Transferring a deep learning model to map sediment and ecological properties based on satellite images from the Wadden Sea to Oman

Maaike Breedveld
6511651

MSc Thesis

February 2024

Supervisors: Elisabeth Addink, Wiebe Nijland,
and Logambal Madhuanand

Contents

Abstract: 3

1. Introduction 3

2. Background 6

2.1 Deep learning models 6

2.1.1 Structure of Convolutional Neural Networks (CNNs) 6

2.1.2 Autoencoders 6

2.2 Deep learning method for tidal flats 7

2.2.1 Autoencoder model 7

2.2.2 Feature extraction 7

3. Data & Methods 8

3.1 Study area 9

3.1.1 Geomorphology & hydrodynamics 9

3.1.2 Ecology 11

3.1.3 Compared to the Wadden Sea 11

3.2 Data collection 11

3.2.1 Field data 11

3.2.2 Satellite images 14

3.3 Preprocessing field data 14

3.3.1 SLC failure data gaps 15

3.3.2. Masking 16

3.4 Models 16

3.4.1 Training model 16

3.4.2 Feature generation 17

3.4.3 Random Forest (RF) model 17

3.5 Scenarios 17

3.5.1. No change 17

3.5.2. Finetuning 17

3.5.3. Freezing 18

3.5.4. From scratch 64 18

3.5.5. From scratch 32 18

3.5.6. Sentinel 18

3.5.7 No change 2015 18

4. Results 19

4.1 Field data 19

4.2 Gap filling & image selection 21

4.3 Training 24
4.4 Features 25
4.5 Cross-validation accuracy Landsat 7 25
4.5.1 Cross-validation accuracy of the environmental variables *25*
4.5.2 Cross-validation accuracy of the different scenarios *28*
4.6 Cross-validation accuracy Sentinel 2 29
5. Discussion 29
5.1 Performance of transfer learning techniques 29
5.2 Landsat versus Sentinel 30
5.3 Comparing to performance in the Wadden Sea 30
5.4. Generalization 31
5.5 Challenges and potential improvements 31
6. Conclusion 32
7. References 32

Abstract:

Worldwide tidal flats fulfil important ecological and economical roles. However, they are under increasingly high pressure. For effective management consistent monitoring of sedimentary and ecological variables is needed. To overcome the limitations of field sampling and traditional satellite image-based methods, a new deep learning-based method was proposed by Madhuanand et al. (2023) to predict sediment and ecological properties of tidal flats based on satellite images. This method uses a ResNet50-based deep learning model to generate features that are used as additional information on top of the original image for a random forest model to predict environmental variables. While satisfactory results were reached for the Wadden Sea it was unknown how well this method would be able to generalize to other tidal flat regions. Here the predictive performance of the method was tested for the tidal flats in Bar Al Hikman, Oman, using Landsat 7 and Sentinel 2 images and different transfer learning techniques. The tested scenarios include literal transfer without finetuning, finetuning, finetuning after freezing, and training from scratch. The prediction accuracy was evaluated for the median grain size, silt content, complete biomass, complete species richness, crab biomass, and crab species richness. It was found that the predictive accuracy of the model was much lower compared to the original accuracy achieved for the Wadden Sea. Similar to for the Wadden Sea the sediment properties had the highest cross-validation accuracy from the tested variables while the accuracy for species richness was low. The scenario using the pre-trained model without any additional training reached the highest cross-validation accuracy. A closer inspection of the results suggested that random forest predictions are sensitive to multi-year data and temporally separated field and imagery data. The remnants of data gaps from the SLC error of the Landsat 7 images were present in the generated features and could thus also have contributed to the low prediction accuracies. Therefore, when using this method focus should be on using satellite images without data gaps, using temporally closely matching field and satellite data, and predicting only one year at a time for the random forest model.

1. Introduction

Tidal flats are characterized by a high biodiversity and productivity and play an important ecological and economic role in many regions. The shallow waters make the tidal flats important nurseries for fish, crabs, and shrimps (Reise, 2012), and provide important foraging areas for waterbirds (Bom et al., 2018). The food and habitat provided are also of economic value by sustaining fisheries (Burt, 2014; Dissanayake et al., 2018). These fisheries are important for both the local population and the export. An example are the crabs in Oman which are valuable seafood items for both the domestic market and the export (Hehanna et al., 2013). Overall, the yearly global value of the services provided by intertidal mudflats is estimated to be around US$ $5.2 \cdot 10^{12}$ 2007$ (Dissanayake et al., 2018; Costanza et al., 2013). However, worldwide tidal flats are under pressure caused by coastal development, sea-level rise, coastal erosion, reduced sediment fluxes, subsidence, eutrophication, non-nutrient pollutants, and overfishing (Lever at al., 2001; Murray et al., 2019). This causes changes in biodiversity and species richness in many coastal environments (Lever et al., 2001).

Consistent monitoring of sedimentary and ecological variables is important for effective management of these regions (Miloslavich et al., 2018). Due to the nature of tidal flats, sampling can be challenging. If possible, it is often time-consuming and expensive. In recent years satellite images have been used for monitoring due to their ability to cover large areas at a time and reach hard-to-reach locations. While satellite images have a low resolution, they can still give information about sediment and ecological characteristics on the ground by using their relation to large-scale geomorphological structures and spectral information.

Tidal flats have a diverse geomorphology with a combination of channels and flats. Sediment and ecological properties can be linked to these geomorphological structures. Sediment properties are related to the energy conditions of the environment. At locations with higher flow velocities like large deep channels courser grained sandy sediments are often present, while the less energetic shallow flats contain more muddy sediment. Previous research has shown that sediment characteristics can be linked to the tidal channel distribution and intertidal DEM (Choi et al., 2011). The spatial distribution of macrobenthos is also closely related to environmental variables like surface elevation (Lee et al., 2013), sediment characteristics (Compton et al., 2013; Yoo et al., 2007), and temperature (Koo et al., 2005, 2007). Van der Wal et al. (2008) found for example that in their study area in the Netherlands the total macrobenthos biomass could be partly explained by the median grain size, mud content, and elevation. The temperature on tidal flats can be influenced by exposure time and water content in the sediment (Koo et al., 2007). Characteristics like moisture content and the presence of channels or flats can be taken from satellite imagery.

The uniform spectral nature of tidal flats however makes traditional spectral analysis challenging. Deep learning methods have been shown to be able to learn to extract complex structures and information from satellite images (Willcock et al., 2018). Therefore, deep learning models have been used increasingly for complex analysis tasks. One such deep learning-based method was developed by Madhuanand et al. (2023) to predict sediment and ecological properties from satellite images for the tidal flats in the Wadden Sea. Using a relatively low demanding method they reached satisfactory results. If their method would be able to generalize for other intertidal mudflats this could elevate the need for extensive field sampling and aid effective management.

There are however two main restrictions that can limit the use of deep learning models (Iman et al., 2023). Firstly, successful training of a deep learning model requires an extensive training dataset of at least thousands to tens of thousands of training images. Obtaining the required amount of data can be time-consuming or even impossible (Zhuang et al., 2020). By using a related existing dataset to initialize the model this training cost can be reduced. Secondly, the training of a deep learning model takes extensive time and process power. Deep transfer learning (DTL) was developed to help alleviate these restrictions and improve the performance of deep neural networks (Iman et al., 2023).

The concept of transfer learning in neural networks was first introduced in the 1990s by Lorien Pratt with the goal of improving the learning speed of neural networks (Pratt, 1993). It is inspired by the way humans learn new tasks by building on their previous experiences. It builds on concepts from psychology, neurobiology, and symbolic machine learning (Pratt, 1933). Similar to how synapses in the brain come pre-wired, the idea behind transfer learning is to use initialized weights from a network pre-trained on a source dataset as an improved starting point for training compared to randomly initialized weights. Intuitively, successful learning requires some connection between the two learning activities, in other words, the source and target domains should be linked by a higher-level common domain (Zhuang et al., 2020; Weiss et al., 2016).

Many transfer learning methods make use of the fact that different layers in convolutional neural networks (CNNs) learn features with different degrees of generalization (Yosinski et al., 2014; Neyshabur et al., 2020). The shallower parts of the network deal with more general features (e.g. edge detection, colour blobs) and are similar for most CNNs trained on images (Yosinski et al., 2014). Deeper into the network the learned features become more dataset-specific. For transfer deep learning the focus is generally on these general features which are applicable for both the source and target dataset. Multiple studies show that models pre-trained on a source dataset can effectively be used to improve training on a different target dataset (Iman et al., 2023). Most works utilize the CNN activations from the fully connected layers while using features from convolutional layers has received limited attention (Hu et al., 2015).

DTL is applied in a large variety of situations with a large variety of approaches. There are different ways to categorize DTL based on the homogeneity of the source and target data, the label-setting aspects, or the applied approaches (Iman et al., 2023). Based on the used approach DTL can be divided into four categories: instance-based, feature-/mapping-based, parameter-/network-/model-based, or relational-/adversarial-based (Pan and Yang, 2010; Tan et al., 2018). The most used approaches within DTL are model-based approaches (Iman et al., 2023). Model-based approaches focus on adjusting the network and include pretraining, freezing, finetuning, and adding fresh layers. These methods can be applied on their own but are usually combined (Iman et al., 2023). The easiest way to transfer information is literal weight transfer where the weights of the pre-trained model are transferred and used for the new task. This does not allow the model to learn any new information within the target dataset. Therefore, this is often combined by methods that require the model to also train on the target dataset like finetuning, freezing, and/or adding new layers to the model (Iman et al., 2023).

There are some complications that might occur when transferring a deep learning model. One of these problems is negative transfer (Zhuang et al., 2020). This happens when the source and target dataset or task are too dissimilar and transferring reduces the performance of the model. Another problem is catastrophic forgetting (Iman et al., 2023). This refers to the loss of learned skills by a trained model after further training on a target dataset. Freezing the lower-level layers of a transferred model or only training newly added layers should reduce the risk of catastrophic forgetting (Iman et al., 2023). Indeed, previous research pointed out that models sometimes perform less when applied to a different geographical or temporal setting (Tong et al., 2021).

The aim of this study is to transfer the deep learning method developed by Madhuanand et al. (2023) to Bar Al Hikman, Oman, to test if this method can be generalized and applied to a different geographic region. To answer this question the cross-validation accuracy of the median grain size, silt content, complete biomass, complete species richness, crab biomass, and crab species richness was evaluated and compared for several scenarios. Scenarios include training from scratch and the transfer learning methods of literal transfer, finetuning, and freezing. These scenarios will be tested on Landsat 7 images which match the temporal range of this study and on a Sentinel 2 image for which the method was developed but which does not match the temporal range. To answer the main question the following sub-questions will be answered:

- How can the scan line error of the Landsat 7 images be resolved?
- What is the cross-validation accuracy of each learning technique for the sediment and ecological variables using Landsat 7 images?
- What is the cross-validation accuracy of each learning technique for the sediment and ecological variables using a Sentinel 2 image?

2. Background

2.1 Deep learning models

Deep learning models have become increasingly popular because of their ability to describe complex relationships with non-linear data. In particular, convolutional neural networks (CNNs) have become popular for performing tasks like image classification and segmentation (Iman et al., 2021; Pires de Lima & Marfurt, 2019; Hu et al., 2015).

2.1.1 Structure of Convolutional Neural Networks (CNNs)

A convolutional neural network is a type of deep neural network that contains at least one convolutional layer (Ketkar & Moolayil, 2021). A convolutional layer uses a mathematical operation called a convolution which convolves the input image or feature map with a kernel or filter. The output of this convolution is a new feature map. The kernel can have different sizes and has weights and biases associated with it. These weights and biases are optimized during the training of the network. CNNs also typically contain pooling layers. These are layers that reduce the spatial dimensions of the feature maps while maintaining the main important structures. This down-sampling reduces the dependence on the location of features and reduces overfitting (Goodfellow et al., 2016). Most commonly max pooling is used which takes the maximum value within a chosen local window to produce the downscaled output map. Several convolutional and pooling layers are often followed by one or several fully connected layers at the end of the model.

CNNs and other deep learning models learn by backpropagation. By evaluating a cost or loss function the model evaluates how well the output corresponds to the desired output. During training the objective of the model is to minimize this cost function and find the global, or in practice often local, minimum in the loss landscape. One of the most well-known methods to achieve this is the gradient descent (Rumelhart et al., 1986). It calculates the partial derivative of the total error with respect to each weight in the network for each input-output combination. It then accumulates the partial derivatives of all these input-output combinations to update the weights proportionally. As this method evaluates all training data before taking one step down the loss domain this method is computationally expensive. Therefore, many models now use the stochastic gradient descent (SGD). It works similarly to the gradient descent, but it is done for a random sample of the training data called a batch. This speeds the model up while keeping the accuracy similar.

An important variable that determines the performance of the deep learning model is the learning rate. The learning rate determines the step size of the gradient descent. If the learning rate is too high the gradient descent can overshoot the local minimum resulting in a model that fails to converge (LeCun et al., 1998). On the other hand, if the learning rate is too small training will take a long time. As a result, it is important to choose an appropriate learning rate for a given task. An often-applied method uses a dynamic learning rate that adapts to the stage of learning. The models often start with a higher learning rate to quickly approach a local minimum. When this local minimum is approached the learning rate is reduced to prevent overshooting and help the model converge.

2.1.2 Autoencoders

An autoencoder is a specific type of neural network that is useful for learning representations of the data without supervision (Bank et al., 2023). As a special type of encoder/decoder architecture it consists of an encoder, which compresses the input through encoded layers, and a decoder, which up-samples the encoded layers again to create an output. Autoencoders learn to reconstruct the input images. By trying to minimize the difference between the input and the output images they

learn key features that contain spatial and textural information describing the input image. These features containing spatial and textural information can then be used as additional information next to the spectral data for predicting environmental and ecological variables from the satellite images (Madhuanand et al., 2023).

2.2 Deep learning method for tidal flats

Madhuanand et al. (2023) developed a deep learning method to predict sediment and ecological properties from satellite images in the Wadden Sea. They used an autoencoder model to generate features from Sentinel 2 images which could be used as additional information to train a random forest model. In the following part their method will be described in more detail.

2.2.1 Autoencoder model

To learn representative features of the input data a regularized version of the autoencoder called the variational autoencoder (VAE) was used. The backpropagated loss is calculated as a combination of the reconstruction and regularization loss as in the equation below.

$$L = \lambda_1 L_m + \lambda_2 L_{kl}$$

Where L_1 is the reconstruction loss, L_2 is the regularization loss, and λ_1 and λ_2 are weights. The reconstruction loss is derived from the mean squared error (MSE). It tries to optimize the model by minimizing the differences between pixels in the reconstructed image and the input. The regularization loss is calculated from the divergence loss which tries to keep the learned distributions close to a standard distribution. The two losses were combined with a weight. The optimal performing weights determining the relative contribution of the reconstruction and regularization loss to the total backpropagated loss were determined by Madhuanand et al. (2023) to be 1 and 0.001 respectively with the reconstruction loss thus having a higher weight.

The used VAE uses the ResNet-50 structure. ResNet-50 is a residual neural network. These types of neural networks contain skip connections that bypass layers in the model. These skip connections are implemented to increase the performance of very deep neural networks (He et al., 2016). They are used to extract both low-level and high-level features (Madhuanand et al., 2023). The structure of the model can be seen in Figure 1. The ResNet-50 model contains 49 convolutional layers and one fully connected layer. It also contains 2 pooling layers, one max pooling after the first convolutional layer and one average pooling before the fully connected layer. After the patches move through this encoder it moves through several up-sampling layers. These up-sampling layers used rectified linear unit (ReLU) activation to introduce non-linearity (Madhuanand et al., 2023).

2.2.2 Feature extraction

Each of the blocks in the encoder contains features that describe the input patches. When moving through the encoder the number of features increases while their dimensions decrease. The first layer containing 64 features of 32x32 pixels each was selected to be used to generate the features for the random forest model and is up-sampled through bilinear interpolation. This layer was chosen to minimize uncertainties introduced by up-sampling and to keep the number of features low to prevent increasing computational demands. To reduce the border effect when the features are up-sampled, the patches are created with an overlap of 30%.

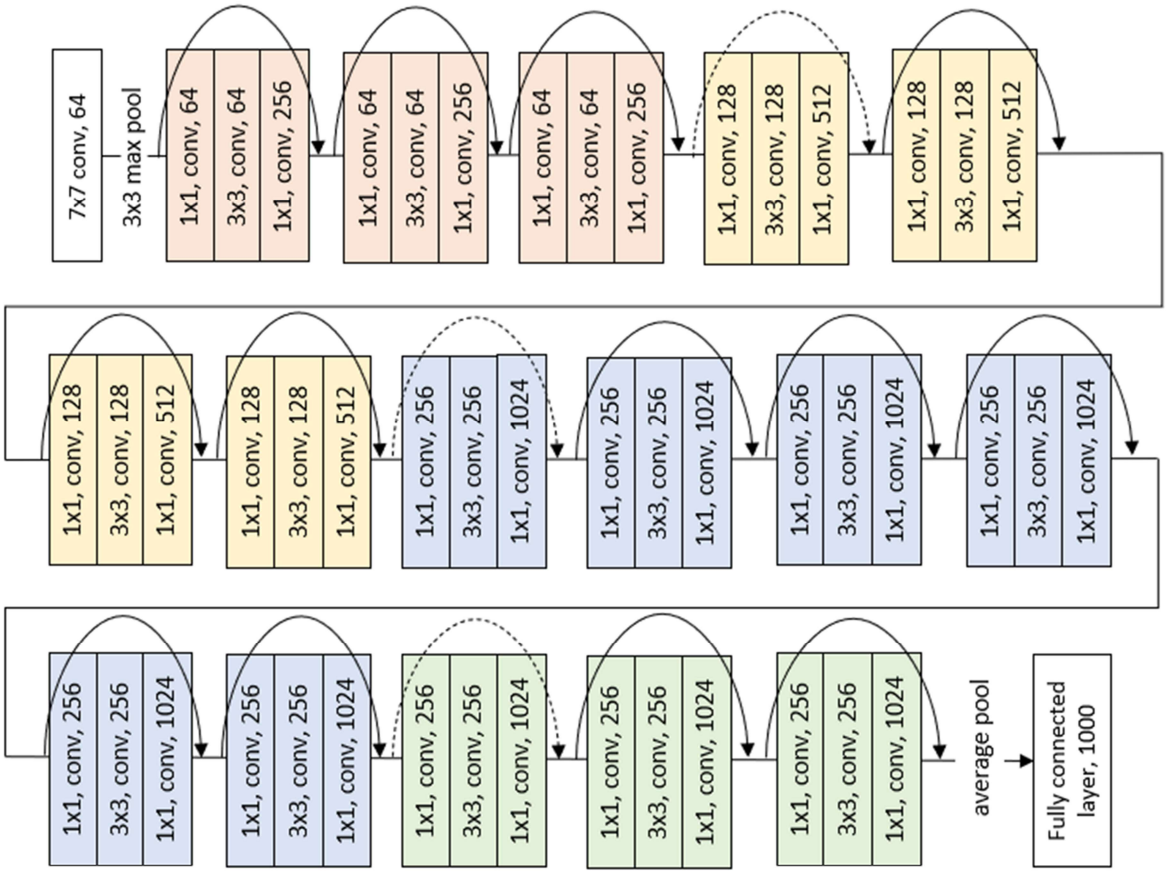

Figure 1 Structure of the ResNet-50 architecture.

2.2.3 Random forest model

The prediction of the variable of interest is done using a random forest regression model. The method was chosen because it had proven to be able to fit models with many input variables even with non-linearity or collinearity. To tune the random forest using cross-validation the data is split into a training dataset (90%) and a validation dataset (10%). A k-fold cross-validation was implemented to optimize the number of observations when fitting the trees. They specifically used a 10-fold cross-validation which was repeated three times to reduce variance in the model performance.

3. Data & Methods

In the section the used methods will be described. This section will start with a description of the study area (section 3.1). This is followed by a description of the data collection (section 3.2) and pre-processing steps (section 3.3). After this, the used models and used parameters are discussed (section 3.4. Finally, a short description is given of the different scenarios (section 3.5). An overview of the used methods and procedures can be found in the flow chart on the next page (figure 2).

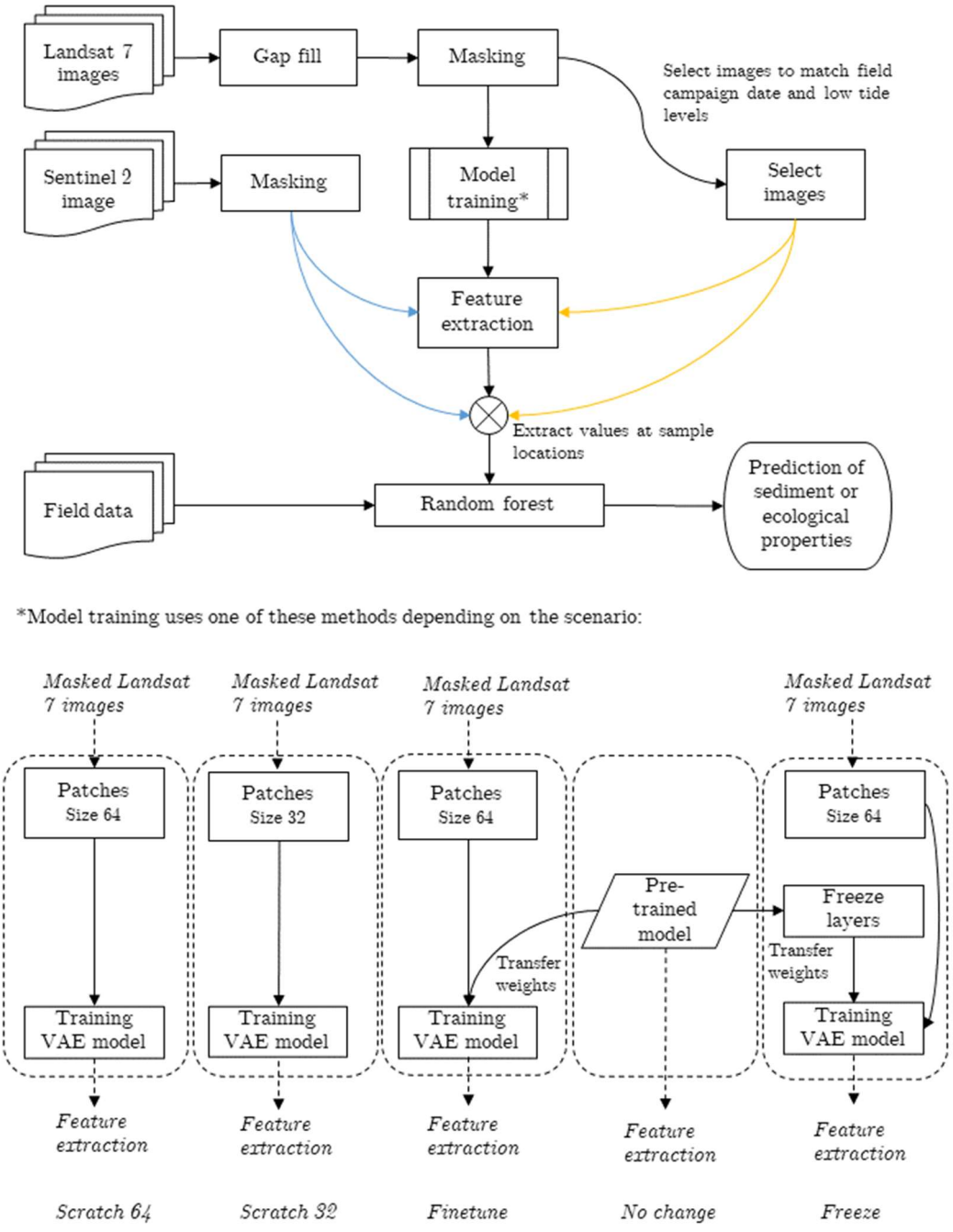

Figure 2 Overview of the used workflow. At the top the overall workflow. The blue lines depict data used for Sentinel 2 scenarios while the yellow lines represent the Landsat 7 scenarios. The training of the deep learning model depends on the used scenario and is explained in more detail in the bottom part for each method.

3.1 Study area

3.1.1 Geomorphology & hydrodynamics

This study focussed on the intertidal mudflats in Barr Al Hikman, Oman. The mudflats are located on the narrow continental shelf along the Arabian Peninsula in the Arabian Sea and form the connection between the inland sabkhas and the Arabian Sea. They can be divided into three subareas: Khawr, Shannah, and Filim (figure 3). Together they have an estimated area of around 190 km^2. The mudflats

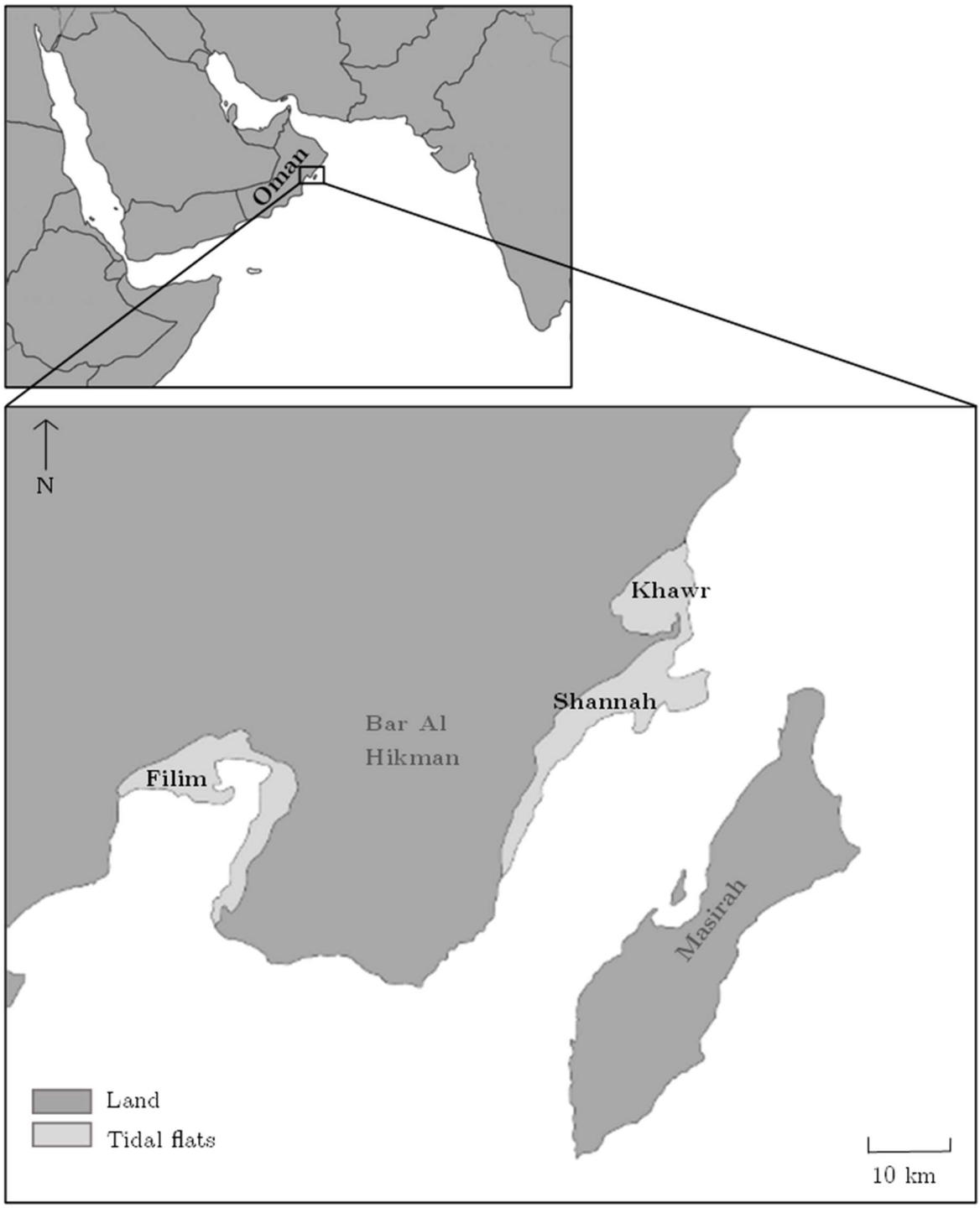

Figure 3 Location of the studied tidal flats in Filim, Khawr, and Shannah in Bar Al Hikman in Oman.

contain bare mudflats, seagrass meadows, intertidal pools, reef structures, and channels (Bom et al., 2018). Each of these features uniquely influences the water dynamics, soil structure, and physical and chemical properties in its area and thereby influences the spatial distribution of flora and fauna.

The sediment on the mudflats originates from both the sea as well as from wind-blown deposits (Bom et al., 2018). Environmental parameters were estimated for a part of Shannah by Bom et al. (2020). They estimated that the intertidal elevation differences are up to 2.2 m and that the median grain size ranges between 136 and 249 µm. The sediment depth ranges between 0 and 20 cm with shallower sediment depths at locations with reef structures (Bom et al., 2020). The tidal cycle in Barr

Al Hikman is a combination of diurnal and semidiurnal. The average tidal range is around 1.5 m but can be up to around 3.5 m during springtides.

Along the coast of Oman there is nearly continuous upwelling. The strength of the upwelling is influenced by the monsoon winds (Bom et al., (2018). From June to August monsoon winds monsoon winds come from the south-west. This causes increased upwelling around the coast bringing nutrient-rich water and increasing productivity (Bom et al., 2018). From December to February the winds change to north-easterly which reverses the water currents.

3.1.2 Ecology

Barr Al Hikman is well-known for its large diversity of marine life and abundant birdlife. The area is an important habitat for many species of fish and crustaceans of economic value (Bom et al., 2018). Primary production on the intertidal mudflats takes place by phytoplankton, seagrasses, epiphytes, and microphytobenthos. Estimates take the primary production of the Barr Al Hikman area, including the subtidal and mangroves, to be up to 160.000 tons g C yr^{-1} placing it among the most productive exclusive economic zones (EEZs) that include internationally important intertidal mudflats larger than 5000 ha (Bom et al., 2018).

Benthic invertebrates (e.g. bivalves, gastropods, polychaetes, and crustaceans) are important for the ecology of intertidal flats. They form an important link between the primary producers and secondary consumers including fish and birds and are usually the starting point for the characterization of food webs in intertidal mudflats (Bom et al., 2018). Of most interest are the macrozoobenthic invertebrates which are all benthos > 1 mm. They form the main food supply for birds and marine predators in Barr Al Hikman (Bom et al., 2018). Barr Al Hikman has a benthic community with at least 97 identified species. Most of these species belong to the gastropods, bivalves, or brachyuran crabs. A field campaign in 2008 showed standing stock densities of the same order of magnitude as for other international intertidal mudflats (Bom et al., 2018). The same field campaign also showed that the main benthic biomass of Barr Al Hikman consists mainly of gastropods and bivalves, while crustaceans and polychaetes contribute less to the total biomass. Over 78% of the biomass density was made up of three species: *Pirenella arabica, Cerithium scabridum,* and *Pillucina fischeriana.* With the exception of crabs little is known about the interannual variations in the benthic community.

3.1.3 Compared to the Wadden Sea

When comparing the tidal flats in Bar Al Hikman to the studied area in the Wadden Sea the variables of interest differ substantially. The tidal flats in Oman have a much higher median grain size with a mean of 190 µm in Oman compared to 145 µm in the tested region in the Wadden Sea (Madhuanand et al., 2023). The silt content of the tidal flats in Oman is also much lower compared to the Wadden Sea, with a mean of 6.3% to 13% (Madhuanand et al., 2023). For the ecological variables, the tidal flats in Oman have a lower biomass and richness compared to the Wadden Sea.

3.2 Data collection

3.2.1 Field data

Field data was collected by Bom et al. (2017, 2020) during eight field campaigns in 2008 and 2011-2015. Geocoded ecological data on the biomass and taxonomy of macrobenthos was collected during each of these campaigns. During the 2008 campaign all macrobenthos were collected while the campaigns from 2011-2015 focused only on crabs. Sediment data on median grain size and silt content was collected during the 2011 campaign. The 2008 campaign took samples in each of the regions Filim, Khawr, and Shannah. The other campaigns sampled only in Shannah. An overview of

Table 1 Number of field data points for each field campaign in Oman.

Campaign	# of points	# of points with ecological data
2008	282	254
2011	458	125
2012 March	84	7
2012 November	440	227
2013	137	97
2014	168	78
2015 March	115	61
2015 November	121	93

Campaign	# of points	# of points with sediment data
2011	66	66

the number of field data points for each campaign can be found in Table 1. The locations of the field samples for each campaign are given in Figure 4 for sediment properties and Figure 5 for the ecological variables.

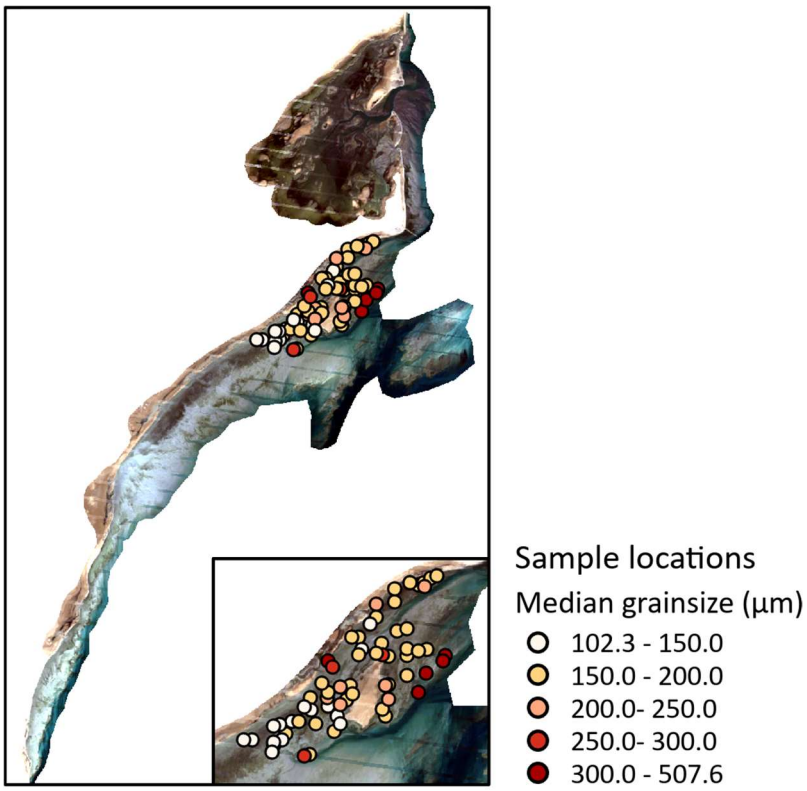

Figure 4 Field collection locations of the sediment properties collected during the field campaign in 2011. Image acquisition date is 14 February 2008 and is in true colour.

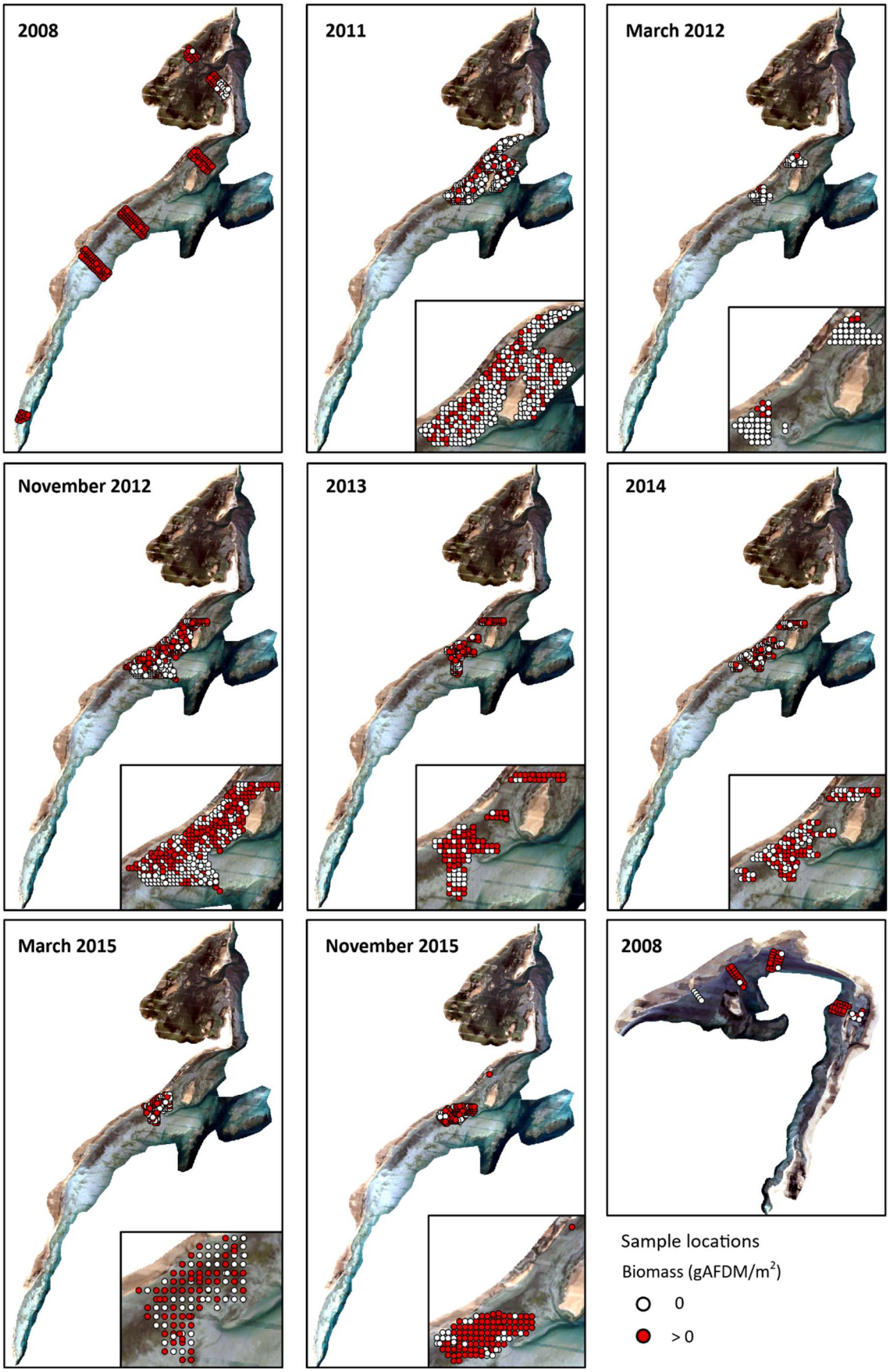

Figure 5 Locations of the field sampling points. The points with no recorded biomass are excluded in the analysis. Image acquisition date is 14 February 2008 and are shown in true colour.

3.2.2 Satellite images

Satellite images were collected to match the temporal availability of the field data between 2008 and 2015. While the original method used Sentinel 2 images, this satellite only launched mid 2015 making it unsuitable for the collection of sufficient training images within the desired time range. Instead, Landsat 7 images were used. Launched in 1999 and operational until 2022 this satellite covered the entire desired temporal range. While Landsat 7 ETM+ data suffers from a failed scan line corrector (SLC) from 2003 onwards other open-source sensors were deemed unsuitable because of their inability to cover the entire desired temporal range (Sentinel 2 and Landsat 8) or failure to cover the desired region (Spot). Based on the results of Madhuanand et al. (2023) only the blue, green, red, and near-infrared(NIR) bands were used. These bands have a spectral range of 0.45-0.52 µm for the blue band, 0.52-0.60 µm for the green band, 0.63-0.69 µm for the red band, and 0.77-0.90 µm for the NIR band. The spatial resolution of the images is 30m.

The models were trained using Landsat 7 images for the years 2008 and 2011-2015 which correspond to the years during which field data was collected in Oman. Only satellite images with exposed tidal flats and no cloud cover over the tidal flats were selected. A total of 21 images complied with these requirements.

To test the effect of the different sensor and SLC error the prediction ability was also tested using a Sentinel 2 image, which the model was originally developed for in the Wadden Sea. As mentioned before, because of the launch date of Sentinel 2 not enough images were available within the field campaign timeframe to train the VAE model. Therefore, only one image was collected corresponding to the field date of the last campaign in December 2015. The selected image was acquired on 18 December 2015. It was chosen based on its relatively good exposed tidal flats and close temporal proximity to the field data.

The image taken on 18 December 2015 was selected This image was then used to generate features and train the random forest model. The spectral ranges of the four used bands are 0.458-0.523 µm for the blue band, 0.543-0.578 µm for the green band, 0.650-0.690 µm for the red band, and 785-899 µm for the NIR band and they have a spatial resolution of 10 m.

3.3 Preprocessing field data

The biomass and taxonomy of each collected macrobenthos was stored separately in the dataset. For each sample location, the biomass of all the collected macrobenthos was summed to get the total biomass. The species richness was calculated by evaluating the total number of different species for each sample location based on the recorded taxonomy. Points with no recorded macrobenthos were excluded from the analysis which resulted in a slightly lower number of data points (table 1). Because of the limited number of points for each campaign for the crab data these points were all combined in the random forest model, resulting in 688 data points for crab biomass and crab richness. Since the November 2012 campaign did have a similar number of points as used by Madhuanand et al. (2023) the predictive performance for crab data of this year was also evaluated separately from the other crab data. For the complete biomass and richness from 2008 there were 254 data points. All 66 sediment samples could be used.

The histograms of the field data showed that the distribution is skewed towards the lower values with some higher values extremes (section 4.1, figure 7). In particular, the biomass shows this pattern strongly. A logarithmic correction was applied to create a more normal distribution (figure 8).

3.4 Preprocessing satellite images

To prepare the Landsat 7 images for training of the VAE model a few steps had to be taken which will be described in more detail below.

3.3.1 SLC failure data gaps

From 2003 onwards Landsat 7 ETM+ images suffer from a failed scan line corrector which results in large strips of data gaps within the images. Before these images can be used by the model these data gaps have to be filled. Different techniques have been suggested to deal with these data gaps (e.g. Hossain et al., 2015; Scaramuzza & Barsi, 2005; Yin et al., 2016). These techniques broadly fall into two categories. Single image techniques use interpolation techniques to fill the data gaps using the data surrounding the gaps. A drawback of these techniques is that they create made-up data and that inherent textures tend to get lost. Multiple image techniques use a second image to fill the data gaps. Since the data gaps do not occur at the same location for each image, data from different images can be used to fill the gaps. The advantage of this technique is that it uses real data with inherent textures. As deep learning models use texture this is an important advantage. A disadvantage is that there is a spectral offset between the different images and that geomorphological changes can occur between the two moments in time. Histogram matching is a method that aims to reduce these spectral differences. However, it is quite sensitive to differences in radiance due to for example clouds, snow, or reflection of water (Scaramuzza & Barsi, 2005).

After testing different filling techniques and visually comparing the results it was found that the multiple-image techniques produced the best results (see section 3.1). As histogram matching was computationally more intensive while not visually improving results this study chooses to fill the data gaps by simply using data from other images without any corrections. When choosing the specific image to fill the gaps three criteria were taken into account. The most important criterion was that the image chosen to fill the gaps must fill the data gaps as much as possible. This means that the data gaps had to be filled completely or at a maximum only left a limited amount of 1-to-2-pixel wide gaps at the edge of the area of interest (figure 6). The second criterion was that the filling image had to

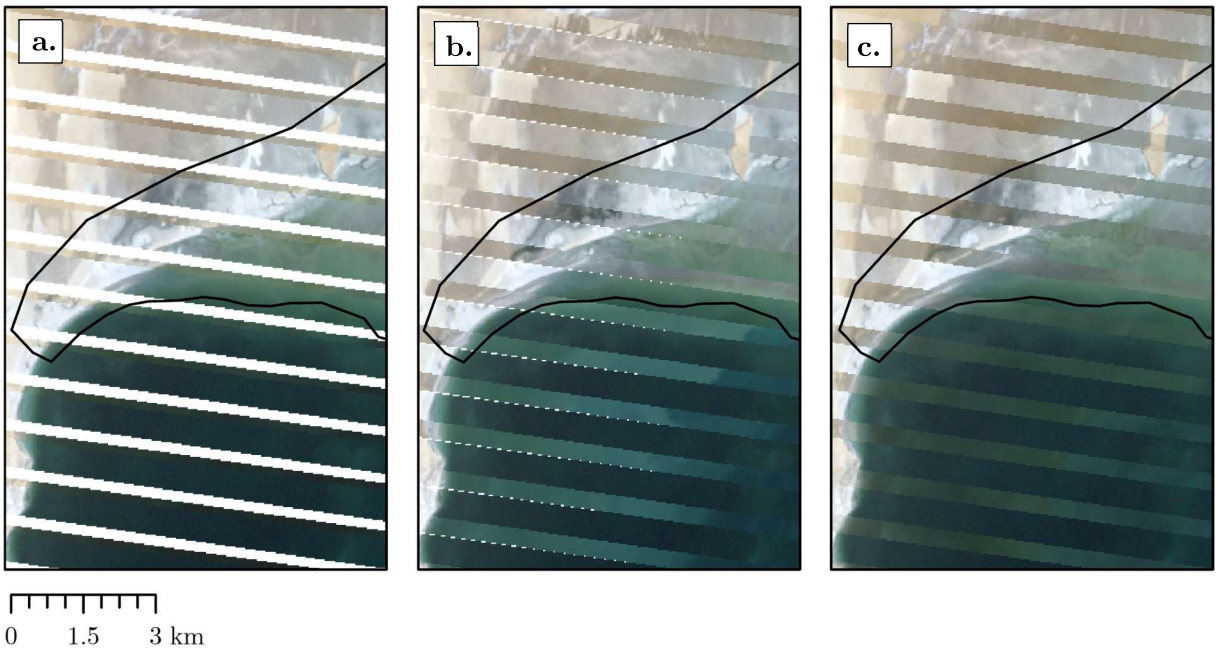

Figure 6 Examples of types of gap fills between two true colour images at the most southwest tip of Filim. a) unsuitable filling image; filling images does not fill the gaps , (b) accepted filling image; filling image left a limited amount of 1-to-2-pixel wide gap, and c) preferred filling image; filling image completely fills the data gaps is the study area.

be taken within a specific time frame from the target image. When evaluating this criterium seasonal years were used where a new year starts after the monsoon season taking place from June to September. Preferably images taken within in same seasonal year were used and the filling image should not be more than two seasonal years away from the target image. This aims to reduce the effect of changes occurring over time like changing geomorphology. Finally, images with the lowest tidal difference were preferred to minimize discontinuities caused by different tidal elevations. Leftover 1-2 pixel wide gaps in a limited part of the study area were filled using a median filter with a 5 by 5 window.

3.3.2. Masking

Before the images were passed to the model the land and open sea were masked out. The extent of the tidal flats was based on the maximum extension of the tidal flats within our data set. The landward extension was taken from two images taken during high tide. A small border around the land and sea was kept to account for some error. The tidal flats in the region of Filim were separated from Khawr and Shannah by this masking.

3.4 Models

The models used are developed by Madhuanand et al. (2023). They consist of a training model based on Resnet 50, a model that generates features, and a random forest model that predicts the variables of interest. They will be described in more detail below.

3.4.1 Training model

The training model uses the pre-processed Landsat 7 images to train the deep learning model based on Resnet 50. First, each image was cut into smaller patches which form the input from which the deep learning model will learn. These patches had a size of 64x64 or 32x32 pixels depending on the scenario being tested (see section 3.5) and had an overlap of 10%. This resulted in a total number of 9,009 patches for a patch size of 64x64 and 36,582 patches for a patch size of 32x32. These patches were then divided into a training (80%) and validation (20%) dataset. As a result, the training dataset for scenarios using a patch size of 64x64 contained 7,207 patches while the remaining 1,802 patches were used for validation. For scenarios with a patch size of 32x32, the training dataset contained 29,265 patches while the validation dataset contained 7,317 patches. The training model was then trained for 100 epochs at the end of which the model was saved to be used to generate features. The hyperparameters used for each scenario can be found in Table 2.

Table 2 Hyperparameters used for training and the image source used for training and feature generation for each of the different scenarios.

Scenario	Patch size	Learning rate	Epochs	Batch size	Training images	Feature generation images
No change	64	-	-	-	-	Landsat 7
Finetuning	64	0.000001	100	32	Landsat 7	Landsat 7
Freezing	64	0.000001	100	32	Landsat 7	Landsat 7
Scratch 64	64	0.00001	100	32	Landsat 7	Landsat 7
Scratch 32	32	0.00001	100	32	Landsat 7	Landsat 7
No change 2015	64	-	-	-	-	Landsat 7
Sentinel no change	64	-	-	-	-	Sentinel 2
Sentinel finetuning	64	0.000001	100	32	Landsat 7	Sentinel 2
Sentinel freezing	64	0.000001	100	32	Landsat 7	Sentinel 2
Sentinel scratch 64	64	0.00001	100	32	Landsat 7	Sentinel 2
Sentinel scratch 32	32	0.00001	100	32	Landsat 7	Sentinel 2

3.4.2 Feature generation

The models trained by the training model and the model from Madhuanand et al. (2023) trained on the Wadden Sea were used to generate features from a given input image. These features are taken from the first layer which contains 64 features of 32x32 pixels. To match the input size these features are up-sampled to 64x64 pixels. For each field campaign an input image was chosen to be close to the field data collection date and to have the lowest possible tidal elevation. The tidal elevation was estimated from the tidal gauge located in Masirah. An overview of the used Landsat 7 images can be found in Table 3 and seen in Figure 4. For the Sentinel scenarios (section 3.5.6) 64 features were generated for the single Sentinel 2 image.

3.4.3 Random Forest (RF) model

From the produced 64 features and the original images with four spectral bands values were extracted at the locations of the field data. The random forest model then built a random forest to predict the field data variable of interest based on these values. A random forest with 800 trees and a maximum depth of 20 was used. The minimum samples for a split was set to 10 with the minimum samples per leaf being 2. The hyperparameters for the RF model can be found in Table 3. Compared to the setting used by Madhuanand et al. (2023) the number of splits used was decreased to 5 and the number of repeats to 2 to account for the lower number of field data points available.

3.5 Scenarios

Different transfer methods were tested in the paper. These are referred to as scenarios throughout the paper. They differ in how the model used for the feature generation was trained and which images were used to generate the features. An overview of the used hyperparameters and used training data per scenario can also be found in Table 2 in section 3.4.1.

3.5.1. No change

In this scenario, the easiest and basic transfer learning method of literal weight transfer without any further training was applied. Thus, the model from Madhuanand et al. (2023) trained on the Wadden Sea was used for feature generation without any further training on the tidal flats in Oman. As the model from Madhuanand et al. (2023) used a patch size of 64x64 this scenario also uses a patch size of 64x64.

3.5.2. Finetuning

The target tidal flats in Oman differ from the tidal flats in the Wadden Sea. These regional differences can be important for environmental predictions but are not yet learned by the model trained on the Wadden Sea. The expectation is that finetuning of the model on images of the tidal flats of Oman will increase the predictive performance. Thus, the trained model from Madhuanand et al. (2023)

Table 3 Hyperparameters used for the random forest model.

Hyperparameters	Values
Nbr. of trees	800
Max. depth	42
Min. samples split	10
Min. samples leaf	2
Nbr. of splits	5
Nbr. of repeats	2
Random state	42

was used to initialize the weights of the training model which was then trained on images from Oman. Since the model has been pre-trained the learning rate was reduced by a factor of 10 compared to training from scratch to prevent weights from changing too quickly and forgetting of learned information (Li et al., 2020). Therefore, a learning rate of 0.000001 was used for this scenario.

3.5.3. Freezing

Literature suggests finetuning to be a suboptimal transfer learning technique because of catastrophic forgetting (Iman et al., 2023). Freezing has been suggested as a technique to prevent catastrophic forgetting by not updating the shallow layers which contain more general features likely suitable for similar tasks (Iman et al., 2023). For this scenario, the weights were again initialized by the weights from Madhuanand et al. (2023). However, in contrast to the previous scenario now all layers with the exception of the last two were frozen. This way the model could keep the general information learned on the Wadden Sea dataset while the last two layers could adapt to new region-specific information. Again, since the model has been pre-trained the lower learning rate of 0.000001 could be used.

3.5.4. From scratch 64

This scenario uses the same model structure as the model used by Madhuanand et al. (2023) and a patch size of 64x64. However, instead of using the pre-trained weights the model was trained from scratch from randomly initialized weights to see how the model structure and setup would perform on the new region without transfer learning.

3.5.5. From scratch 32

The pixel resolution of Landsat 7 of 30 m is much larger than the 10 m pixel resolution of the Sentinel 2 images that the model was originally based on. As a result, a patch size of 64x64 pixels is quite large for the tidal flat region of Oman. It results in only 7,207 training patches and 1,802 validation patches which is quite little to train a deep learning model, for which, as said, preferably tens of thousands of patches are used. To see the effect of using a lower patch size the model structure was adapted to work with a patch size of 32x32 pixels which resulted in a much larger training dataset of 29,265 training patches and 7,317 validation patches. The hyperparameters are kept the same as in the scratch 64 scenario.

3.5.6. Sentinel

The quality of the Landsat 7 images could influence the performance of the model as inconsistencies between the original image and the image used to fill the SLC error gaps can be picked up and learned by the model. To get an idea of this influence the same scenarios mentioned above were also applied to a Sentinel 2 image. This is the same type of image that the model was developed for in the Wadden Sea. Compared to Landsat 7, Sentinel 2 images have a higher resolution of 10 m. As Sentinel 2 image collection started only at the end of 2015 the temporal resolution is reduced for these scenarios and only a single image collected in December 2015 was used.

3.5.7 No change 2015

To also test the effect of the lower temporal resolution without the added effect of the different sensor one Landsat 7 scenario was repeated while using only one Landsat 7 image for the feature generation. Similar to the Sentinel scenarios the image used for this was the December 2015 image. Because of its results, the trained model of Madhuanand et al. (2023) was used like in the other ‘no change’-scenarios.

4. Results

4.1 Field data

The median grain size in the study area ranged between 102.3 and 507.6 µm with a mean of 190.8 µm. The silt content ranged from 1.08% to 15.2% with a mean of 6.30%. The biomass, measured as ash-free dry mass (AFDM), ranged from 0.027 to 207.8 gAFDM/m^2 with a mean of 22.2 gAFDM/m^2 full biomass collected in 2008 and from 0.0002 to 1.7 gAFDM/m^2 with a mean of 0.10 gAFDM/m^2 for only crabs collected during the 2011-2015 campaigns. The full species richness ranged between 1 and 15 with a mean of 4. The crab richness ranged between 1 and 6 and had a mean of 1.

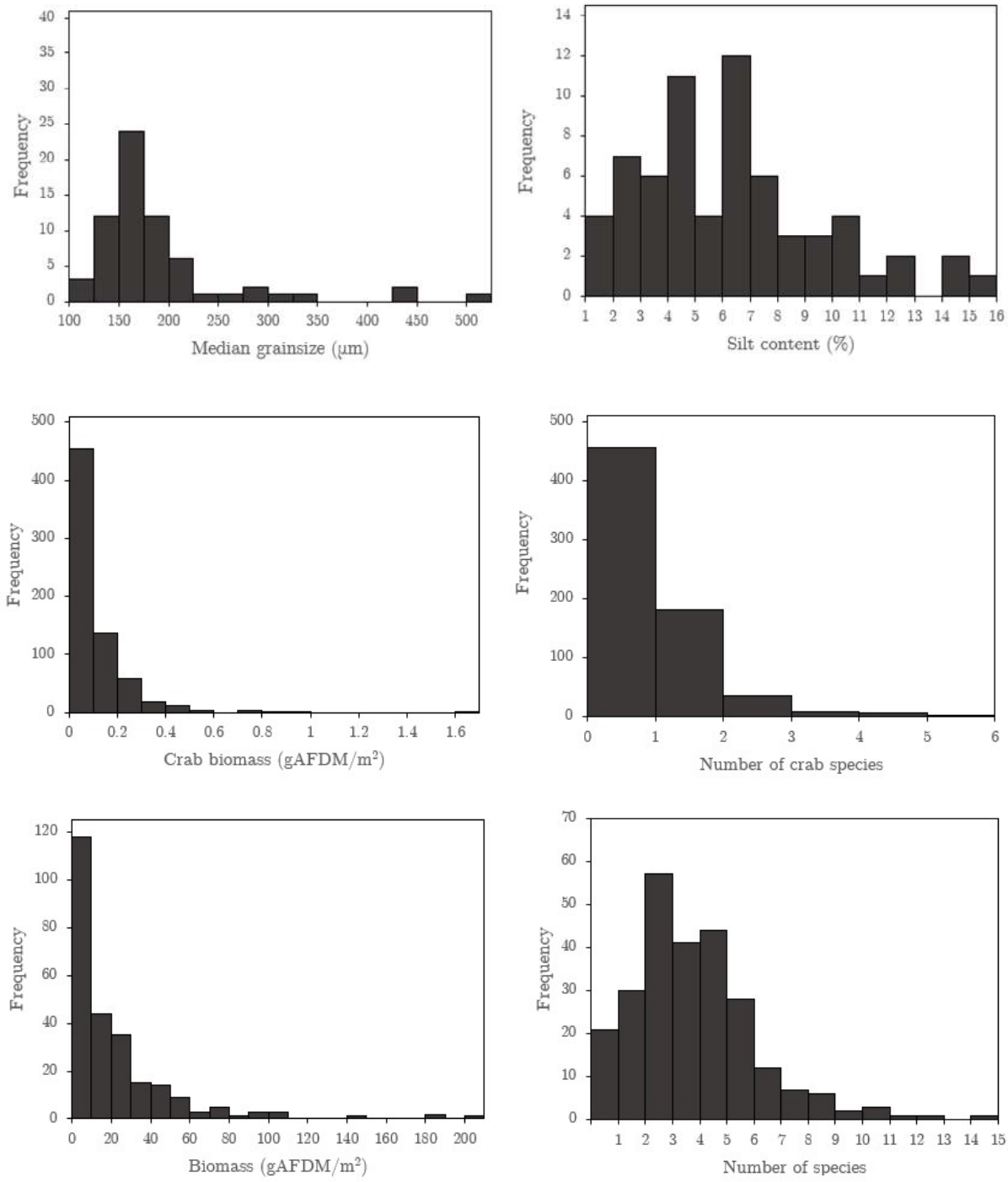

Figure 7 Histograms of the field data show a skewed distribution towards the left with some higher extremes.

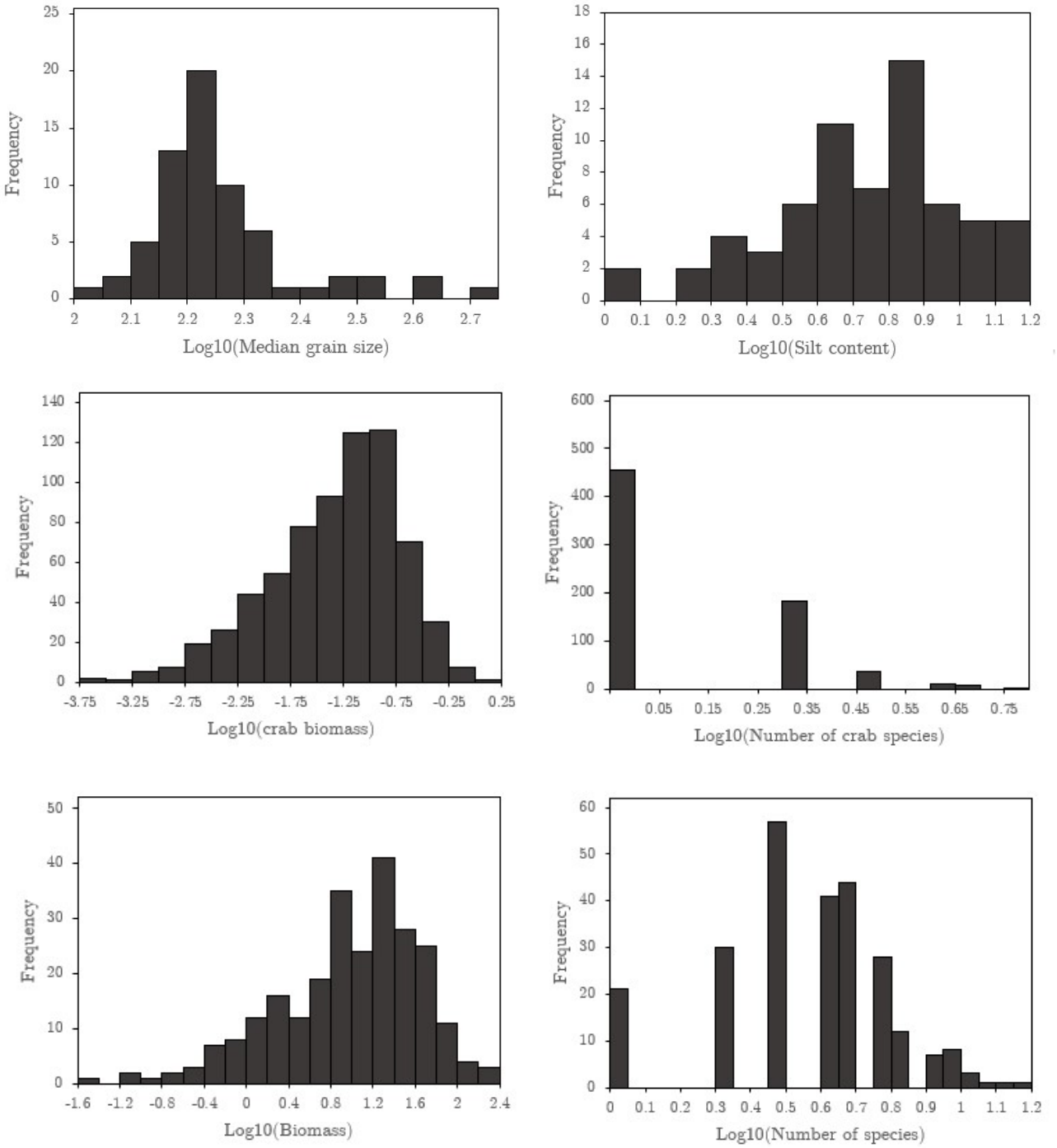

Figure 8 Histograms of the data after the logarithmic transformation. The distributions shifted towards the right and have fewer extreme outliers.

As can be seen from the histogram in Figure 7 the distributions of all variables are skewed toward the left with some extreme outliers for the higher variables. In particular, for the biomass, crab biomass, and crab richness are the lower values much more frequent. After the logarithmic transformation the amount of extreme outliers is reduced (figure 8). Only for the crab species richness does the transformation not seem to change the distribution strongly. The effect of the logarithmic transformation was most pronounced for both biomass variables. After the transformation their distribution seems slightly skewed to the right.

4.2 Gap filling & image selection

Comparing the results of different gap-fill methods it can be seen that there are clear differences (figure 9). When using an interpolation technique the gap fills are smoother than the surroundings. If a second image is used to fill the gaps the filled regions have more texture. However, the spectral differences are more pronounced. The difference in water level between two images creates a clear difference between the original image and the filled stripes.

To find the most optimal images to combine for multi-image gap filling an overview was created which evaluated each image combination based on the criteria mentioned in section 3.3.1. This overview can be seen in Table 4. It gives an overview of how well an image fills the gaps, how many seasonal years they are separated, and the tidal elevation difference. Of the 21 images that had to be filled, 13 were filled with an image acquired within a seasonal year. 5 images were selected that did not fill the entire data gap but left a 1-to-2-pixel wide gap to be filled with interpolation.

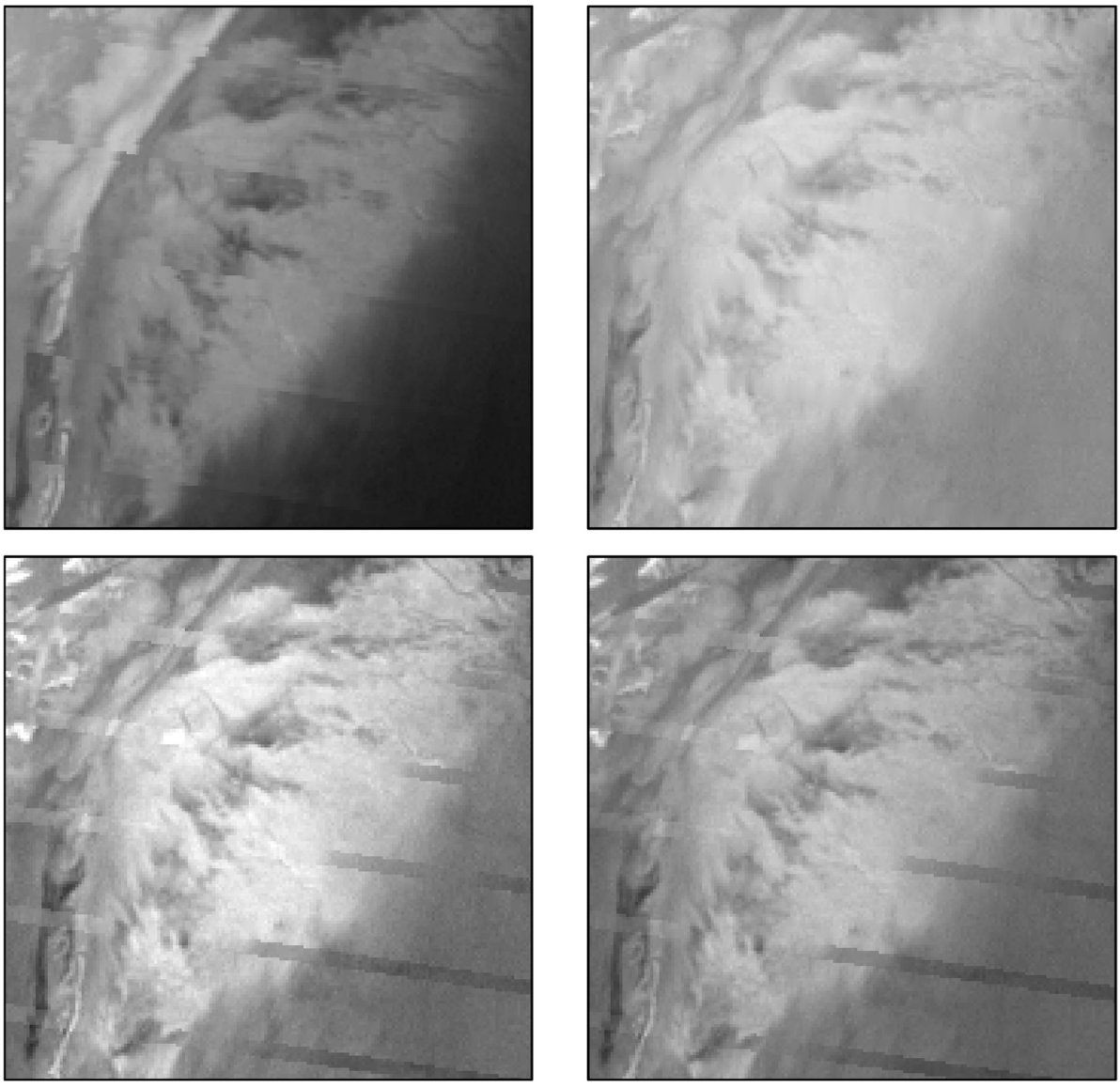

Figure 9 The results of different gap fill techniques. (a) interpolation using median focal statistics with a 7x7 window, (b) interpolation using inverse distance with a 7x7 window, (c) filling with a different image without corrections, (d) filling with a different image using global histogram matching.

Table 4 Images used for filling the SLC data gaps. Black line separate different seasonal years separated by the monsoon season. The closer to the black squares, the close the acquisition dates are. Within the black boxes containing the black squares are images within the same seasonal year. The error gaps of the image on the left are filled with the image on top. Values are absolute tidal differences.

	18/12/2015	02/12/2015	24/05/2015	01/02/2015	15/12/2014	29/11/2014	06/06/2014	13/01/2014	28/12/2013	18/05/2013	02/05/2013	16/04/2013	23/11/2012	07/11/2012	29/04/2012	13/04/2012	20/10/2011	26/03/2011	10/03/2011	01/03/2008	14/02/2008
18/12/2015																					
02/12/2015			32	466	52	33	119	347	222	55	326	21	514	296	158	393	405	444	173	346	894
24/05/2015		32		434	84	1	87	315	254	87	358	11	482	264	190	425	373	476	141	378	926
01/02/2015		466	434		518	433	347	119	688	521	792	445	48	170	624	859	61	910	293	812	1360
15/12/2014		52	84	518		85	171	399	170	3	274	73	566	348	106	341	457	392	225	294	842
29/11/2014		33	1	433	85		86	314	255	88	359	12	481	263	191	426	372	477	140	379	927
06/06/2014		119	87	347	171	86		228	341	174	445	98	395	177	277	512	286	563	54	465	1013
13/01/2014		347	315	119	399	314	228		569	402	673	326	167	51	505	740	58	791	174	693	1241
28/12/2013		222	254	688	170	255	341	569		167	104	243	736	518	64	171	627	222	395	124	672
18/05/2013		55	87	521	3	88	174	402	167		271	76	569	351	103	338	460	389	228	291	839
02/05/2013		326	358	792	274	359	445	673	104	271		347	840	622	168	67	731	118	499	20	568
16/04/2013		21	11	445	73	12	98	326	243	76	347		493	275	179	414	384	465	152	367	915
23/11/2012		514	482	48	566	481	395	167	736	569	840	493		218	672	907	109	958	341	860	1408
07/11/2012		296	264	170	348	263	177	51	518	351	622	275	218		454	689	109	740	123	642	1190
29/04/2012		158	190	624	106	191	277	505	64	103	168	179	672	454		235	563	286	331	188	736
13/04/2012		393	425	859	341	426	512	740	171	338	67	414	907	689	235		798	51	566	47	501
20/10/2011		405	373	61	457	372	286	58	627	460	731	384	109	109	563	798		849	232	751	1299
26/03/2011		444	476	910	392	477	563	791	222	389	118	465	958	740	286	51	849		617	98	450
10/03/2011		173	141	293	225	140	54	174	395	228	499	152	341	123	331	566	232	617		519	1067
01/03/2008		346	378	812	294	379	465	693	124	291	20	367	860	642	188	47	751	98	519		548
14/02/2008		894	926	1360	842	927	1013	1241	672	839	568	915	1408	1190	736	501	1299	450	1067	548	

Gaps filled completely

Some gaps (max. 1-2 pixel wide)

Gaps not filled

Seasonal years.

Used image combination

The images selected for feature generation can be found in Table 5. Most images had a tidal elevation below the mean tidal level of 1471 m (table 4). Two images were selected with a slightly higher tidal elevation in the absence of a better alternative. The image taken on 14 February 2008 has a tidal elevation corresponding to low water levels during spring tide. An overview of the images can be found in Figure 10.

Table 5 Tidal elevation of the images used to generate features.

Date	Tidal elevation (mm)
02/12/2015	1242
01/02/2015	1422.5
29/11/2014	1208.5
28/12/2013	768
23/11/2012	1560.5
13/04/2012	845.5
20/10/2011	1555.5
14/02/2008	329

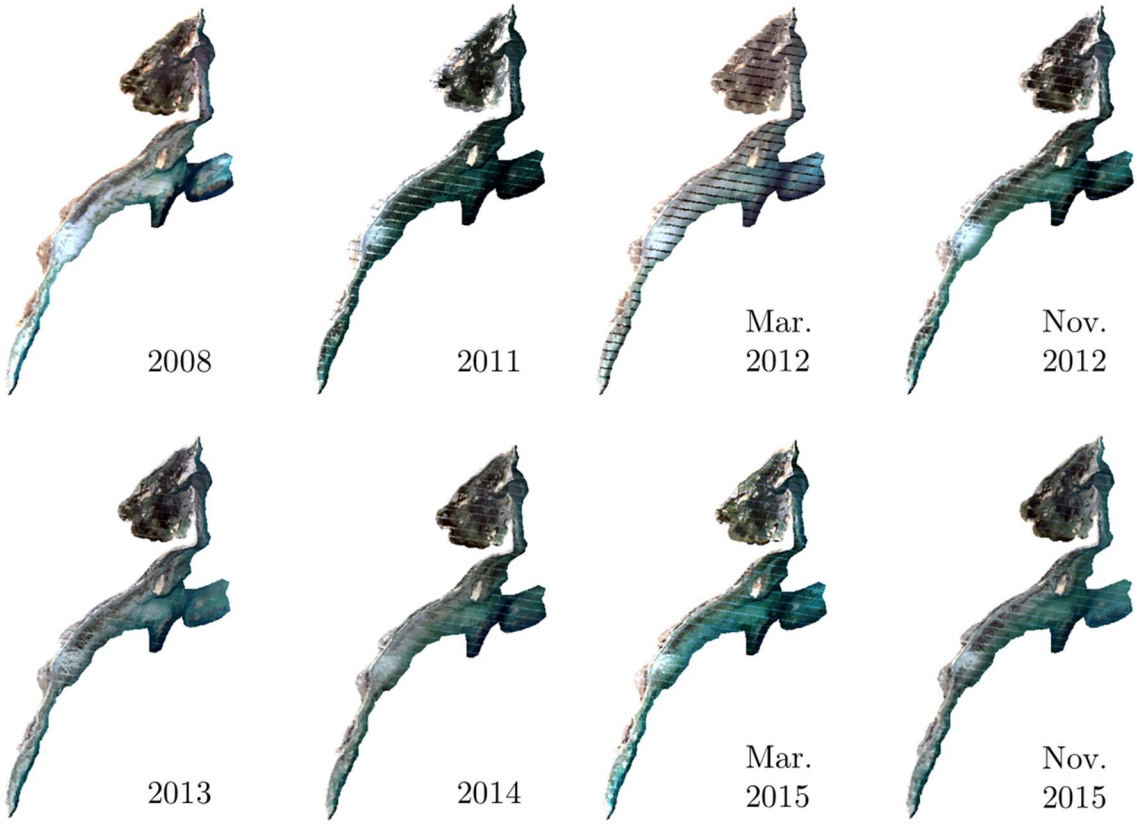

Figure 10 Landsat 7 images used for feature extraction shown in true colour. Striping from the filled data gaps is still visible.

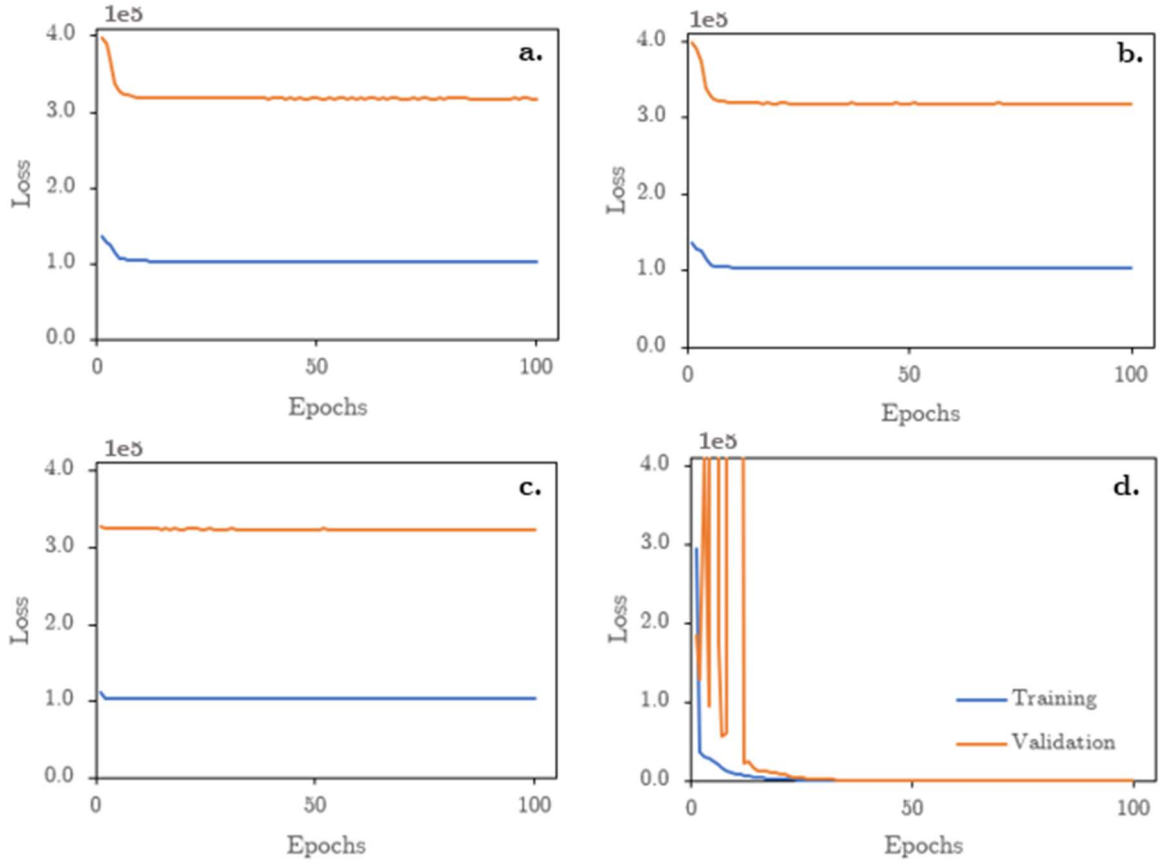

Figure 11 Training and validation loss during the training of the four different scenarios (a) finetuning, (b) freezing, (c) scratch 64, and (d) scratch 32.

4.3 Training

For all four models trained the training loss showed a steep initial loss drop followed by a slower loss drop which until around epoch 20 after which the additional loss drop is minimal (figure 11). For the models trained from scratch the strongest loss drop occurred after the first epoch. For the model trained from patches with a patch size of 64x64 this initial strong loss drop is followed by a gradual drop that seems to continue to drop very slightly the rest of the epochs. For the model trained from scratch with a patch size of 32x32 the steep loss drop after the first epoch is followed by a slower loss drop until around epoch 20 after which the loss decreases only minimally. For the models that are finetuned from the pretrained model on the Wadden Sea the initial strong loss drop lasts for the first five epochs. After this the loss drops slightly slower for a few epochs before it reaches an almost stable state around 15 epochs. The final training loss was similar for the models trained from scratch for a patch size of 64x64, the finetuned model, and the partially frozen model. The model trained on patches of 32 pixels reaches a much lower training loss.

The validation loss shows similar patterns. The validation loss for the model trained on patches of 64x64 pixels shows a steep initial drop followed by a much more varying but overall declining trend. The model trained on the patches of 32x32 shows a strong fluctuation for the first 10 epochs but with an overall decreasing trend. After these first 10 epochs, the loss follows a slower decline until it reaches a mostly stable state after 50 epochs. The final validation loss of this model was much lower than for the other models. The validation loss for the models based on the pre-trained model were very similar (Figures 5a and b). The initial loss was larger than for the other two models. After a steep

initial drop the loss seems to quickly converge after around 10 epochs at a slightly lower level than for the model trained from scratch on patch of 64x64.

4.4 Features

While for each scenario about 23-29 of the 64 generated features are empty for the tidal flats area, the remaining features show a diverse range of patterns. A limited number of features only showed a few select small regions with values while the rest of the flats were empty. This type of feature occurred 7 times for the model trained on the Wadden Sea and 1-2 times for the other models. Some features seem to enhance edges with a specific aspect, others seem to sharpen the edges, and some others create larger more blurred areas. While some features seem to extract larger scale similar areas, others detect and enhance smaller structural variations. Some features seem to enhance the striping effect of the Landsat images. Distinct geomorphological structures are visible in the features like channels and the edges of tidal flats. A small selection of features created from different scenarios is given in Figure 12.

The features generated by the finetuned model and the partly frozen model are similar in the sense that the same features contain information or are empty for the tidal flat region. Their values are also similar with the finetuned model having slightly lower values with a mean of 1.15 compared to 1.20 for the partly frozen model. The features generated by the model trained on the Wadden Sea had much higher values with a mean of 4.24 for the non-empty features. The models trained from scratch have a mean of 1.11 and 0.99 for a patch size of 64x64 and 32x32 respectively.

4.5 Cross-validation accuracy Landsat 7

4.5.1 Cross-validation accuracy of the environmental variables

The cross-validation accuracy of the sediment properties exceeds the ecological predictions (table 6). For each scenario, the cross-validation accuracy was highest for the silt content with predictions ranging from 16.1% for the ‘scratch 32’-scenario to 25.0% for the 'no change’-scenario. The predictions for median grain size ranged from 8.6% to 21.0% with the same scenarios performing the worst and best respectively.

For the ecological predictions the best cross-validation accuracy was reached for the complete biomass collected during the 2008 campaign. With predictions ranging from 16.3% to 19.0% these accuracies exceed those on crab biomass for which the cross-validation accuracy ranged from 3.08% to 5.57%. When only using the crab biomass of November 2012 the crab biomass cross-validation accuracies were higher ranging from 8.36% to 14.1%. The cross-validation accuracy was lowest for the crab richness ranging from 1.71% to 3.56%. The cross-validation accuracy for the complete species richness ranged from 3.17% to 3.82%.

The cross-validation accuracy of the ‘no change 2015’-scenario is lower than that of the ‘no change’-scenario for most variables. The crab biomass and crab richness now have negative R2 values. The cross-validation accuracy of the silt content and the full biomass had the best cross-validation accuracy under these conditions with 13.4% and 12.9% respectively. The complete species richness performed better compared to the ‘no change’-scenario with a cross-validation accuracy of 9.0%.

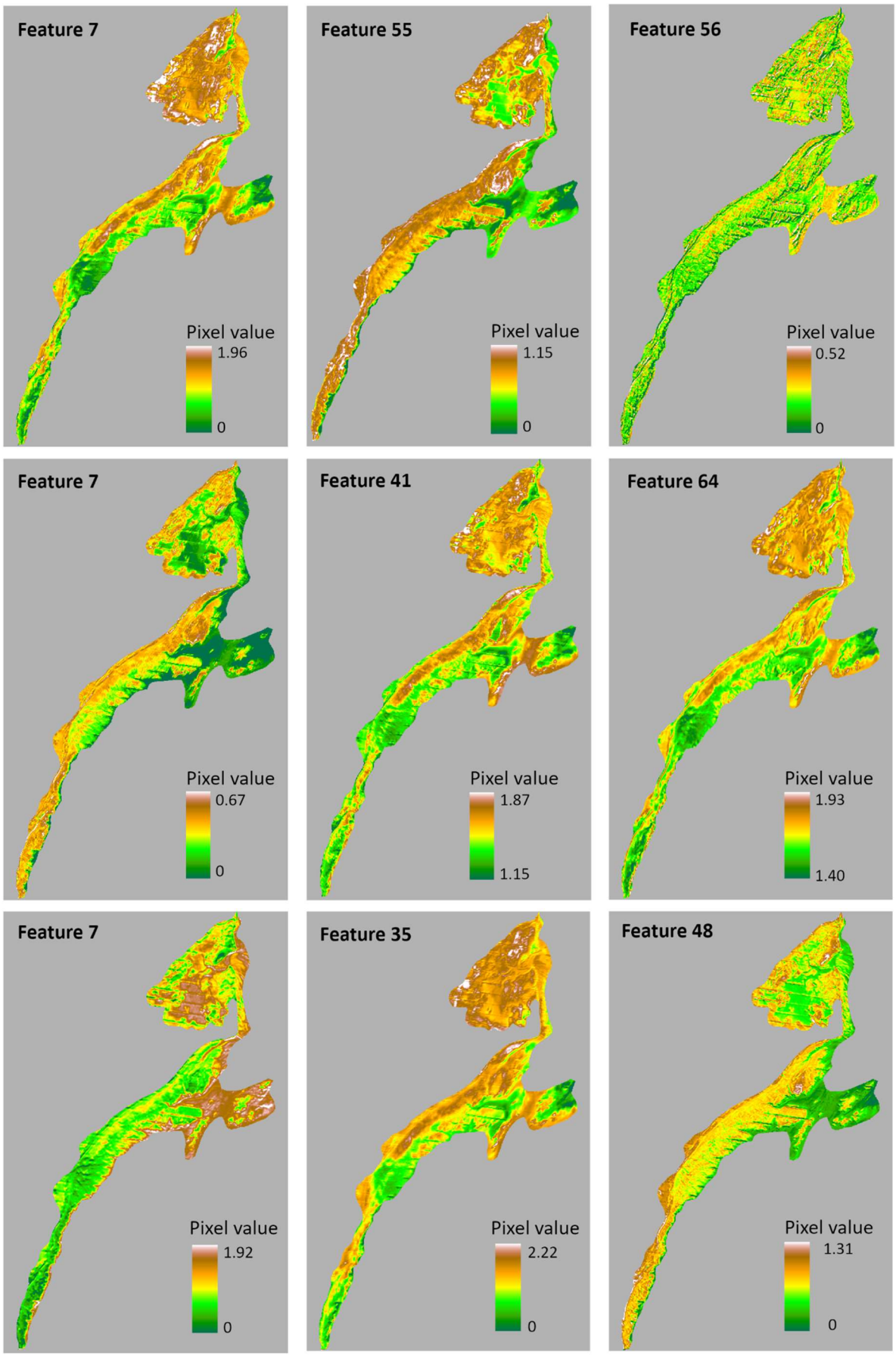

Figure 12 Features generate for the image collected on 14 February 2008 with top: scratch 64, middle: scratch 32, and bottom: freeze.

Table 6 Cross-validation accuracies (R2) for the different experiments.

Experiment	Median grainsize 2011	Silt 2011	Biomass (all) 2008	Richness (all) 2008	Biomass crab 2011-2015	Richness crab 2011-2015	Biomass crab nov. 2012
No change	**0.210**	**0.250**	0.186	**0.0382**	**0.0557**	**0.0356**	**0.141**
Finetuning	0.198	0.214	0.163	0.0372	0.0363	0.0254	0.120
Freezing	0.183	0.209	0.164	0.0381	0.0403	0.0253	0.126
Scratch 64	0.173	0.201	**0.190**	0.0326	0.0439	0.0331	0.0836
Scratch 32	0.0861	0.161	0.182	0.0317	0.0308	0.0171	0.113
No change 2015	0.0779	0.134	0.129	0.0903	-0.0119	-0.199	0.0725
Sentinel no change	0.134	0.113	0.0974	**0.0368**	**-0.0047**	-0.190	0.00486
Sentinel finetuning	0.149	0.0676	0.0893	0.0238	-0.0106	**-0.185**	0.0071
Sentinel freezing	0.152	0.063	0.0904	0.0226	-0.0102	-0.185	0.0072
Sentinel scratch 64	**0.180**	**0.158**	**0.1099**	0.0303	-0.01224	-0.204	**0.0147**
Sentinel scratch 32	0.154	0.097	0.0813	-0.0070	-0.0175	-0.196	-0.02041

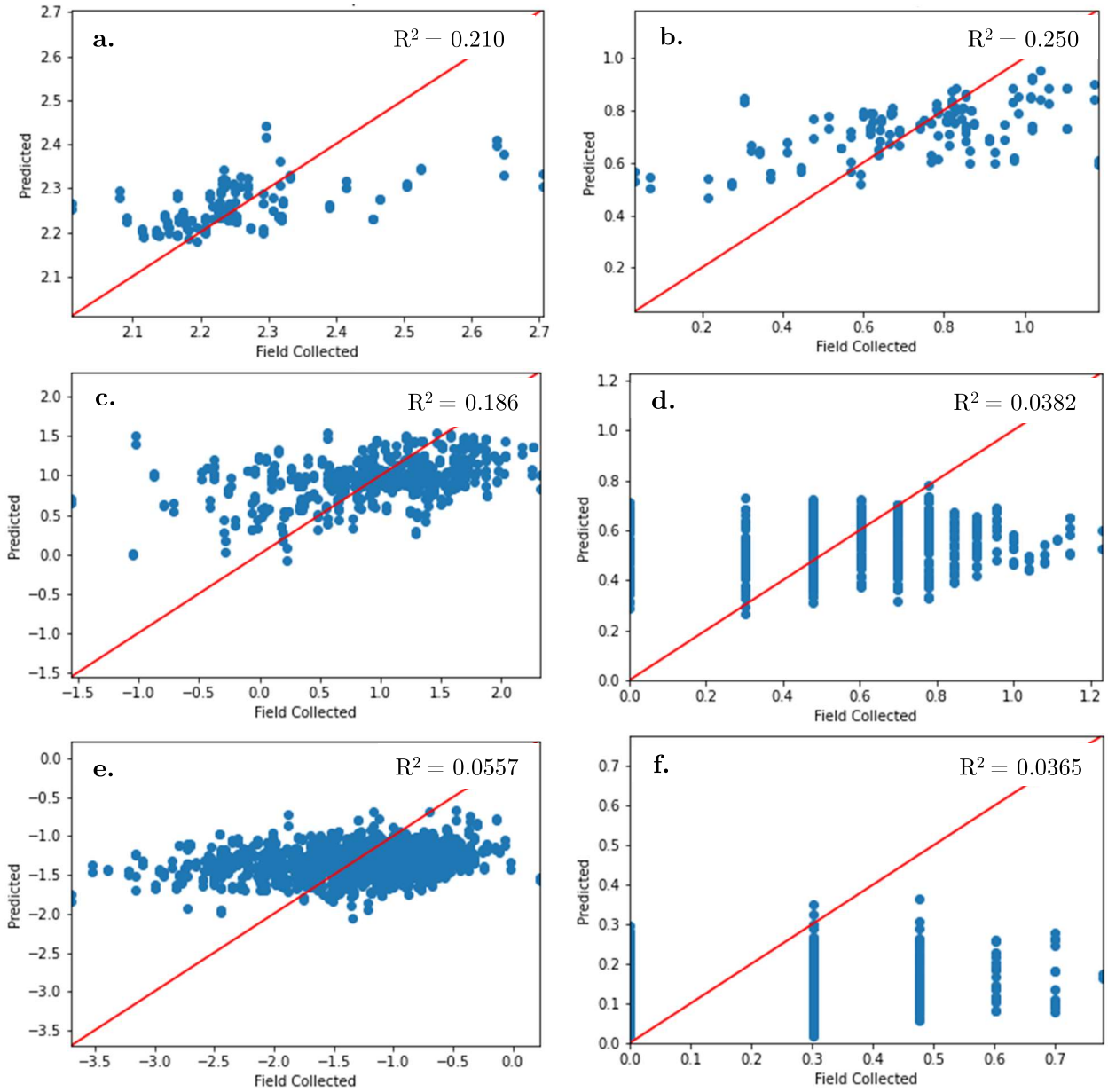

Figure 13 Predicted values against the field samples for the 'no change'-scenario. (a) Median grain size, (b) silt content, (c) full biomass, (d) full richness, (e) crab biomass, and (f) crab richness.

When comparing the predicted variables with the real field values it can be seen that the higher values are often underestimated by the model while the lower values are overestimated (figure 13). When plotted against the field values the predicted ecological variables form a horizontal cloud instead of the desired 1-to-1 line. For the sediment properties there seems to be a small positive trend with the predictions for the high field variables being higher than expected. However, the range of the predicted variables is still lower than the range of the field variables.

4.5.2 Cross-validation accuracy of the different scenarios

The 'no change'-scenario had the highest cross-validation accuracy for all but one variable. Only for the complete biomass collected in 2008 did the 'scratch 64'-scenario result in a higher cross-validation accuracy. The 'scratch 32'-scenario resulted in the lowest cross-validation accuracy for most variables. Only for the complete biomass and the November 2012 crab biomass did it reach higher

cross-validation accuracies than the ‘finetuning’- and ‘freezing’-scenario and the ‘scratch 64’-scenerio respectively.

4.6 Cross-validation accuracy Sentinel 2

The scenarios using Sentinel 2 images for the feature generation and random forest model resulted in lower cross-validation accuracies for almost all tested variables compared to when Landsat 7 images were used. On average the cross-validation accuracy was 8.67 percent points worse. The cross-validation accuracy reached by the Sentinel scenarios was lower than the highest accuracy reached by the Landsat scenarios for each variable. When comparing the cross-validation accuracies of the Sentinel scenarios to their Landsat counterpart only the accuracy of the median grain size from the ‘scratch 64’- and ‘scratch 32’-scenario was higher. For crab biomass and species richness the cross-validation accuracy reached negative R2 values meaning the model could not fit that data.

There was no scenario that had a convincingly higher overall cross-validation accuracy compared to the other scenarios. The ‘scratch 64’-model had the highest cross-validation accuracy for the sediment properties and complete biomass but had among the lowest accuracies for the other ecological variables. The scenario with the highest cross-validation accuracy on the Sentinel 2 image was the same as for the Landsat 7 images for the complete biomass, complete species richness, and crab biomass but differed for the other variables.

5. Discussion

5.1 Performance of transfer learning techniques

It was expected that the 'freezing’-scenario would have the highest predictive performance followed by the ‘finetuning’-scenario. Both these models were pre-trained on a related source dataset with higher quality images without data gaps while also getting the opportunity to learn region-specific information during finetuning on the target dataset. However, it was found that the pre-trained model without any finetuning resulted in higher cross-validation accuracies. As literal transfer without finetuning is also the most straightforward transfer learning method this would be promising to be able to apply these models without having to train it yourself.

When looking at the Landsat scenarios the ‘finetuning’- and ‘freezing’-scenarios did outperform the models trained from scratch for the sediment properties, complete richness, and November 2012 crab biomass. It is thus likely that the pre-training of the model did provide valuable information useful for predicting the sediment and ecological variable.

The ‘scratch 32’-scenerio had the worst performance. This scenario was included to increase the number of training patches. It was expected that this would result in a better trained model and better predictive results, at least compared to training from scratch with a patch size of 64x64. From the loss graphs this scenario seemed to indeed do best during training reaching much lower loss values compared to the models using less but larger patches. However, the cross-validation accuracy of this scenario was overall the lowest. Besides the lower accuracy of the prediction, the training of this model also took much more time.

On the Sentinel images the ‘scratch 64’-scenario outperformed the ‘no change’-model. The ‘no change’-scenario did have the second highest cross-validation accuracies. The ‘scratch 32’-scenario had again the lowest performance.

5.2 Landsat versus Sentinel

Because of the suboptimal image quality of the Landsat 7 image due to the SLC error Sentinel 2 images were also used to predict the environmental properties. As these images have a higher quality and the original model was based on these images it was expected that the predictive results would be higher or at least similar to the Landsat 7 images. However, it was found that the cross-validation accuracies obtained based on the Sentinel 2 image were much lower compared to those obtained based on Landsat 7. This was especially surprising for the 'no change'-scenario. While the other models were trained at least partly on Landsat 7 data the model used for the 'no change'-scenario was only trained on Sentinel 2.

An explanation could be that the temporal matching of the images and field data is important. For the Landsat scenarios the field data was predicted based on images with an acquisition time as temporally close as possible, often within the same seasonal year. For the Sentinel scenarios, however, the predictions for all variables had to be based on a single image collected in 2015. As a result, there exists a time gap between the field data collection and the image that is used to predict the field data. The complete biomass collected in 2008 for example had to be predicted based on the image from 2015, almost 7 years later.

This is partly supported by the results of the 'no change 2015'-scenerio. When similar to in the Sentinel 2 scenario only one Landsat 7 image, also taken in December 2015, was used the cross-validation accuracy was also reduced strongly. The results from this scenario are more similar to the Sentinel 2 scenarios. This suggests that indeed using an image that is not temporally close to the field data decreases the accuracy of the predictions based on it.

Over time environmental variables like sedimentary and ecological variables can change. They may even vary seasonally or change suddenly due to extreme events. Stormy seasons can for example be associated with erosion while sedimentation can take place during calmer seasons (Belliard et al., 2019). Tidal channels may also migrate on a time scale of a couple of years (Zhao et al., 2022). Biomass may also change seasonally (Beukema, 1974) or year-to-year (Beukema et al., 1993). When the conditions have changed between the sampling and the acquisition date of the satellite image this causes a mismatch between the data. This can then reduce the cross-validation of the random forest model.

5.3 Comparing to performance in the Wadden Sea

Similar to the findings of Madhuanand et al. (2023) the model shows a higher predictive performance for the sediment variables compared to the ecological variables. Overall, however, the model had a lower predictive performance for the tidal flats in Oman compared to in the Wadden Sea. For the sediment properties and species richness the highest cross-validation accuracy reached in Oman is lower than the lowest cross-validation accuracy reported for the same variable by Madhuanand et al. (2023). The highest cross-validation accuracies of the complete biomass did fall within the range found by Madhuanand et al. (2023). Crab biomass and richness were not predicted for the Wadden Sea but the 2012 crab biomass does fall within the cross-validation accuracy range for biomass given by Madhuanand et al. (2023).

For the Wadden Sea a difference in predictive performance between areas was already observed (Madhuanand et al., 2023). This difference in predictive performance was attributed to a difference in the distribution of the data. As the range and the mean of the predicted variables differ between Oman and the Wadden Sea. As a result, the distribution might also be different, and this might have affected the performance of the model.

For Oman the availability of field data is also limited. While for the Wadden Sea each field campaign had over 200 field samples for Oman only the field campaign of 2008 and November 2012 had over 200 field samples. For the sediment characteristic only 66 field samples were available to train and test the random forest model which could have limited the predictive performance. For the crab biomass and richness the field data of the different years was combined to increase the number of field data points. However, this reduced the temporal resolution of the field data. By combining different years differences between years can increase the noise included in the model reducing the prediction accuracy.

5.4. Generalization

There have been many experiments on different datasets that report positive gains from transfer learning. Mensink et al. (2012) for example performed multiple experiments comparing different pre-trained models with training from scratch and found that pre-trained models outperformed the models trained from scratch for all experiments. Similar results were also found by Neyshabur et al. (2020) where the pre-trained models were also able to reach higher accuracies compared to models trained from scratch. The results of this study are partly in line with these observations. For the Landsat scenarios the pre-trained model indeed reached higher accuracies than the models trained from scratch. However, this was not the case for all variables, especially when also comparing the finetuning and freezing scenarios. For the Sentinel scenarios the model trained from scratch reaches higher accuracies compared to the pre-trained scenarios for most variables. Thus, the gains from using a pre-trained model seem to be lower during these experiments than as described literature.

The decrease in the accuracy of a pre-trained when applied to a new target dataset compared to its accuracy on the source dataset has also been previously described. Research on land cover classifications found that when applying a model trained on a dataset from one continent to a different continent the prediction performance of the model decreased (Tong et al., 2021). This problem of generalization is thought to be caused by a difference in the spectral distribution between the training and target images. Seasonal changes can also influence the generalization of DL models (Tong et al., 2021). A model trained on one season can be expected to have a lower performance when applied to images of a different season.

5.5 Challenges and potential improvements

The existence of the SLC failure gaps makes the Landsat 7 images suboptimal. By combining images from two different moments in time with a different tidal elevation the edge stripes keep being detectable. Differences in the spectral data between the two years generate edges that are picked up by the model in the same way that spectral differences within one image for example between flats and channels get picked up. If an input image has clear stripes the autoencoder model will try to reconstruct this resulting in it learning to create features related to these stripes. Features generated by edge detection like filters for example often do not only show the edges of channels and flats but also the edges of these stripes. Indeed, the effect of the stripes was visible in the generated features that were used in the random forest model, for example in feature 56 in Figure 12. This can cause confusion for the random forest model since these edges are an artifact from the input images and do not reflect the environmental conditions in which the field variables were collected. As a result the accuracy of the model will likely decrease. In total 45% of the field data points were located on or within a 5-pixel radius from these data gaps, which strongly affected the prediction accuracies of the random forest model.

Comparing the results of the ‘no-change 2015’-scenarion and the ‘Sentinel no change’-scenario the effect of the quality difference between the Sentinel 2 and Landsat 7 images used does not seem very

clear. However, since both of these scenarios use a single image temporally separated from their data they are both suboptimal. As a result, further investigation into the effect of the quality of the satellite images would be advised. Furthermore, an evaluation of the performance of the model in a different area with higher-quality satellite image availability and a larger field data set would be recommended.

6. Conclusion

This study looked at the transferability of the deep learning method to predict sediment and ecological variables developed by Madhuanand et al. (2023) for the Wadden Sea. The method is comprised of a VAE deep learning model and a random forest model. The VAE model is trained using satellite images and produces features. The random forest model then combines these features with the original satellite image and field data to predict sedimentary and ecological variables. In this study the cross-validation accuracy of the model was evaluated for the tidal flats in Bar Al Hikman, Oman, using Landsat 7 and Sentinel 2 images and field data collected in 2008 and 2011-2015. The cross-validation accuracies for the variables median grain size, silt content, biomass, species richness, crab biomass, and crab species richness were evaluated. Different scenarios representing different transfer learning techniques were tested which differed in the training of the deep learning model used to extract features. The cross-validation accuracies obtained using literal transfer without change, finetuning, freezing and training from scratch were compared.

The best-performing scenario used the pre-trained model without any further training in the study area in Oman. The model trained from scratch with a patch size of 64x64 had the highest accuracies on the Sentinel images. Higher cross-validation accuracies were found when the random forest model was trained on single-year data and when images temporally close to the field campaign data were used.

Similar to the findings for the Wadden Sea the cross-validation accuracy was highest for the sediment properties. The cross-validation accuracy was lowest for the species richness. The cross-validation accuracies obtained for Oman were lower than those obtained by Madhuanand et al. (2023) for the Wadden Sea. This reduced accuracy when DL models are applied to a different region is in line with previous research on the generalization of deep learning methods.

The Landsat 7 images used in this study were suboptimal because of the data gaps caused by the SLC failure. These had to be filled with data from different images creating edges with spectral differences picked up by the deep learning model and visible in the generated features. As a result, they likely influenced the prediction accuracy.

While the cross-validation accuracies of the transfer attempts of this study were lower than preferred, indications have been found that using an improved dataset may result in better predictions. Focus should be on using satellite images without data gaps, using single-year data in the random forest model, and using temporally closely matched field and satellite data.

7. References

Bank, D., Koenigstein, N., & Giryes, R. (2023). Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, 353-374.

Belliard, J. P., Silinski, A., Meire, D., Kolokythas, G., Levy, Y., Van Braeckel, A., Bouma, T.J. & Temmerman, S. (2019). High-resolution bed level changes in relation to tidal and wave forcing on a

narrow fringing macrotidal flat: Bridging intra-tidal, daily and seasonal sediment dynamics. *Marine Geology*, 412, 123-138.

Beukema, J. J. (1974). Seasonal changes in the biomass of the macro-benthos of a tidal flat area in the Dutch Wadden Sea. *Netherlands Journal of Sea Research*, 8(1), 94-107.

Beukema, J. J., Essink, K., Michaelis, H., & Zwarts, L. (1993). Year-to-year variability in the biomass of macrobenthic animals on tidal flats of the Wadden Sea: how predictable is this food source for birds?. *Netherlands journal of sea research*, 31(4), 319-330.

Bom, R. A., de Fouw, J., Klaassen, R. H., Piersma, T., Lavaleye, M. S., Ens, B. J., Oudman, T., & van Gils, J. A. (2017). Food web consequences of an evolutionary arms race: Molluscs subject to crab predation on intertidal mudflats in Oman are unavailable to shorebirds. *Journal of Biogeography*, 45(2), 342-354.

Bom, R.A., Philippart, C.J.M., Van der Heide, T., de Fouw, J., Campuysen, C.J., Dethmer, K., Folmer, E.O., Stocchi, P., Stuut, J.B.W., Van der Veer, H.V., & Al Zakwani, I. (2018). Barr Al Hikman: a pristine coastal ecosystem in the Sultanate of Oman: Current state of knowledge and future research challenges. NIOZ.

Bom, R. A., van Gils, J. A., Molenaar, K., Kwarteng, A. Y., Victor, R., & Folmer, E. O. (2020). The intertidal mudflats of Barr Al Hikman, Sultanate of Oman, as feeding, reproduction and nursery grounds for brachyuran crabs. *Hydrobiologia*, 847, 4295-4309.

Burt, J. A. (2014). The environmental costs of coastal urbanization in the Arabian Gulf. *City*, 18(6), 760-770.

Choi, J. K., Eom, J., & Ryu, J. H. (2011). Spatial relationships between surface sedimentary facies distribution and topography using remotely sensed data: Example from the Ganghwa tidal flat, Korea. *Marine Geology*, 280(1-4), 205-211.

Compton, T. J., Holthuijsen, S., Koolhaas, A., Dekinga, A., ten Horn, J., Smith, J., Galama, Y., Brugge, M., Van der Wal, D., Van der Meer, J., Van der Veer, H., & Piersma, T. (2013). Distinctly variable mudscapes: distribution gradients of intertidal macrofauna across the Dutch Wadden Sea. *Journal of Sea Research*, 82, 103-116.

Costanza, R., De Groot, R., Sutton, P., Van der Ploeg, S., Anderson, S. J., Kubiszewski, I., Farber, S., & Turner, R. K. (2014). Changes in the global value of ecosystem services. *Global environmental change, 26*, 152-158.

Dissanayake, N. G., Frid, C. L., Drylie, T. P., & Caswell, B. A. (2018). Ecological functioning of mudflats: global analysis reveals both regional differences and widespread conservation of functioning. *Marine Ecology Progress Series*, 604, 1-20.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. *MIT press.*

He, K., Xiangyu, Z., Shaoqing, R., Jian, S. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition (CVPR),* 770-778.

Hossain, M. S., Bujang, J. S., Zakaria, M. H., & Hashim, M. (2015). Assessment of Landsat 7 Scan Line Corrector-off data gap-filling methods for seagrass distribution mapping. *International Journal of Remote Sensing*, 36(4), 1188-1215.

Hu, F., Xia, G. S., Hu, J., & Zhang, L. (2015). Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11), 14680-14707.

Iman, M., Arabnia, H. R., & Branchinst, R. M. (2021). Pathways to artificial general intelligence: a brief overview of developments and ethical issues via artificial intelligence, machine learning, deep learning, and data science. *Advances in Artificial Intelligence and Applied Cognitive Computing: Proceedings from ICAI'20 and ACC'20*, 73-87.

Iman, M., Arabnia, H. R., & Rasheed, K. (2023). A review of deep transfer learning and recent advancements. *Technologies*, 11(2), 40.

Ketkar, N., & Moolayil, J. (2021). Convolutional Neural Networks. *Deep Learning with Python*. *Apress, Berkeley, CA,* 197-242.

Koo, B. J., Kwon, K. K., & Hyun, J. H. (2005). The sediment-water interface increment due to the complex burrows of macrofauna in a tidal flat. *Ocean Science Journal*, 40, 221-227.

Koo, B. J., Kwon, K. K., & Hyun, J. H. (2007). Effect of environmental conditions on variation in the sediment-water interface created by complex macrofaunal burrows on a tidal flat. *Journal of Sea Research*, 58(4), 302-312.

LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R. (1998). Efficient BackProp. In: Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science,* 1524*.*

Lee, S., Park, I., Koo, B. J., Ryu, J. H., Choi, J. K., & Woo, H. J. (2013). Macrobenthos habitat potential mapping using GIS-based artificial neural network models. *Marine pollution bulletin*, 67(1-2), 177-186.

Levin, L. A., Boesch, D. F., Covich, A., Dahm, C., Erséus, C., Ewel, K. C., Kneib, R.T., Moldenke, A., Palmer, M.A., Snelgrove, P., Strayer, D., & Weslawski, J. M. (2001). The function of marine critical transition zones and the importance of sediment biodiversity. *Ecosystems*, 4, 430-451.

Li, H., Chaudhari, P., Yang, H., Lam, M., Ravichandran, A., Bhotika, R., & Soatto, S. (2020). Rethinking the hyperparameters for fine-tuning. *ArXiv preprint arXiv:2002.11770.*

Madhuanand, L., Philippart, C. J., Wang, J., Nijland, W., de Jong, S. M., Bijleveld, A. I., & Addink, E. A. (2023). Enhancing the predictive performance of remote sensing for ecological variables of tidal flats using encoded features from a deep learning model. *GIScience & Remote Sensing*, 60(1), DOI: 10.1080/15481603.2022.2163048

Mehanna, S. F., Khvorov, S., Al-Sinawy, M., Al-Nadabi, Y. S., & Al-Mosharafi, M. N. (2013). Stock assessment of the blue swimmer crab Portunus pelagicus (Linnaeus, 1766) from the Oman Coastal Waters. *International Journal of Fisheries and Aquatic Sciences*, 2(1), 1-8.

Mensink, T., Uijlings, J., Kuznetsova, A., Gygli, M., & Ferrari, V. (2021). Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 9298-9314.

Miloslavich, P., Bax, N. J., Simmons, S. E., Klein, E., Appeltans, W., Aburto-Oropeza, O., Andersen Garcia, M., Batten, S.D., Benedetti-Cecchi, L., Checkley Jr., D.M., Chiba, S., Duffy, J.E., Dunn, D.C., Fischer, A., Gunn, J., Kudela, R., Marsac, F., Muller-Karger, F.E., Obura, D., & Shin, Y. J. (2018). Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. *Global Change Biology*, 24(6), 2416-2433.

Murray, N. J., Phinn, S. R., DeWitt, M., Ferrari, R., Johnston, R., Lyons, M. B., Clinton, N., Thau, D., & Fuller, R. A. (2019). The global distribution and trajectory of tidal flats. *Nature*, 565(7738), 222-225.

Neyshabur, B., Sedghi, H., & Zhang, C. (2020). What is being transferred in transfer learning?. *Advances in neural information processing systems*, 33, 512-523.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.

Pires de Lima, R., & Marfurt, K. (2019). Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sensing*, 12(1), 86.

Pratt, L.Y. (1993). Transferring previously learning backpropagation neural networks to new learning tasks. *Rutgers The State University of New Jersey, School of Graduates Studies.*

Reise, K. (2012). Tidal flat ecology: an experimental approach to species interactions. *Springer Science & Business Media,* 54.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.

Scaramuzza, P., & Barsi, J. (2005). Landsat 7 scan line corrector-off gap-filled product development. In *Proceeding of Pecora,* 16, 23-27.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III. *Springer International Publishing*, 270-279.

Tong, X. Y., Xia, G. S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237, 111322.

Van der Wal, D. P. M. J. H. D., Herman, P. M. J., Forster, R. M., Ysebaert, T., Rossi, F., Knaeps, E., Plancke, Y.M.G. & Ides, S. J. (2008). Distribution and dynamics of intertidal macrobenthos predicted from remote sensing: response to microphytobenthos and environment. *Marine Ecology Progress Series*, 367, 57-72.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 1-40.

Willcock, S., Martínez-López, J., Hooftman, D. A., Bagstad, K. J., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G., Voigt, B., Villa, F., Bullock, J.M., & Athanasiadis, I. N. (2018). Machine learning for ecosystem services. *Ecosystem services*, 33, 165-174.

Yin, G., Mariethoz, G., & McCabe, M. F. (2016). Gap-filling of landsat 7 imagery using the direct sampling method. *Remote Sensing*, 9(1), 12.

Yoo, J. W., Hwang, I. S., & Hong, J. S. (2007). Inference models for tidal flat elevation and sediment grain size: a preliminary approach on tidal flat macrobenthic community. *Ocean Science Journal*, *42*, 69-79.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *Advances in neural information processing systems*, 27.

Zhao, B., Liu, Y., Wang, L., Liu, Y., Sun, C., & Fagherazzi, S. (2022). Stability evaluation of tidal flats based on time-series satellite images: A case study of the Jiangsu central coast, China. *Estuarine, Coastal and Shelf Science*, 264, 107697.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43-76.