_Coherent_Semantic_Attention_for_Image_Inpainting_ICCV_2019_paper.html

Loshchilov, I., & Hutter, F. (2019). *Decoupled weight decay regularization*. eprint arXiv, Cornell University. Retrieved from https://arxiv.org/abs/1711.05101

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). *Rectifier nonlinearities improve neural network acoustic models*. Retrieved from http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf

Madec, G. (2016). *NEMO ocean engine. Note du Pole de modélisation, Version 3.6 27*. Paris, France: Institut Pierre-Simon Laplace (IPSL).

Mao, X. J., Shen, C. H., & Yang, Y. B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29 (NIPS 2016)* (pp. 2810–2818). Barcelona, Spain: Curran Associates Inc., 57 Morehouse Lane Red Hook NY United States. Retrieved from http://papers.nips.cc/paper/6172-image-restoration-using-very-deep-convolutional-encoder-decoder-networks-with-symmetric-skip-connections

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. 1. A discussion of principles. *Journal of Hydrology*, *10*, 282–290.

O'Dea, E. J., Arnold, A. K., Edwards, K. P., Furner, R., Hyder, P., Martin, M. J., et al. (2012). An operational ocean forecast system incorporating NEMO and SST data assimilation for the tidally driven European North-West shelf. *Journal of Operational Oceanography*, *5*, 3–17.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory* (p. 14). Cambridge, MA: MIT.

Otto, L., Zimmerman, J. T. F., Furnes, G. K., Mork, M., Saetre, R., & Becker, G. (1990). Review of the physical oceanography of the North Sea. *Netherlands Journal of Sea Research*, *26*(2–4), 161–238. https://doi.org/10.1016/0077-7579(90)90091-T

Peeck, H. H., Proctor, R., & Brockmann, C. (1983). Operational storm surge models for the North Sea. *Continental Shelf Research*, *2*(4), 317–329. https://doi.org/10.1016/0278-4343(82)90024-3

Proudman, J., & Doodson, A. T. (1924). The principal constituent of the tides in the North Sea. *Philosophical Transactions of the Royal Society A*, *244*, 185–219.

Radford, A., Metz, L., & Chintala, S. (2015). *Unsupervised representation learning with deep convolutional generative adversarial networks*. eprint arXiv, Cornell University. Retrieved from https://arxiv.org/abs/1511.06434

Riley, P. (2019). Three pitfalls to avoid in machine learning. *Nature*, *572*, 27–29. Retrieved from https://www.nature.com/articles/d41586-019-02307-y

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), Medical image computing and computer-assisted intervention–MICCAI 2015 (Vol. 9351, pp. 234–241). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Schulz-Stellenfleth, J., & Stanev, E. V. (2010). Statistical assessment of ocean observing networks—A study of water level measurements in the German Bight. *Ocean Modelling*, *33*(3–4), 270–282.

Soetje, K. C., & Brockmann, C. (1983). An operational numerical model of the North Sea and the German Bight. In J. Siindermann, & W. Lenz (Eds.), *North Sea dynamics* (pp. 95–107). Berlin, Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-642-68838-6_6

Tonani, M., Sykes, P., King, R. R., McConnell, N., Péquignet, A.-C., O'Dea, E., et al. (2019). The impact of a new high-resolution ocean model on the Met Office North-West European Shelf forecasting system. *Ocean Science*, *15*, 1133–1158. https://doi.org/10.5194/os-15-1133-2019

Wahl, T., Haigh, I. D., Woodworth, P. L., Albrecht, F., Dillingh, D., Jensen, J., et al. (2013). Observed mean sea level changes around the North Sea coastline from 1800 to present. *Earth Science Review*, *124*, 51–67. https://doi.org/10.1016/j.earscirev.2013.05.003

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, *32*, 2088–2094. https://doi.org/10.1002/joc.2419

Yu, J. H., Lin, Z., Yang, J. M., Shen, X. H., Lu, X., & Huang, T. S. (2018). *Generative Image Inpainting with Contextual Attention.* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5505 -5514, Salt Lake City, Utah, IEEE. Retrieved from http://openaccess.thecvf.com/content_cvpr_2018/papers/Yu_Generative_Image_Inpainting_CVPR_2018_paper.pdf

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*(1), 1–7. https://doi.org/10.1038/s41467-019-11786-6

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss Functions for image restoration with Neural Networks. *IEEE Transactions on Computational Imaging*, *3*(1), 47–57. https://doi.org/10.1109/TCI.2016.2644865