


## Article

# On the generalization ability of data-driven models in the problem of total cloud cover retrieval

Mikhail Krinitskiy <sup>1</sup> \* , Marina Aleksandrova <sup>1</sup>, Polina Verezhenskaya <sup>1</sup>, Sergey Gulev <sup>1</sup>, Alexey Sinitsyn <sup>1</sup>, Nadezhda Kovaleva <sup>1</sup>, Alexander Gavrikov <sup>1</sup>

<sup>1</sup> Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia

\* Correspondence: krinitsky@sail.msk.ru; Tel.: +7-499-1247928

**Abstract:** Total cloud cover (TCC) retrieval from ground-based optical imagery is a problem being tackled by a few generations of researchers. The number of human-designed algorithms for the estimation of TCC grows every year. However, there is not very much progress in terms of quality, mostly due to the lack of systematic approach to the design of the algorithms, to the assessment of their generalization ability, and to the assessment of the TCC retrieval quality. In this study, we discuss the optimization nature of data-driven schemes for TCC retrieval. In order to compare the algorithms, we propose the framework for the assessment of the algorithms characteristics. We present several new algorithms that are based on deep learning techniques: a model for outliers filtering, and a few models for TCC retrieval from all-sky imagery. For training and assessment of data-driven algorithms of this study, we present the Dataset of All-Sky Imagery over the Ocean (DASIO) containing over one million of all-sky optical images of visible sky dome taken in various regions of the World Ocean. The research campaigns contributed to DASIO collection took place in the Atlantic ocean, the Indian ocean the Red and Mediterranean seas, and also in the Arctic ocean. Optical imagery collected during these missions are accompanied by standard meteorological observations of cloudiness characteristics made by experienced observers. We assess the generalization ability of the presented models in several scenarios that differ in terms of the regions selected for the train and validation subsets. As a result, we demonstrate that our models based on convolutional neural networks deliver superior quality compared to all previously published schemes. As a key result, we demonstrate considerable drop of the ability to generalize the training data in case of strong covariate shift between training and validation subsets of imagery which may take place in case of region-aware subsampling.

**Keywords:** Total cloud cover; all-sky camera; algorithms assessment; neural networks; machine learning; data-driven approach

## 1 Introduction

Clouds are considered playing one of the primary roles in climate regulation due to their impact on radiative [1–3] and latent heat fluxes. Clouds are also crucial for the hydrological cycle in both global and regional scales [4]. Cloud cover plays crucial role regulating climatic feedback, thus cloud cover may be exploited as a diagnostic in sensitivity studies of climate models in different scenarios. Cloud cover variability over the ocean is a key variable for understanding of regional climatic processes, e.g., monsoons, ENSO, ITCZ shift, NAO, PDO, and also multi-scale sea-air interactions in western boundary currents zones and upwelling zones [5,6].

There are a few data sources available for studies of clouds in the ocean. Among the most frequently used are remote sensing archives, reanalyses data, and the observations made at sea from research vessels and voluntary observing ships. Each of these data sources has its own advantages and flaws. Satellite observations may be considered accurate, and they are uniformly scattered spatially and temporally, though their time series are limited starting from the early 1980-s [7–12]. Satellite measurements are also characterized by different flaws, e.g., underestimating cloudiness over sea

ice under nighttime conditions [13] which may be crucial in the Arctic. Reanalyses data is uniformly sampled as well. However, although the models applied in reanalyses for diagnostic cloud cover estimation continuously improve, they need further development and validation [13,14]. Reanalyses were shown to underestimate total cloud cover compared to measurements provided by land-based weather stations, and observations over the ocean [13,15]. Possible cause of this underestimation may be the overestimated downward short-wave radiation that is taken into account within the schemes for cloud coverage computations [16].

The best data source for the climatological studies of clouds is the archives of observations made at voluntary observing ships (VOS) which are organized in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) [17,18]. The very first visual observations at sea were in the middle of the XIX century, though they are rare until the XX. Most studies use the ICOADS observations dated from the early 1950-s [19–21] due to the change of cloudiness codes in the late 1940-s [22]. This change of codes reduces the validity of climatic studies relying on long-term homogeneity of the time series of cloudiness characteristics over the ocean in XX century. The key disadvantage of ICOADS records is their temporal and spatial inhomogeneity. Most observations are attributed to the North Atlantic's and North Pacific's major sea traffic routes. In contrast, central regions of the Atlantic and the Pacific are not covered by the measurements tightly enough. The Southern ocean coverage is poor as well [23].

Visual observations of clouds considered the most reliable at the moment [24,25]. The observations over the ocean are conducted every three of six hours at UTC time divisible by 3 hours. This procedure provides four or eight measurement records a day per observing ship. Observed parameters include total cloud cover (TCC) and low cloud cover, morphological characteristics of clouds, and the estimate of cloud-base height. The total cloud coverage is estimated by a meteorology expert based on the visible hemisphere of the sky. Estimating the total cloud cover, the expert considers the temporal characteristics of the observed clouds along with their additional parameters, e.g. observed precipitation, preceding types of clouds, light scattering phenomena, *etc.* For TCC retrieval, the observer estimates visually the fraction of the sky dome occupied by clouds. This procedure is described in detail in WMO manual on codes [24], WMO guide to meteorological observations [? ], and in the International Cloud Atlas [25]. The procedure of the estimation of TCC implies that depending on the cloud types, one does or does not account for the sky gaps in the clouds into the sky fraction of the observed scene. For example, the gaps are accounted as clear sky in case of low clouds or convective clouds, e.g. cumulus and stratocumulus clouds. In contrast, the sky gaps are not taken into account in case of cirrus, cirrocumulus and almost all sub-types of altocumulus clouds. This feature of TCC estimation procedure introduces the uncertainty to the results of automated TCC retrieval schemes.

Cloud characteristics remain one of a few meteorological parameters subsets still observed by experts visually whereas most of the other indices are measured automatically today. The procedure is hard to automate due to the high amount of non-formalized rules of thumb and heuristics that are learned by an expert as a result of long-term practice. This way the expert estimating cloud characteristics adjusts his/her understanding of cloud formation processes, relates them to the state of clouds theory, and learns the correspondence of the observed visual scenes to the underlying physics dictated by the theory. The whole experience further results in somewhat consistent measure of quantitative characteristics that are recommended by WMO as the most reliable source of information about clouds. The flaws of the approach of visual estimation are obvious: it suffer from the subjectivity; the learning curve mentioned above may result in biased estimates; the approach itself is highly time-consuming and requires massive human resources even today, in the era of Artificial Intelligence and advanced computer vision.

In this study, we discuss mostly the problem of data-driven TCC retrieval, although the classification of observed clouds is also an intriguing problem addressed in a number of studies employing data-driven methods along with expert-designed and fused approaches [26–31].

## 1.1 On the optimization nature of known schemes for TCC retrieval from all-sky optical imagery

A number of automated schemes were proposed in latest 20 years, beginning with the pioneering work of Long et al. presented along with the optical package for the all-sky imagery retrieval in 1998 [32] and described in detail in 2006 [33]. Since the first scheme of Long et al., numerous variations of algorithms were described for estimating some of quantitative characteristics based on the different *in situ* measurements such as all-sky imagery [32,34–39] or downward short-wave radiation [40]. Most of these algorithms are designed by experts introducing their understanding of the physical processes resulting in the all-sky imagery similar to the one presented in 1. Given all-sky imagery acquired, in most simple cases, an index is calculated pixel-wise, e.g., red-to-blue ratio (RBR) in the series of papers of Long et al. [32,33,41] or in following studies [42–44], or the ratio  $\frac{B-R}{B+R}$  in [36,37,45], or even a set of indices [46]. Then, an empirical threshold is applied for the classification of pixels dividing them into two classes: "cloudy" and "clear sky" ones. A few schemes with more complex algorithm structure were presented lately [35,39,41,47,48]. These schemes were introduced mostly for tackling the flaw of simple yet computationally efficient schemes taking sun disk and circumsolar region of an all-sky image into account of cloudy pixels.

In contrast with expert-designed algorithms for TCC retrieval, just a few data-driven schemes were presented lately for estimating TCC [49–51] or for clouds segmentation in optical all-sky imagery [52]. The researchers may consider the problem of TCC retrieval being solved or to be too simple for being addressed with complex machine learning algorithms. However, none of the presented approaches demonstrated any kind of significant improvement in terms of the quality of TCC estimation.

The only exception here is the method presented by Krinitskiy [51] which was claimed to demonstrate almost human-like level of TCC estimation accuracy. There is however an error at the validation stage resulting in incorrect quality assessment. Current study may be considered corrigendum to the conference paper of Krinitskiy [51].

At this point we need to note, that the so-called data-driven methods for the approximation of some variable (say, TCC) do not differ considerably from the ones that are designed by an expert. In case of an expert-designed method, generally one introduces an understanding of the underlying processes that form the source data (say, optical ground-based imagery) and its features (say, relations between red, green and blue channels of a pixel registering clear sky or a part of a cloud). Then, these human-engineered features are used in some sort of an algorithm for the computation of an index or multiple indices which then are aggregated over the image to a quantitative measure. The algorithm may be considered simple [32,33,36,38,48], adaptive [39,47] or designed to be complex to some extent in attempt of taking some advanced spatial features into account [34,35,41,43,44] or in attempt of correcting the distortion of imagery or other features of imagery that are not consistent with the initial researcher's assumptions [36,43,44,46]. However, all these methods rely on the aggregation step at some point that is commonly implemented as a variation of thresholding. The threshold value(s) is(are) empirical and should be adjusted minimizing the error of TCC estimates or maximizing the quality of the method that is proposed in the corresponding study. This adjustment stage is commonly described vaguely [43] or briefly [33], though in this sense, the above mentioned expert-designed algorithms are essentially data-driven and inherently have optimization nature, thus the optimization is a key part of the development of these schemes.

In case of the schemes employing machine learning (ML) methods [49–51], the algorithms are inherently of optimization nature since the essence of almost any supervised ML algorithm is the optimization of empirical costs based on train data set.

In the context of this study, the mentioned train dataset consists of ground-based all-sky imagery with the corresponding TCC ("labels" hereafter) estimated by an expert in the field concurrently with the picture shooting. It is worth mentioning that this labeling procedure is subject to noise. There are multiple sources of noise and uncertainties in labels and imagery itself:

- Subjectivity of an observer. As mentioned above, the whole expert experience may impact the quality of TCC estimates. The uncertainty introduced by a human observer has not been assessed thoroughly yet. There is only hope that this uncertainty is less than 1 okta (one eighth of the whole sky dome, the unit of TCC dictated by WMO [24,25]), though from the subjective experience of the authors, the uncertainty may exceed 1 okta when one scene is observed by multiple experts.
- Violation of the observations procedure. Ideally, an expert need to observe the sky dome in the environment clear from obstacles, which may not always be the case not only in strong storming conditions at the sea, but even in case of land-based meteorological stations. In our study, every record made in hard-to-observe conditions was commented accordingly, thus no such record is used in the filtered train, validation and test datasets.
- Temporal discrepancy  $\Delta t$ ,  $s$  between the moment of an observation and the moment of corresponding imagery acquisition. This time gap can never be zero, thus there is always a room for the decision, which  $\Delta t$  is short enough for the expert records to be still correct for the corresponding all-sky image.
- Reduced quality of imagery. There is always a room for improvement in terms of the resolution of optical cameras, their light sensitivity and corresponding signal-to-noise ratio in low-light cases (e.g., for registering in nighttime conditions). The conditions of imagery acquisition may play its role as well, since raindrops, dust and dirt may distort the picture significantly, and may be considered strong noise in source data.
- Reduced relevance of the acquired imagery to the TCC estimation problem. It should be mentioned that in strongly waving conditions in the ocean, an optical package tightly mounted to the ship may partly register sea surface instead of sky dome.
- Reduced relevance of the acquired imagery to the labeling records. Ideally, TCC labeling site should be collocated with the optical optical package performing the imagery acquisition. This is not exactly the case in some studies [35].

All these factors may be considered introducing noise to the labels or imagery of the datasets that are essentially the basis of all the data-driven algorithms mentioned above. The impact of these factors may be reduced by modifying the observations and imagery acquisition procedures, or by data filtering. Some factors may be addressed by the releases of more strictly standardized procedures for cloud observations, though the WMO guide seems strict and straightforward enough [24,25].

However, there are factors that are unavoidable, that were not mentioned above. Typical types of clouds, and their amounts differ in various regions of the ocean. Typical states of the atmosphere and its optical depth differs significantly as well, strongly influencing the quality of imagery and its features. Statistically speaking, for the development of a perfect unbiased TCC estimator within the optimization approach, one needs to acquire a train dataset perfectly and exhaustively representing all the cases and conditions that are expected to be met at the inference phase. This requirement seems unattainable in practice. Alternatively, a data-driven algorithm is expected to have some level of generalization ability, thus being capable of inferring TCC for new images acquired in previously unmet conditions. The effect of significant changes in the source data is well known in machine learning and in the theory of statistical inference as covariate shift. The capability of an algorithm of generalization is called generalization ability. This ability may be expressed in terms of discrepancy between the quality of the algorithm being assessed on different datasets within one downstream task.

## 1.2 On the climatology of clouds

Climatology of clouds shows strong difference between characteristics of clouds in various regions of the World Ocean not only in terms of total cloud cover, but in terms of typical cloud types and their seasonal variability. In Aleksandrova et al. 2018 [23], the climatology of total cloud coverage for the period of 1950-2011 is presented.

The climatology is based on ICOADS [17,18] data and limited to periods from January till March, and from July till September. In tropics, the average TCC varies from 1.5 to 3 okta, whereas average

TCC in middle latitudes and in sub-polar regions are from 6.5 to 7.5 okta. In middle latitudes and sub-polar regions of the ocean, seasonal variability results in an increasing TCC in summer compared to winter seasons. In contrast, in tropics and subtropics, average TCC in summer is lower compared to winter periods, which is especially noticeable in the Atlantic ocean. The most striking seasonal variations are registered in the Indian ocean which is characterized by strong monsoon circulation. That is, in the northern regions of the Indian ocean, in winter dry seasons, mean seasonal TCC varies from 1 to 3 okta, whereas in wet summer seasons average seasonal TCC is from 5 to 6.5 okta. Time series of total cloud coverage for distinct regions of the World Ocean demonstrate significant difference in characteristics of the inter-annual variability of TCC. Cloud coverage in the southern Atlantic rarely drops under 80%-90%, whereas the inter-annual seasonally averaged TCC variability may exceed 50%. In the central Pacific, some years may be considered outliers with TCC significantly exceeding the mean values due to ENSO [21].

At the same time, averaged TCC values is not representative enough within the scope of our study. One needs to consider the regimes of cloudiness over the regions of the ocean. In the middle latitudes, both the regime of long-lasting broken clouds (4 to 6 okta), and the regime of interchanging short periods of overcast and clear-sky conditions may result in the same seasonal mean TCC. The diversity of the regimes of cloudiness may be shown by the empirical histograms of the fractional TCC for various regions of the World Ocean (see [23], fig. 5).

Types of clouds are not distributed evenly over the ocean as well. This feature strongly impacts the average TCC, and also the characteristics of the acquired imagery. The frequency of different types of clouds vary between tropics, mid-latitudes, and sub-polar regions. The difference between western and eastern regions of oceans is significant as well, especially when one considers low clouds [53].

Cumulonimbus denoted by two codes in WMO manual on codes ( $C_L9$  for *cumulonimbus calvus* and  $C_L3$  for *cumulonimbus capillatus*) [24] are observed frequently in tropics, and are rare in mid-latitudes and can almost never registered in sub-polar regions of the World Ocean. At the same time, there is a considerable difference even between the distributions of the two sub-types of cumulonimbus over the ocean: records of *cumulonimbus calvus* over the ocean is twice as frequent as the observations of *cumulonimbus capillatus*. There is also a significant difference in spatial distributions of these two sub-types: *cumulonimbus capillatus* are registered sometimes in the North Atlantic and Northern regions of the Pacific ocean, whereas *cumulonimbus calvus* are observed almost only in tropics and equatorial zone. *Cumulonimbus calvus* maximum frequency is attributed to central regions of the oceans, whereas *cumulonimbus capillatus* are more frequent in coastal zones.

*Cumulus* clouds include two WMO codes:  $C_L1$  for *cumulus humilis* and  $C_L2$  for *cumulus mediocris* or *cumulus congestus*. The frequency of *cumulus* clouds is high in western and central regions of subtropics and tropics of the oceans. In mid-latitudes, *cumulus* clouds are not that frequent in general, and even less frequent in summer. There are, however, exceptions of eastern part of the North Atlantic and northern regions of the Pacific ocean, where cold-air outbreaks are more frequent, thus the conditions for *cumulus* clouds are more favorable. Generally, *cumulus humilis* are less frequent over the ocean compared to *cumulus mediocris* and *cumulus congestus*.

In contrast with *cumulus* clouds, *stratus* clouds are much more frequent in mid-latitudes (WMO codes  $C_L5$  for *stratocumulus* other than *stratocumulus cumulogenitus*, and  $C_L6$  for *stratus nebulosus* or *stratus fractus* other than *stratus* of bad weather). They are also frequent in eastern regions of subtropics over the ocean. In some studies, these two codes are considered as one type [54], however, their spatial distributions differ considerably. *Stratocumulus* clouds ( $C_L5$ ) are registered most frequently in eastern regions of the oceans' subtropical zones, whereas *stratus* clouds ( $C_L6$ ) are mostly attributed to mid-latitudes, especially in summer. *stratus* clouds ( $C_L6$ ) are rare in eastern regions of subtropics of the oceans (excluding some of the upwelling zones). There are also *stratus fractus* or *cumulus fractus* of bad weather (WMO code  $C_L7$ ) which are frequently observed in mid-latitudes in winter, when the synoptic activity is strong. Sometimes, clouds of  $C_L7$  are registered in low latitudes, in the stratiform precipitation regions [55].



Sometimes there are even no low clouds (WMO code  $C_L0$ ). This code is frequently registered in the coastal region of the ocean, in the Arctic and in Mediterranean sea.

As one may notice from the brief and incomplete climatology of clouds above, different types of clouds are distributed strongly uneven over the World Ocean. In case one collects a dataset in a limited number of the regions of the ocean for the optimization of a data-driven algorithm, the resulting scheme may lack the generalization ability.

### 1.3 On the data-driven algorithms for TCC retrieval from all-sky optical imagery

A few data-driven methods were presented lately for estimating TCC from all-sky optical imagery [49,50,52]. In [50], the only improvement compared to simple schemes [33] is the application of clustering algorithm in the form of superpixel segmentation step. This step allows the authors to transform the scheme to an adaptive one. However, one still needs to compute the threshold value for each superpixel. In [52], a probabilistic approach for clouds segmentation is proposed employing Principal Components Analysis (PCA) approach along with Partial Least Squares (PLS) model. The whole approach may be expressed as a PLS-based supervised feature engineering resulting in the pixel-wise linearly computed index claimed to characterize the probabilistic indication of the “belongingness” of a pixel to a specific class (i.e. cloud or sky). For this index to be technically interpreted as a measure of probability, it is normalized to the  $[0, 1]$  range linearly. The model described in this study, is similar to logistic regression with the only reservation that log-regression model has strong probabilistic foundations resulting in both logistic function and binary cross-entropy loss function. The logistic function naturally transforms the covariates to the probability estimates within the  $[0, 1]$  range without any normalization. Thus, the model proposed in [52] has questionable probabilistic foundations compared to well-known logistic regression. However, the study [52] is remarkable being the first (as of our best knowledge) formulating the problem of cloud cover retrieval as a pixel-wise semantic segmentation employing a simple ML method. In [49], the authors employ the state of the art (at the time of the study) neural architecture namely U-net [56] for semantic segmentation of clouds in optical all-sky imagery. The two latter approaches are very promising in case one has a segmentation mask as a supervision. Worth mentioning that labeling of all-sky images in order to create cloud mask is very time-consuming. In our experience, this kind of labeling of one image may take 15 to 30 minutes of an expert depending on the amount of clouds and their spatial distribution. As of our best knowledge, no ML-based algorithms were presented that are capable of estimating TCC directly without preceding costly segmentation labeling.

One more issue of most of the presented schemes for TCC estimation is the lack of universal quality measure. In some studies, the quality measure is not even introduced [46]. Other studies with the problem formulated as a semantic segmentation of clouds, employ typical pixel-wise quality measures well-known from the computer vision segmentation tasks, such as Precision, Recall, F1-score, and misclassification rate [50,52]. This decision may be motivated by the models applied and by the state of the computer vision. However, the definition of a quality measure should never depend on the way the problem is solved. In some studies, the quality measures are used that are common for regression problems, e.g., correlation coefficient [49], MSE or RMSE [43]. In probability theory, these measures usually imply the specific assumptions about the distribution of the target value (TCC), and also the assumptions about the set of all possible outcomes and their type (real values). In case of TCC, these assumptions are obviously not met. It is also obvious that the assessment of the quality of TCC retrieval by any valid algorithm (that does not produce invalid TCC) is biased in case of the events labeled as 8 okta or 0 okta. Since the set of possible outcomes of TCC is limited, any non-perfect algorithm underestimates TCC for the 8-okta events and overestimates TCC for the 0-okta events. Thus, any quality metric is biased by design being calculated using the deviation of the result of an algorithm from the expert label. We are confident that one should never use a biased-by-design quality measure. Thus, in case one employs the quality measures of regression problems (MSE, RMSE,

correlation coefficient, determination coefficient, *etc.*), it would be consistent with solving the problem as regression, which is not always the case for the studies mentioned above.

In our understanding, the problem of TCC retrieval should be formulated as a classification since the set of possible outcomes is finite and discreet. Alternatively, one may formulate the problem as ordinal regression [57]. In these cases, accuracy (event-wise, rather than pixel-wise) or other quality measures of classification problem may be the right choice. In our study, we balance the datasets prior to the training and quality assessment, thus accuracy may be considered suitable metric. In case of ordinal regression, the categorical scale of classes is implied, which shows an order between the classes. It is exactly the case in the problem of TCC retrieval, since the classes TCC are ordered in such a way that the label "1 okta" denotes more clouds compared to the label "0 okta"; "2 okta" is more than "1 okta", and so on. In this case, the conditional distribution of the target variable  $P(TCC|event)$  is still not defined, which would be necessary for the formulation of a loss function and quality measures (MSE, RMSE *etc.*) within the approaches of Maximum Likelihood Estimator or Maximum a Posteriori Probability Estimator, similar to regression statistical models. However, the "less or equal than one-okta error accuracy" ("Leq1A" hereafter) is frequently considered as additional quality measure in the problem of TCC retrieval [43,49]. In our understanding, this metric is still not valid and may be biased due to the reasons given above, though we provide its estimates for our results to be comparable with other studies.

In the context of the introduction given above, the contributions of our study are the following:

1. We present the framework for the assessment of the algorithms for TCC retrieval from all-sky optical imagery along with the results of our models;
2. We present a novel scheme for estimating TCC over the ocean from all-sky imagery employing the model of convolutional neural networks within two problem formulations: classification and ordinal regression;
3. We demonstrate the degradation of the quality of data-driven models in case of strong covariate shift.

The rest of the paper is organized as follows: in Section 2, we describe our dataset of All-Sky Imagery Over the Ocean (DASIO); in Section 3.1, we describe the neural models we propose in this study and the design of the experiment for the assessment of their generalization ability; in Section 4, we present the results of the experiment. Section 5 summarizes the paper with the conclusions and provides an outlook.

## 2 Data

### 2.1 Dataset of All-Sky Imagery Over the Ocean (DASIO)

Since early 2000-s, we collect all-sky imagery over the ocean along with the concurrent expert estimates of a set of meteorological parameters recommended by WMO. Among the other parameters, our experts observe and register TCC, low cloud cover and other cloud characteristics. Starting from 2014, imagery acquisition was automated using the optical package designed and assembled in our laboratory [51]. We name it SAILCOP, which stands for "Sea-Air Interactions Laboratory Clouds Optical Package". SAILCOP is capable of acquiring the optical imagery of the visible sky dome. In fig. 1, one registering head of the package is presented along with the mounting points of both of the optical heads on our research vessel. In fig. 2, some examples of the imagery are presented. One may notice that the positions of the superstructures of a vessel visible in the images in fig. 2 are not the same due to variations of the mounting points of cameras, and also due to variations of the host vessel itself. The mounting points, however, are fixed within a mission, thus in case one would like to apply masking, the two masks (one per camera) are unique for a mission. The optical heads of SAILCOP are wired with the management computer. Each head is equipped with GPS sensor and

positioning sensors including accelerometer. As a result, SAILCOP is capable of taking pictures at the moments of nearly horizontal positioning of the vessel. Also, each pair of all-sky images is attributed with the GPS coordinates, date and time (UTC), and some additional technical information. In normal functioning regime, SAILCOP takes images synchronously from two heads with time discrepancy not exceeding 15ms. In fig. 3 (c,d) we present an example of the images acquired simultaneously from two cameras. One may see that they are almost the same, however, the two cameras were positioned in some distance apart from each other, thus the clouds are presented in these snapshots from slightly different angles. In some studies, this effect is exploited in the schemes for estimating cloud base height [58,59], though the distance between paired instruments in these studies may be considerably larger compared to linear scale of a research vessel. Typical period of imagery acquisition in SAILCOP is 20s. In tab. 1, we present the short description of the research missions resulting in DASIO collection. In Appendix A, we also present the complete maps of the missions.

DASIO collection obviously does not represent all the regions of the World Ocean. However, the regions of this dataset include western, central and eastern parts of Indian ocean as well as of Northern Atlantic; north-eastern, central and south-western regions of the Atlantic ocean, Arctic ocean, and the regions of the Mediterranean sea and the Red sea. Typical cloud types and cloud cover levels of these regions differ considerably as mentioned in Section 1.2. Thus using this dataset one may assess generalization ability of a data-driven scheme optimizing it on some subset of DASIO collection and estimating it on hold-out subset.

## 2.2 Data preprocessing and filtering

In this section, we present our approach for data preprocessing and filtering employed in this study. DASIO collection include some outliers as shown in fig. 2. There are also outliers caused by malfunctioning of the cameras, e.g., when the automatic algorithm for adjustment of white balance fails. This may happen in the early morning, or in low-light conditions typically observed concurrently with *stratocumulus* clouds in high latitudes. In our study, we filter the collection applying convolutional variational autoencoder model (CVAE) [60], see Section 3.1. This method allows us to filter out the snapshots that may be considered outliers. We inspect the results of the filtering to decide whether the images marked within our approach are outliers indeed. The criterion here is our expert understanding on whether the image represent its scene of sky dome the way that the major collection do. In Appendix A, we present a subsample of the images marked as outliers by using CVAE model. Some of the images are corrupted and may not be used as training examples, however, others may be considered valid. Expert inspection is still inevitable here.

For the application of the models of our study, we preprocess all the images of DASIO. First, we create a binary mask for each mount point of each mission. A mask is an image of equally sized as DASIO examples. The goal of using the masks is to mask the constructions of a vessel. The masks for some missions are presented in fig. 3 along with the corresponding imagery. The whole set of masks for the imagery is included in DASIO. We also reduce the size of snapshots in order to fit our computational resources. Original size of DASIO examples is 1920x1920 px. In our study, we resize the images to 512x512 px. The masks are resized respectively.

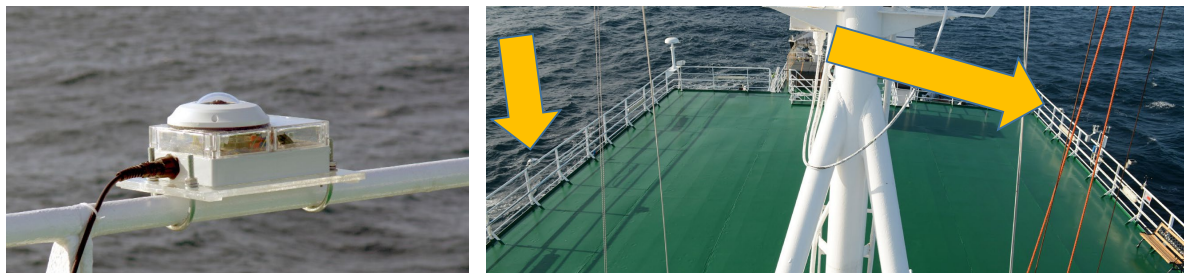
For improving generalization ability of the CVAE model and other data-driven models of our study, we apply augmentation of data. Several studies demonstrated the effectiveness of augmentation for the increasing of the generalization ability of data-driven models lately [56,61–69]. In some studies, even the approach is presented for trainable data augmentation which improves the training process [70–72]. In our study, at the stage of image data augmentation, we apply simple affine transformations along with weak elastic distortions described in [61]. Worth being noted that all the artificial neural networks of this study were trained on NVIDIA GPU using PyTorch framework [73]. However, for improving the computational speed, we stopped using the torchvision (a part of PyTorch project) implementations of imagery augmentations. We re-implemented all the transformations



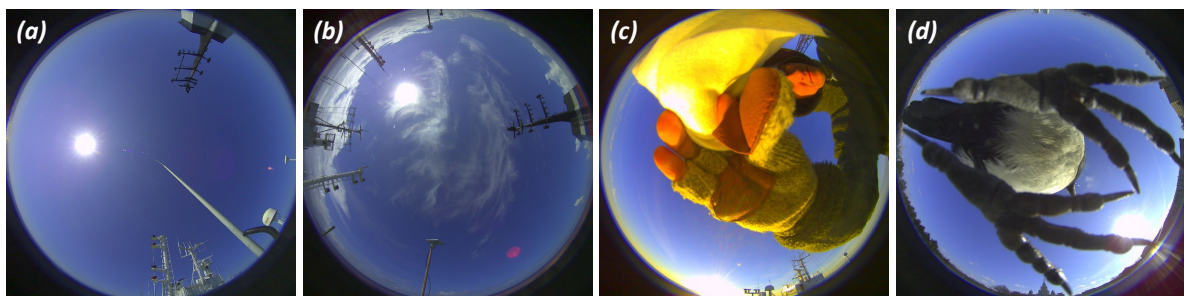
with pure PyTorch so the augmentation is effectively performed on GPU, which is not always the case at the moment for native torchvision. As a result decision, we observed 4X speedup of the training. The code for the augmentations is integrated into the code of our study, which is available on GitHub: <https://github.com/MKrinitskiy/TCCfromAllSkyImagery>. The distortions introduced at the augmentation time are stochastic by design: we sample the magnitudes of the affine and elastic transformations. However, for preserving the consistency of distorted imagery and the masks, we apply the same transformation to the masks as to the images.

Since we exploit artificial neural networks as the data-driven models in our study, we apply source data normalization recommended for the stabilization of training (see, e.g., [74]).

The sampling procedure is rather a part of the experiment design; thus, one may refer to Section 3 for the detailed description.



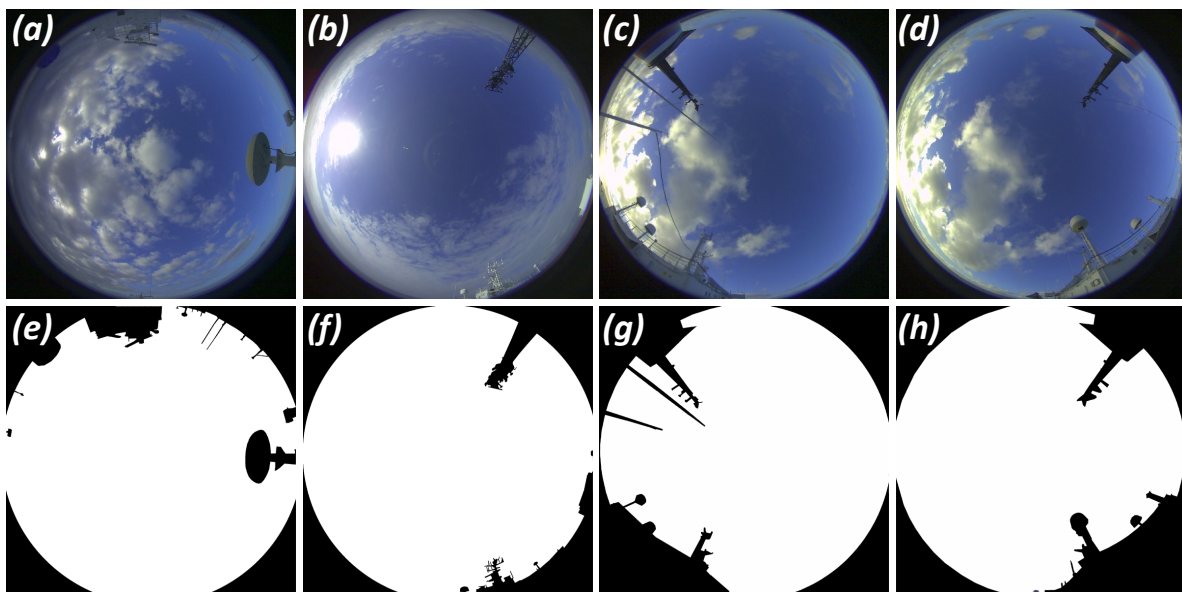
**Figure 1.** The optical package designed for the all-sky imagery acquisition: (left) registering head, and (right) positions of registering heads mounted on the research vessel "Akademik Ioffe" in the research mission "AI-49" in June 2016.



**Figure 2.** Examples of all-sky images acquired by the optical package SAILCOP (see fig. 1): (a),(b) - typical all-sky imagery over the ocean; (c),(d) - rare cases (outliers) when an image represents a scene with a bird or maintenance staff

**Table 1.** Scientific missions resulting in DASIO collection of all-sky imagery over the ocean with the corresponding expert records of meteorological parameters.

Mission name	Departure	Destination	Route
AI-45	17.09.2014 Reykjavik, Iceland	25.09.2014 Rotterdam, Netherlands	Northern Atlantic
AI-49	12.16.2015 Gdansk, Poland	02.07.2015 Halifax, Canada	Northern Atlantic
ANS-31	16.12.2015 Colombo, Sri Lanka	19.01.2016 Kaliningrad, Russia	Indian ocean, Red sea, Mediterranean sea, Atlantic ocean
AI-52	30.09.2016 Gdansk, Poland	03.11.2016 Ushuaia, Argentina	Atlantic ocean
ABP-42	21.01.2017 Singapore	25.03.2018 Kaliningrad, Russia	Indian ocean, Red sea, Mediterranean sea, Atlantic ocean
AMK-70	05.10.2017 Arkhangelsk, Russia	13.10.2017 Kaliningrad, Russia	Northern Atlantic, Arctic
AMK-71	24.06.2018 Kaliningrad, Russia	13.08.2018 Arkhangelsk, Russia	Northern Atlantic Arctic
AMK-79	13.10.2019 Kaliningrad, Russia	05.01.2020 Montevideo, Uruguay	Atlantic ocean Arctic



**Figure 3.** Examples of all-sky imagery (a...d) with the corresponding computation masks (e...h)

### 3 Machine learning models and experiment design

In this section, we present the design of the experiments of our study. We also present the competing architectures of artificial neural networks we employed in the problem of TCC retrieval from all-sky optical imagery over the ocean.

### 3.1 Filtering outliers with Convolutional Variational Autoencoder (CVAE)

Our solution of the problem of outliers filtering rely strongly on the assumption that most of the typical examples of train dataset belong to some compact manifold in some feature space. Intuitively, one may expect a considerable difference between the cardinality of the set of all the possible RGB images of size 1920x1920 ( $\sim 255^{10^7}$ ) and the amount of meaningful all-sky photographs which obviously will not include not only meaningless chaotic images, but also real-world objects, persons, scenes of kinds other than all-sky, *etc.* The assessment of the belongingness of an example to the manifold of a dataset may be performed in terms of some distance measure, e.g., Euclidean distance in the original feature space described by R,G and B components of all the pixels of an image. The dimensionality of this original feature space is  $\sim 10^7$  in case of DASIO examples. Thus, an effect of high dimensionality of the examples takes place which results in insignificant differences of distances between the examples that are close to each other ("seem similar" as images) or spaced apart ("seem dissimilar" as images). This effect is known as "curse of dimensionality". One way for tackling this effect is dimensionality reduction. The requirements for this mapping are simple: (i) it should preserve the relations between the examples of training dataset (similar images should be projected to the points of a new feature space close to each other); (ii) at the same time, if the examples happened to be mapped close to each other in a new feature space, they should appear similar.

Artificial neural network is essentially a function that maps objects from one feature space (e.g., images) to another, so-called feature space of hidden representations. There is a neural model capable of reducing dimensionality of examples without losing much of meaningful information of the examples, namely autoencoder (AE) [75–77]. Since the examples of our study are images, we exploit convolutional autoencoders. An autoencoder generally includes two functional parts: an encoder and a decoder. The encoder transforms the examples extracting meaningful features and mapping the examples into hidden representation feature space. The dimensionality of this feature space is commonly lower than the original dimensionality of the examples. It is the case in our study. The decoder part decodes (reconstructs) the examples based on their hidden representations. The technical task for an autoencoder is to reproduce the examples with the lowest errors. Training an autoencoder is no different from any other artificial neural network or other statistical models: one exploits a gradient-based optimization procedure for optimizing the loss function in the space of parameters of the neural network. In our study, we employ MSE as a reconstruction loss for the autoencoder model.

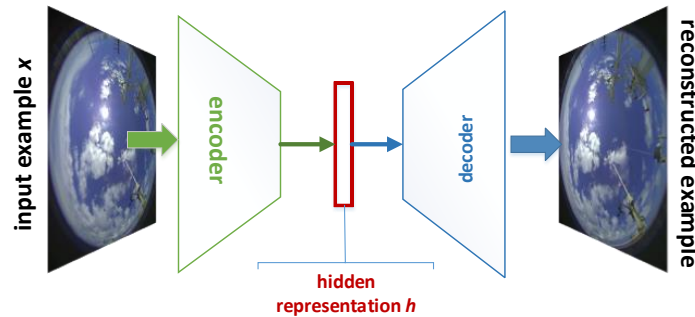
Though autoencoders are applied widely for anomaly detection, an autoencoder in its simple form (a.k.a. "vanilla" autoencoder, which is jargon, though allows to identify the one) is not enough for filtering outliers. Just the second requirement is met in case of "vanilla" autoencoder: examples that are close to each other in the hidden representation feature space, are similar. However, the opposite is not true: subsets of similar examples may be projected into clusters spaced apart. In our study, we overcome this issue by exploiting variational autoencoders [60] in the form of convolutional variational autoencoder (CVAE). Variational autoencoders were shown to find a special kind of mapping that preserves the continuity of the feature space, meaning that the first requirement is met: similar examples are projected to the close points of the hidden representation feature space. This property is achieved by introducing the assumption that each feature of the hidden representation of the dataset's examples is distributed normally. Technically this assumption can be met by using an additional loss component of KL divergence between the sample distribution of hidden representation features and normal distribution. The loss component and a technical way for implementing this approach (known as "reparameterization trick") were proposed in [60]. Resulting loss function is presented in eq. 3 and includes reconstruction loss MSE along with the KL term. In fig. 4, we present the general architecture of the CVAE implemented in our study.

$$h_i = \mathcal{E}(x_i, \theta_e), \quad (1)$$

$$\hat{x}_i = \mathcal{D}(h_i, \theta_d), \quad (2)$$

$$\mathcal{L}(x_i, h_i, \hat{x}_i) = \text{MSE}(x_i, \hat{x}_i) + \text{KL}(h_i), \quad (3)$$

where  $h_i$  is hidden representation of an example  $x_i$ ;  $\mathcal{E}(\cdot)$  and  $\mathcal{D}(\cdot)$  are encoder and decoder parts of the autoencoder;  $\theta_e$  and  $\theta_d$  are parameters of the encoder and the decoder respectively;  $\mathcal{L}(\cdot)$  is CVAE loss function.



**Figure 4.** General architecture of the CVAE model for outliers filtering. See more details in Appendix A.

In our study, we employ several approaches for improving the convergence of artificial neural networks. We employ Adam optimizer [78] with learning rate scheduling strategy named SGDR (stochastic gradient descent with warm restarts) presented in [79] that implies cosine annealing of learning rate with warm restarts. In addition to SGDR we apply exponential decay of the maximum learning rate. Resulting learning rate curve and typical training loss curve are presented in Appendix A in fig. A3.

### 3.2 Neural models for TCC retrieval from all-sky imagery

Convolutional neural networks (CNNs) demonstrated a huge leap forward in image recognition [62,63,80], semantic segmentation [56,81–84] and other visual tasks. In most of the problems related to image processing, CNNs are capable of achieving the highest quality with a gap unbridgeable by classic computer vision approaches. In some of the problems, CNNs reach human-like or even super-human quality today. Being data-driven models, CNNs seem to be perfect candidates for the application in the problem of TCC retrieval from all-sky imagery. In our study, we propose an advanced architecture of CNN motivated by latest results in the field of neural architectures. In contrast with a number of contemporary studies, we were looking for a state of the art CNN architecture demonstrating best results in a range of benchmark visual problems. In <https://paperswithcode.com/>, a perfect up-to-date collection of studies is presented. Although the resource is not of academic sort, it does not provide a comprehensive overview and does not perform benchmarking itself, it is still a tool for the selection of best known approaches at the moment. Not to mention the set of benchmarks like MNIST [85], CIFAR-10/100 [63], Imagenet [86] and up to 165 others that are commonly considered by researchers when testing the proposed approaches and model architectures. The approaches presented in state of the art studies at the moment are of several types: data augmentation strategies, strategies of training of neural networks, approaches for neural architecture search, meta-learning, and a set of approaches that pursue a goal of decreasing the computational overhead of neural networks without too much loss of the quality. One of the best neural designs we came up with, studying state of the art architectural solutions, is the PyramidNet proposed in [87]. Key contribution of this work is the conscious design of the layers resulting in computational efficiency of the PyramidNet along with increased quality compared to other architectures with similar computational costs [88,89]. In our



study, we tested both ResNet-like architecture and the one based on PyramidNet. No statistically significant difference was found. In this paper, we do not demonstrate the results of the ResNet-like architecture as thorough comparison of different architectures in the problem of TCC retrieval from all-sky imagery is not the topic of this study.

### 3.2.1 TCC retrieval as clasification

As mentioned in Section 1.3 we consider promising and valid two formulations of the TCC retrieval problem. The first and very straightforward problem design is classification since the set of possible outcomes is finite and discreet. In this case, the target variable (TCC) is one-hot-encoded, that is, transformed into the form of a row-vector of  $K$  elements, where  $K$  is the number of classes of the problem. All elements of this vector are zero, and just one element of this vector equals 1, that corresponds to the label number. For example, the target vector for an image labeled as  $TCC = 3$  okta will be the following:  $[0, 0, 0, 1, 0, 0, 0, 0]$ . As a result of this transformation, the one-hot-encoded target vector represents probability distribution of a multi-variate target variable with an assumption of Bernoulli distribution for each of the components. In this formulation, a neural network approximates the parameters  $p$  of Bernoulli distribution for each of the components of target vector. In the Maximum Likelihood Estimation approach, this sequence of assumptions results in loss function of multinomial cross-entropy (see eq. 6). Note that a neural network that maps features of examples into the feature space of one-hot-encoded target variable is just a parametric function  $\mathcal{F}(x_i, \theta_{NN})$ , where  $\theta_{NN}$  are the parameters of the network. Worth mentioning that the architecture of the neural network is not dependent on the loss function. Thus, one may consider the same architecture with different formulation of the problem resulting in different loss function and different behavior of the network.

$$h_i = \text{SoftMax}(\mathcal{F}(x_i, \theta_{NN})), \quad (4)$$

$$\text{SoftMax}_k(z) = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}}, \quad (5)$$

$$\mathcal{L}_{PC}(h_i, t_i) = - \sum_{j=1}^K t_{ij} * \log(h_{ij}), \quad (6)$$

where  $x_i$  is the image raw data;  $\text{SoftMax}(\cdot)$  is the function transforming a vector to the form meeting the requirements to represent a probability distribution (e.g., summation to one over the vector);  $h_i$  is the estimates of parameters of Bernoulli distribution of one-hot-encoded target variable TCC denoted as  $t_i$  (also known as "ground truth");  $K$  is the number of classes ( $K = 9$  for TCC retrieval following the recommendations of WMO). The subscript PC regards the formulation particularity which is "Pure Classification" here. When a network is trained, the estimates  $h_{ij}$  may be regarded as a measure of probability of an event described by features  $x_i$  to be of class  $C_j$ , where  $\{C_j\}_{j=1}^K$  is a set of classes of the problem.

The network described above is referred hereafter as PNetPC which stands for **P**yramid**N**et for **P**ure **C**lassification. The approach of classification using artificial neural networks is not novel; on the contrary, classification is one of the most common problems in machine learning. The novelty of the presented approach is the formulation of the TCC retrieval problem as a classification.

### 3.2.2 TCC retrieval as ordinal regression

As an alternative, we consider the TCC retrieval problem as ordinal regression, since TCC classes has natural order. There are a few competing approaches for solving ordinal regression with artificial neural networks [90–96]. As mentioned above, the architecture of a network does not depend on its loss function. It will define the behavior of the network rather than its architecture. Thus, one may apply any of the approaches mentioned in the studies [57,90–96] with own network architecture. In



our study, we exploit PyramidNet within the approach of ordinal regression. Following the survey [57] and the original paper [91], we implemented our neural model for solving the problem of TCC retrieval within the approach of ordinal regression in the form with no assumptions made on the distribution of target value. Alternatively, one may consider the problem to be soft classification: if an object  $x_i$  is labeled as class  $C_j$ , then it is inherently labeled as class  $C_{j-1}$  and all the other lower classes until the  $C_1$ . Technically, this formulation implies just two minor changes: a new encoding of target variables, and a new loss function. The new encoding implies that all the elements of target vector of an example labeled as  $C_j$  are equal to 1 in positions  $j$  and lower. For example, the target vector for an image labeled as  $TCC = 3$  okta will be the following:  $[1, 1, 1, 0, 0, 0, 0]$ . As the problem is considered soft classification, each of the elements of target vector considered an independent random variable with Bernoulli distribution. Thus, loss function in this case will be just the sum of individual binary cross-entropy loss functions for all the  $K$  components of target vector as shown in eq. 8.

$$h_i = \mathcal{F}(x_i, \theta_{NN}), \quad (7)$$

$$\mathcal{L}_{OR}(h_i, t_i) = - \sum_{j=1}^K (t_{ij} * \log(h_{ij}) + (1 - t_{ij}) * \log(1 - h_{ij})), \quad (8)$$

where the notation is the same as in eq. 6. The subscript OR regards the formulation particularity which is "Ordinal Regression" here. Note that there is no  $SoftMax(\cdot)$  in this case. However, each of the independent components of the target vector should still represent a Bernoulli parameter estimate, thus the activation function of the very last layer of the network  $\mathcal{F}(x_i, \theta_{NN})$  needs to be sigmoid  $\sigma(\xi) = \frac{1}{1+\exp(-\xi)}$  or similar alternative (e.g.,  $\tanh(\cdot)$  normalized accordingly).

The network described above is referred hereafter as PNetOR which stands for **PyramidNet** for **Ordinal Regression**. As of our best knowledge, the problem of TCC retrieval has never been solved using any data-driven models within the approach of Ordinal Regression. As one may see in Section 4, ordinal regression delivers supreme quality compared to classification.

### 3.3 Experiment design

In this section, we propose the framework for the assessment of the quality and generalization ability of data-driven models in the problem of TCC retrieval. Given a model capable of estimating TCC for an expert-labeled all-sky image, one may compare the expert label with the model estimate. As mentioned in Section 1.3,  $MSE$ ,  $MAE$ , correlation coefficient or determination coefficient are questionable quality measures for data-driven models in the problem of TCC retrieval. We propose measuring the quality with accuracy (eq. 9) in case of balanced validation dataset.

$$Acc = \frac{\sum_{i=1}^N [T\hat{C}_i = t_i]}{N}, \quad (9)$$

where  $N$  is the number of examples of validation dataset;  $T\hat{C}_i$  is model estimate of TCC;  $t_i$  is the expert-defined label for TCC (ground truth). In addition to accuracy, one may assess the quality using a common measure "less or equal than one-okta error accuracy" ( $Leq1A$ , see eq. ??). However, as mentioned in Section 1.3, this measure may be biased by design of the problem.

$$Leq1A = \frac{\sum_{i=1}^N [|T\hat{C}_i - t_i| \leq 1]}{N}, \quad (10)$$

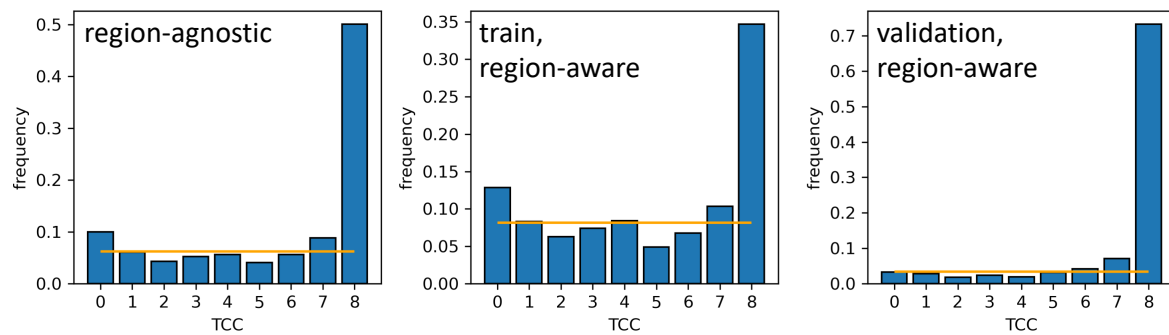
DASIO collection is strongly unbalanced: the number of examples labeled with 8 okta is approximately 50% (see region-agnostic TCC distribution in fig. 5). In our study, all the subsets (training and validation) are balanced the following way: we randomly subsample the subset of the examples labeled as 8-okta resulting in the number of them equals to the average number of the rest

classes (marked with orange lines in fig. 5). The other classes are subsampled or oversampled to reach the same numbers of examples. This target-balancing procedure is applied in all the experiments described further.

For the comparison of different data-driven algorithms in the problem of TCC retrieval, we propose training them on the same subset, and assess their quality on the same validation subset of imagery. In our study, we applied this procedure for comparing the performance of the two proposed models, namely PNetPC and PNetOR. In this scenario, we do not impose any restrictions on the regions where the data were collected for train and validation subsets. Note, however, that objects of DASIO collection may be correlated when being close in time. In time series analysis, there is a known effect of naturally correlated examples that are temporally related. This is especially the case when the imagery is taken with the period of order of 20s, which is considerably shorter compared to the typical period for decorrelating of spatial cloud characteristics that was estimated to be  $\sim 15$  minutes [58]. Thus, the common data science procedure of completely random train-validation split will produce strongly correlated (i.e., not independent) subsets, and thus the estimates of the quality will be too optimistic at the end of the day. This effect influenced strongly the results presented in [51]. Instead of completely random subsampling, one needs to employ the procedure of block-split. In our study, we consider each day of observations as a block of data. This way, we overcome the issue of auto-correlated time series of TCC and all-sky imagery. DASIO collection contains 250 days of observations so far, thus there are 250 daily blocks. In case of non-restricted splitting procedure, we split these 250 blocks on train and validation subsets. As the fractions of classes may vary depending on the split realization, we did our best to select the split that produces the most equally distributed train and validation subsets. Particularly, we sample the blocks for subsets with subsequent computation of KL divergence between the TCC distributions produced. After 10000 sampling attempts, we consider the best split that results in lowest KL divergence. This way, the fractions of TCC classes are almost the same in train and validation subsets. Although, we still cannot guarantee the absence of covariate shift this way, at least we guarantee the absence of target variable shift.

For the assessment of the generalization ability of data-driven models, we propose the comparison of the quality estimates estimated within the region-agnostic approach described above with the ones achieved with a different strategy of train-validation split. The region-aware approach implies the restrictions on the regions of imagery acquisition. In this scenario, strong covariate shift is expected, as mentioned in climatology of clouds in Section 1.2. In particular, we train the models on the subset limited by latitudes from  $45^{\circ}\text{S}$  till  $45^{\circ}\text{N}$ . In this scenario, we assess the quality on the data subset acquired to the north of  $45^{\circ}\text{N}$ . As one may see in the maps of the missions contributed to DASIO collection, the regions selected for the training set this way include central Atlantic, Mediterranean and Red seas, and all the observations in the Indian ocean. The regions selected for the validation set include northern Atlantic (the section along the  $60^{\circ}\text{N}$ , and coastal regions) and Atlantic sector of Arctic ocean. In fig. 5, the distributions of TCC values are presented for region-agnostic and region-aware scenarios.

Each configuration of data-driven models (PNetPC, PNetOR) is trained and evaluated multiple times (typically 5 to 9 due to high computational costs of CNNs) for estimating the uncertainty of the quality measures. In this section of our study, we employ the same approaches as in CVAE section 3.1 for improving the convergence of artificial neural networks. We use Adam optimizer [78] with SGDR learning rate scheduling strategy [79] and exponential decay of the maximum learning rate. Resulting learning rate curve and typical learning curves for PNetOR are presented in Appendix A in fig. A4. We do not show learning curves for PNetPC in this paper. They are similar to the ones of PNetOR with the reservation that PNetOR delivers slightly better quality as one can see in the Results section 4.



**Figure 5.** Histograms of the empirical distributions of TCC in different scenarios: region-agnostic (train and validation distributions are the same); and region-aware. In all the three cases, orange lines denote the average level for classes of 0-7 okta that is used for target-balancing each of the subsets; 8-okta class is downsampled to this level in accordance with the procedure proposed in Section 3.3.

## 4 Results and discussion

In this section, we present the results of the numerical experiments described in Section 3. For each of the data-driven models proposed in this study, we perform the training and quality assessment several times (typically 5 to 9) for estimating the uncertainty of the quality measures. The results in this section are presented in the following manner: first, in tab. 2 we present the results of PyramidNet-based models PNetPC and PNetOR in the region-agnostic scenario for assessing the influence of the problem formulation. Then, we in tab. 3 we present the results of PNetOR model in both region-agnostic and region-aware scenarios for assessing the generalization ability in case of strong covariate shift.

**Table 2.** Quality estimates for PNetPC and PNetOR data-driven models in region-agnostic scenario.

Model architecture	$Acc_{train}$	$Acc_{val}$	$Leq1A_{train}$	$Leq1A_{val}$
PNetPC	$46.52 \pm 4.6 \%$	$41.43 \pm 2.3 \%$	$86.52 \pm 1.87 \%$	$85.7 \pm 2.1 \%$
PNetOR	$46.44 \pm 0.38 \%$	$42.38 \pm 0.97 \%$	$88.4 \pm 0.23 \%$	$84 \pm 0.18 \%$

From the results presented in tab. 2, a weak superiority of PNetOR is noticeable. We need to mention, however, that this superiority is not statistically significant at least in the case of this small number of runs. Also, in terms of  $Leq1A$ , PNetPC is slightly better than PNetOR. It worth mentioning also that state of the art expert-designed schemes for TCC retrieval demonstrate considerably lower level of quality: in no fair conditions,  $Acc$  for them exceeds 30% [43,51].

**Table 3.** Quality estimates for PNetOR in different scenarios.

Scenario	$Acc_{train}$	$Acc_{val}$	$Leq1A_{train}$	$Leq1A_{val}$
Region-agnostic	$46.44 \pm 0.38 \%$	$42.38 \pm 0.97 \%$	$88.4 \pm 0.23 \%$	$84 \pm 0.18 \%$
Region-aware	$48.35 \pm 0.72 \%$	$36.56 \pm 0.5 \%$	$90.03 \pm 0.42 \%$	$77.44 \pm 0.64 \%$

From the results presented in tab. 3, one can see a considerable difference in the gaps between train accuracy and its validation estimate in different sampling scenarios. In the region-agnostic scenario, the gap is  $\sim 4\%$ , whereas in the region-aware scenario, the gap is  $\sim 12\%$ . We need to remind that this gap is inevitable, and only a perfect model would deliver the same quality on validation set as on train set (sometimes even better in some special cases). However, as we mentioned in Section 3.3 (Experiment design), we assess the capability of a data-driven model to generalize in terms of these gaps. Meaning, if the gap increases significantly, then typical covariate shift between regions is too

strong and the combination of a model flexibility and data variability forces a researcher to collect more data for the model to be applicable in a new region with some level of confidence. In the presented cases, a significant difference is noticeable.

One possible reason for that may be that strong disproportion of TCC classes in region-aware validation subset. One may notice that an automatic scheme may achieve an outstanding quality just predicting 8 okta every time. However, quality estimate in this case would be absolutely unreliable. Worth mentioning that even in this worst case, the CNN-based approach presented in our study is still superior compared to previously published results.

## 5 Conclusions and outlook

In this study, we present the dataset of all-sky imagery over the ocean (DASIO) collected in various regions of the World Ocean. Some of the imagery are attributed by cloud characteristics observed *in situ* by experts. We demonstrate strong covariate shift in this data due to natural climatological features of the regions presented in DASIO. We propose the framework for systematic study of automatic schemes for the retrieval of total cloud cover from all-sky imagery along with the quality measures for the assessment and comparison of the algorithms within the framework.

**Author Contributions:** Conceptualization, project administration, M. Krinitskiy and S. Gulev; methodology, software, validation, formal analysis, data curation, visualization, M. Krinitskiy; investigation, M. Krinitskiy, M. Aleksandrova, P. Verezhenskaya, A. Sinitsyn, A. Gavrikov, N. Kovaleva; funding acquisition, resources and supervision, S. Gulev; writing—original draft preparation, M. Krinitskiy and M. Aleksandrova; writing—review and editing, M. Krinitskiy and N. Kovaleva.”

**Funding:** This work was undertaken with financial support by the Russian Ministry of Science and Higher Education (agreement 05.616.21.0112, project ID RFMEFI61619X0112).

**Acknowledgments:** We are deeply indebted to Svyatoslav Elizarov for his contribution as a consultant on deep learning techniques.

**Conflicts of Interest:** The authors declare no conflict of interest.

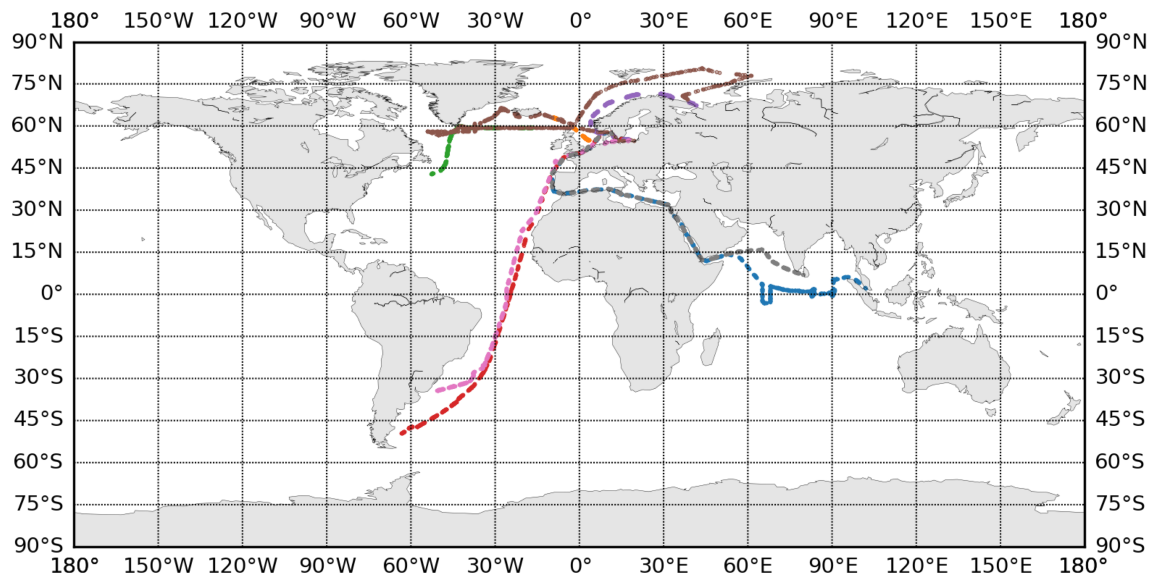
## Abbreviations

The following abbreviations are used in this manuscript:

ENSO	El Niño Southern Oscillation
TCC	Total Cloud Cover
NAO	North Atlantic Oscillation
PDO	Pacific Decadal Oscillation
ITCZ	Intertropical Convergence Zone
VOS	Voluntary Observing Ships
ICOADS	International Comprehensive Ocean-Atmosphere Data Set
WMO	World Meteorological Organization
DASIO	Dataset of All-Sky Imagery Over the Ocean
CVAE	Convolutional Variational Autoencoder
MSE	mean squared error
KL	Kullback–Leibler (divergence)
SGDR	stochastic gradient descent with warm restarts [79]

## A Supplementary materials

In this section, we present the materials we consider important for the reproducibility of our study, though not essential for the main sections of the paper. In fig. A1, we present the map of all of the scientific cruises taken so far, that contributed into the DASIO collection. In fig. A2, we present



**Figure A1.** Routes of scientific missions resulting in DASIO collection.

examples of DASIO collection that were marked as outliers using CVAE model exploited for data filtering presented in Section 3.1.

In tab. A1, we present the details of the architecture of CVAE model we come up with for filtering DASIO outliers. The model is built using ResNet [89] building blocks (namely, `residual_block` and `identity_block`) for improving the training stability and increasing the capability for training itself. More details of the implementation of CVAE model are available in source code of our study at GitHub: <https://github.com/MKrinitskiy/TCCfromAllSkyImagery>.

In fig. A3, we also present the example of the learning rate curve for one of the training runs of CVAE model.

In fig. A4, we present the details of PNetOR model training in different data split scenarios that are characterized by different strength of covariate shift. We do not show the learning curves of all the runs performed in order to assess the uncertainty of the quality estimates. Instead, we demonstrate only one typical run per scenario of train-validation split.



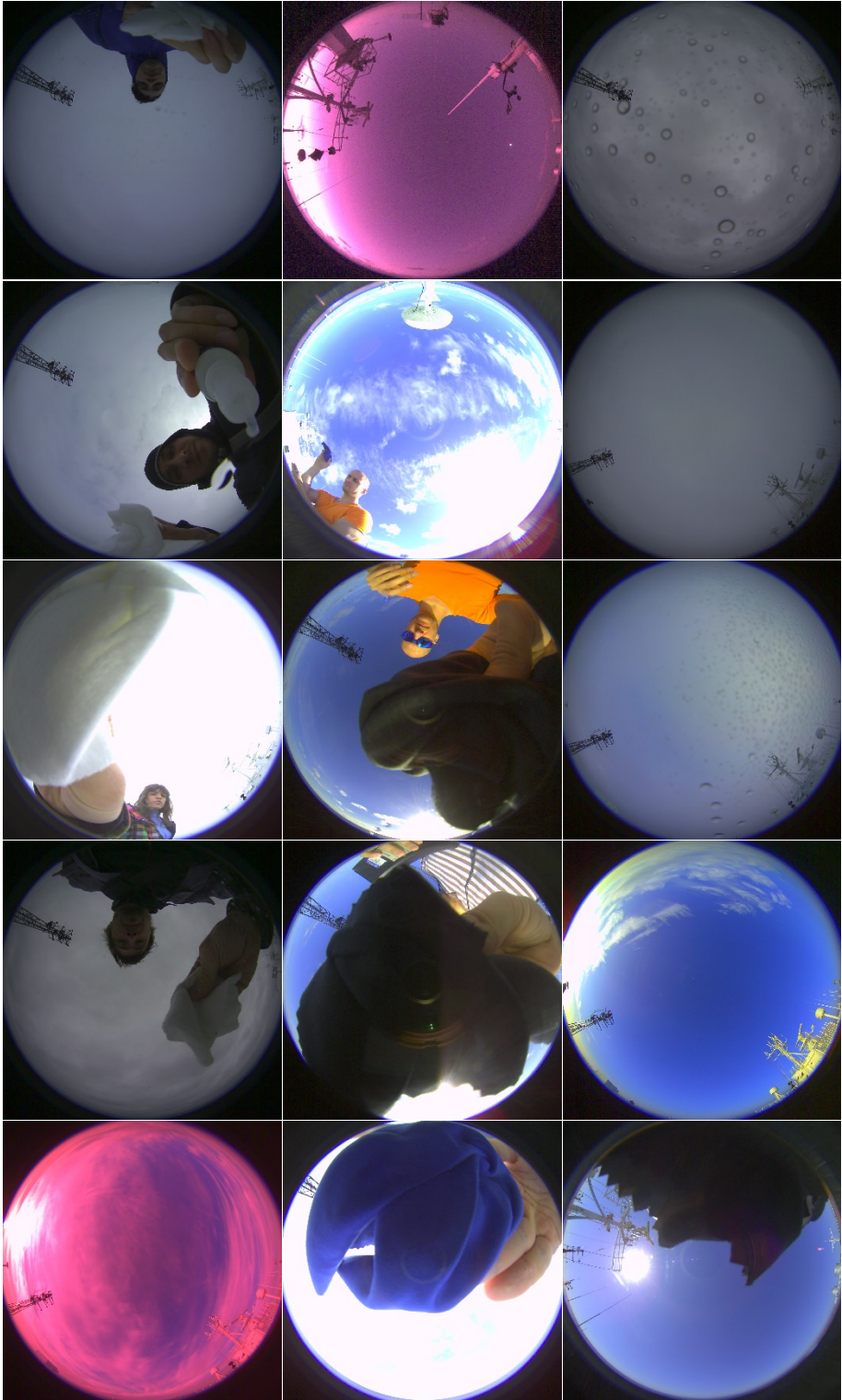
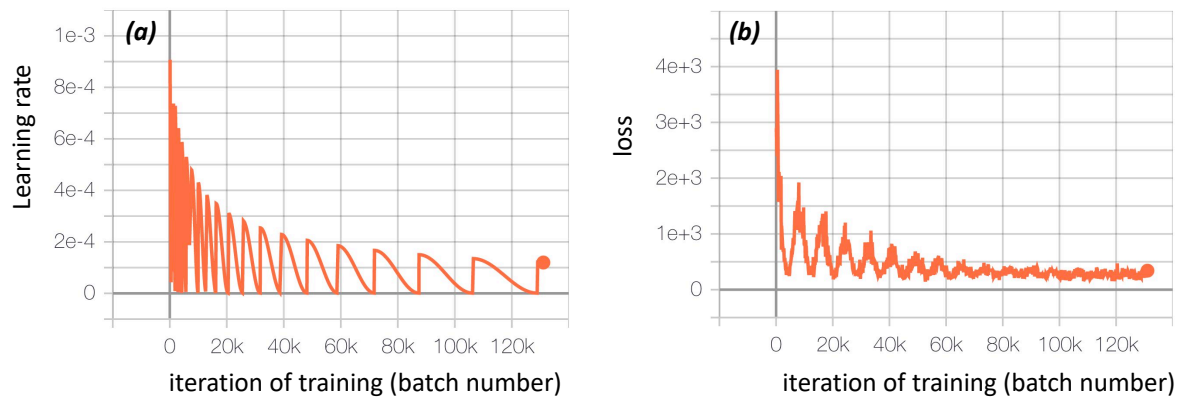


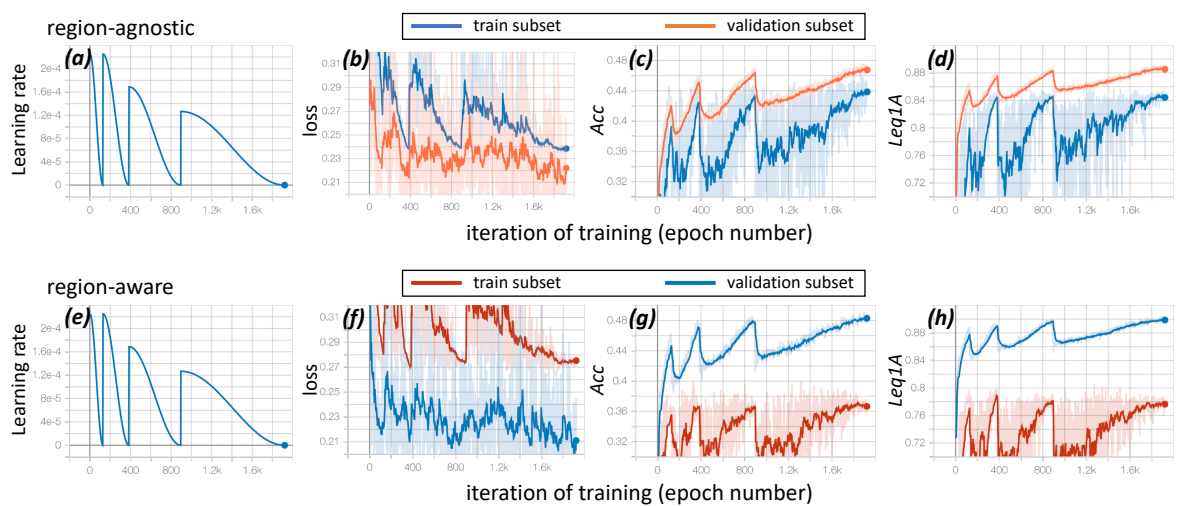
Figure A2. Examples of DASIO collection marked as outliers using the CVAE model exploited in this study.

Table A1. CVAE architecture details

block name	block type	inputs	input size	output size
encoder_input	input	-	256,256,3	256,256,3
mask_input	input	-	256,256,3	256,256,3
residual_block_1c	residual_block	encoder_input	256,256,3	256,256,32
identity_block_1i	identity_block	residual_block_1c	256,256,32	256,256,32
residual_block_2c	residual_block	identity_block_1i	256,256,32	128,128,64
identity_block_2i	identity_block	residual_block_2c	128,128,64	128,128,64
residual_block_3c	residual_block	identity_block_2i	128,128,64	64,64,128
identity_block_3i	identity_block	residual_block_3c	64,64,128	64,64,128
residual_block_4c	residual_block	identity_block_3i	64,64,128	64,64,128
identity_block_4i	identity_block	residual_block_4c	64,64,128	32,32,256
residual_block_5c	residual_block	identity_block_4i	32,32,256	32,32,256
identity_block_5i	identity_block	residual_block_5c	32,32,256	16,16,512
identity_block_6i	residual_block	residual_block_5c	32,32,256	16,16,512
residual_block_6c	identity_block	identity_block_6i	16,16,512	8,8,1024
identity_block_7i	residual_block	residual_block_6c	8,8,1024	8,8,1024
residual_block_7c	identity_block	identity_block_7i	8,8,1024	4,4,1024
enc_gap2d	GlobalAveragePooling2D	identity_block_7i	8,8,1024	1024
bottleneck	fully-connected	enc_gap2d	1024	512
z_mean_fc	fully-connected	bottleneck	512	512
z_var_fc	fully-connected	bottleneck	512	512
z_sampling	normal sampling	z_mean_fc,z_var_fc	(512), (512)	512
dec_input	fully-connected	z_sampling	512	1024
dec_reshape	Reshape	dec_input	1024	32,32,1
dec_identity_block_1i	identity_block	dec_reshape	32,32,1	32,32,1024
dec_identity_block_2i	identity_block	dec_identity_block_1i	32,32,1024	32,32,1024
dec_upsampling_1u	UpSampling2D	dec_identity_block_2i	32,32,1024	64,64,1024
dec_identity_block_3i	identity_block	dec_upsampling_1u	64,64,1024	64,64,512
dec_identity_block_4i	identity_block	dec_identity_block_3i	64,64,512	64,64,512
dec_upsampling_2u	UpSampling2D	dec_identity_block_4i	64,64,512	128,128,512
dec_identity_block_5i	identity_block	dec_upsampling_2u	128,128,512	128,128,256
dec_identity_block_6i	identity_block	dec_identity_block_5i	128,128,256	128,128,256
dec_upsampling_3u	UpSampling2D	dec_identity_block_6i	128,128,256	256,256,256
dec_identity_block_7i	identity_block	dec_upsampling_3u	256,256,256	256,256,256
dec_identity_block_8i	identity_block	dec_identity_block_7i	256,256,256	256,256,256
dec_identity_block_9i	identity_block	dec_identity_block_8i	256,256,256	256,256,128
dec_identity_block_10i	identity_block	dec_identity_block_9i	256,256,128	256,256,128
dec_conv2d_out	Conv2D	dec_identity_block_10i	256,256,128	256,256,3
dec_pw_norm	min-max normalization	dec_conv2d_out	256,256,3	256,256,3
masking	element-wise multiplication	mask_input, dec_pw_norm	(256,256,3), (256,256,3)	256,256,3



**Figure A3.** Training details of the CVAE model: (a) learning rate scheduling strategy implies SGDR [79] with exponential decay of the maximum value of learning rate; (b) typical learning curve (estimated on hold-out validation subset of DASIO).



**Figure A4.** Training details of data-driven models of our study: (a-d) learning curves for the region-agnostic scenario; (e-h) learning curves for the region-aware scenario.

## References

1. Dobson, F.W.; Smith, S.D. Bulk models of solar radiation at sea. *114*, 165–182. [\\_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49711447909](https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49711447909), doi:10.1002/qj.49711447909.
2. Aleksandrova, M.P.; Gulev, S.K.; Sinitsyn, A.V. An improvement of parametrization of short-wave radiation at the sea surface on the basis of direct measurements in the Atlantic. *32*, 245–251. doi:10.3103/S1068373907040048.
3. Josey, S.A.; Pascal, R.W.; Taylor, P.K.; Yelland, M.J. A new formula for determining the atmospheric longwave flux at the ocean surface at mid-high latitudes. *108*. doi:10.1029/2002JC001418.
4. Dufour, A.; Zolina, O.; Gulev, S.K. Atmospheric Moisture Transport to the Arctic: Assessment of Reanalyses and Analysis of Transport Components. *29*, 5061–5081. Publisher: American Meteorological Society, doi:10.1175/JCLI-D-15-0559.1.
5. Hand, R.; Keenlyside, N.; Omrani, N.E.; Latif, M. Simulated response to inter-annual SST variations in the Gulf Stream region. *42*, 715–731. doi:10.1007/s00382-013-1715-y.
6. Hernández, K.; Yannicelli, B.; Montecinos, A.; Ramos, M.; González, H.E.; Daneri, G. Temporal variability of incidental solar radiation and modulating factors in a coastal upwelling area (36S). *92-95*, 18–32. doi:10.1016/j.pocean.2011.07.011.
7. Rossow, W.B.; Schiffer, R.A. ISCCP Cloud Data Products. *72*, 2–20. Publisher: American Meteorological Society, doi:10.1175/1520-0477(1991)072<0002:ICDP>2.0.CO;2.
8. Rossow, W.B.; Schiffer, R.A. Advances in Understanding Clouds from ISCCP. *80*, 2261–2288. Publisher: American Meteorological Society, doi:10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2.
9. Frey, R.A.; Ackerman, S.A.; Liu, Y.; Strabala, K.I.; Zhang, H.; Key, J.R.; Wang, X. Cloud Detection with MODIS. Part I: Improvements in the MODIS Cloud Mask for Collection 5. *25*, 1057–1072. Publisher: American Meteorological Society, doi:10.1175/2008JTECHA1052.1.
10. Foster, M.J.; Heidinger, A. PATMOS-x: Results from a Diurnally Corrected 30-yr Satellite Cloud Climatology. *26*, 414–425. Publisher: American Meteorological Society, doi:10.1175/JCLI-D-11-00666.1.
11. Karlsson, K.G.; Riihelä, A.; Müller, R.; Meirink, J.F.; Sedlar, J.; Stengel, M.; Lockhoff, M.; Trentmann, J.; Kaspar, F.; Hollmann, R.; Wolters, E. CLARA-A1: a cloud, albedo, and radiation dataset from 28 yr of global AVHRR data. *13*, 5351–5367. Publisher: Copernicus GmbH, doi:https://doi.org/10.5194/acp-13-5351-2013.
12. Stubenrauch, C.J.; Rossow, W.B.; Kinne, S.; Ackerman, S.; Cesana, G.; Chepfer, H.; Di Girolamo, L.; Getzewich, B.; Guignard, A.; Heidinger, A.; Maddux, B.C.; Menzel, W.P.; Minnis, P.; Pearl, C.; Platnick, S.; Poulsen, C.; Riedi, J.; Sun-Mack, S.; Walther, A.; Winker, D.; Zeng, S.; Zhao, G. Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel. *94*, 1031–1049. Publisher: American Meteorological Society, doi:10.1175/BAMS-D-12-00117.1.
13. Chernokulsky, A.; Mokhov, I.I. Climatology of Total Cloudiness in the Arctic: An Intercomparison of Observations and Reanalyses. ISSN: 1687-9309 Pages: e542093 Publisher: Hindawi Volume: 2012, doi:https://doi.org/10.1155/2012/542093.
14. Bedacht, E.; Gulev, S.K.; Macke, A. Intercomparison of global cloud cover fields over oceans from the VOS observations and NCEP/NCAR reanalysis. *27*, 1707–1719. [\\_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1490](https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1490), doi:10.1002/joc.1490.
15. Liu, Y.; Key, J.R. Assessment of Arctic Cloud Cover Anomalies in Atmospheric Reanalysis Products Using Satellite Data. *29*, 6065–6083. Publisher: American Meteorological Society, doi:10.1175/JCLI-D-15-0861.1.
16. Free, M.; Sun, B.; Yoo, H.L. Comparison between Total Cloud Cover in Four Reanalysis Products and Cloud Measured by Visual Observations at U.S. Weather Stations. *29*, 2015–2021. Publisher: American Meteorological Society, doi:10.1175/JCLI-D-15-0637.1.
17. Woodruff, S.D.; Worley, S.J.; Lubker, S.J.; Ji, Z.; Freeman, J.E.; Berry, D.I.; Brohan, P.; Kent, E.C.; Reynolds, R.W.; Smith, S.R.; Wilkinson, C. ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *31*, 951–967. [\\_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.2103](https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.2103), doi:10.1002/joc.2103.
18. Freeman, E.; Woodruff, S.D.; Worley, S.J.; Lubker, S.J.; Kent, E.C.; Angel, W.E.; Berry, D.I.; Brohan, P.; Eastman, R.; Gates, L.; Gloeden, W.; Ji, Z.; Lawrimore, J.; Rayner, N.A.; Rosenhagen, G.; Smith, S.R. ICOADS Release 3.0: a major update to the historical marine climate record. *37*, 2211–2232. [\\_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4775](https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4775), doi:10.1002/joc.4775.



19. Berry, D.I.; Kent, E.C. A New Air–Sea Interaction Gridded Dataset from ICOADS With Uncertainty Estimates. 90, 645–656. Publisher: American Meteorological Society, doi:10.1175/2008BAMS2639.1.
20. Norris, J.R. On Trends and Possible Artifacts in Global Ocean Cloud Cover between 1952 and 1995. 12, 1864–1870. Publisher: American Meteorological Society, doi:10.1175/1520-0442(1999)012<1864:OTAPAI>2.0.CO;2.
21. Eastman, R.; Warren, S.G.; Hahn, C.J. Variations in Cloud Cover and Cloud Types over the Ocean from Surface Observations, 1954–2008. 24, 5914–5934. Publisher: American Meteorological Society, doi:10.1175/2011JCLI3972.1.
22. Elms, J.D.; Woodruff, S.D.; Worley, S.J.; Hanson, C. Digitizing historical records for the comprehensive ocean-atmosphere data set (COADS). 4, 4–10.
23. Aleksandrova, M.; Gulev, S.K.; Belyaev, K. Probability Distribution for the Visually Observed Fractional Cloud Cover over the Ocean. 31, 3207–3232. Publisher: American Meteorological Society, doi:10.1175/JCLI-D-17-0317.1.
24. *Manual on Codes - International Codes, Volume I.1: part A- Alphanumeric Codes. (2015: 2011 edition updated);* World Meteorological Organization.
25. *Guide to meteorological instruments and methods of observation, Chapter 15 “Observations on clouds”, 15.2 “Estimation and observation of cloud amount, height and type”;* World Meteorological Organization.
26. Heinle, A.; Macke, A.; Srivastav, A. Automatic cloud classification of whole sky images. 3, 557–567. Publisher: Copernicus GmbH, doi:https://doi.org/10.5194/amt-3-557-2010.
27. Calbó, J.; Sabburg, J. Feature Extraction from Whole-Sky Ground-Based Images for Cloud-Type Recognition. 25, 3–14. Publisher: American Meteorological Society, doi:10.1175/2007JTECHA959.1.
28. Liu, S.; Duan, L.; Zhang, Z.; Cao, X. Hierarchical Multimodal Fusion for Ground-Based Cloud Classification in Weather Station Networks. 7, 85688–85695. Conference Name: IEEE Access, doi:10.1109/ACCESS.2019.2926092.
29. Xiao, Y.; Cao, Z.; Zhuo, W.; Ye, L.; Zhu, L. mCLOUD: A Multiview Visual Feature Extraction Mechanism for Ground-Based Cloud Image Categorization. 33, 789–801. Publisher: American Meteorological Society, doi:10.1175/JTECH-D-15-0015.1.
30. Liu, S.; Li, M.; Zhang, Z.; Xiao, B.; Cao, X. Multimodal Ground-Based Cloud Classification Using Joint Fusion Convolutional Neural Network. 10, 822. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, doi:10.3390/rs10060822.
31. Taravat, A.; Frate, F.D.; Cornaro, C.; Vergari, S. Neural Networks and Support Vector Machine Algorithms for Automatic Cloud Classification of Whole-Sky Ground-Based Images. 12, 666–670. doi:10.1109/LGRS.2014.2356616.
32. Long, C.; DeLuisi, J. Development of an automated hemispheric sky imager for cloud fraction retrievals. pp. 171–174.
33. Long, C.N.; Sabburg, J.M.; Calbó, J.; Pagès, D. Retrieving Cloud Characteristics from Ground-Based Daytime Color All-Sky Images. 23, 633–652. Publisher: American Meteorological Society, doi:10.1175/JTECH1875.1.
34. Yang, J.; Min, Q.; Lu, W.; Yao, W.; Ma, Y.; Du, J.; Lu, T.; Liu, G. An automated cloud detection method based on the green channel of total-sky visible images. 8, 4671–4679. doi:10.5194/amt-8-4671-2015.
35. Kazantzidis, A.; Tzoumanikas, P.; Bais, A.F.; Fotopoulos, S.; Economou, G. Cloud detection and classification with the use of whole-sky ground-based images. 113, 80–88. doi:10.1016/j.atmosres.2012.05.005.
36. Yamashita, M.; Yoshimura, M.; Nakashizuka, T. Cloud cover estimation using multitemporal hemisphere imageries. 35, 826–829. Publisher: Citeseer.
37. Yamashita, M.; Yoshimura, M. Ground-based cloud observation for satellite-based cloud discrimination and its validation. 39, B8.
38. Kalisch, J.; Macke, A. Estimation of the total cloud cover with high temporal resolution and parametrization of short-term fluctuations of sea surface insolation. pp. 603–611. Publisher: Schweizerbart’sche Verlagsbuchhandlung, doi:10.1127/0941-2948/2008/0321.
39. Krinitskiy, M.; Sinitsyn, A. Adaptive algorithm for cloud cover estimation from all-sky images over the sea. 56, 315–319. doi:10.1134/S0001437016020132.



40. Luo, L.; Hamilton, D.; Han, B. Estimation of total cloud cover from solar radiation observations at Lake Rotorua, New Zealand. *84*, 501–506. doi:10.1016/j.solener.2010.01.012.
41. Long, C.N. Correcting for Circumsolar and Near-Horizon Errors in Sky Cover Retrievals from Sky Images. *4*, 45–52.
42. Peng, Z.; Yu, D.; Huang, D.; Heiser, J.; Yoo, S.; Kalb, P. 3D cloud detection and tracking system for solar forecast using multiple sky imagers. *118*, 496–519. doi:10.1016/j.solener.2015.05.037.
43. Kim, B.Y.; Cha, J.W. Cloud Observation and Cloud Cover Calculation at Nighttime Using the Automatic Cloud Observation System (ACOS) Package. *12*, 2314. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute, doi:10.3390/rs12142314.
44. Lothon, M.; Barnéoud, P.; Gabella, O.; Lohou, F.; Derrien, S.; Rondi, S.; Chiriaco, M.; Bastin, S.; Dupont, J.C.; Haeffelin, M.; Badosa, J.; Pascal, N.; Montoux, N. ELIFAN, an algorithm for the estimation of cloud cover from sky imagers. *12*, 5519–5534. doi:10.5194/amt-12-5519-2019.
45. Yamashita, M.; Yoshimura, M. ANALYSIS ON LIGHT QUANTITY AND QUALITY BASED ON DIVERSE CLOUD CONDITIONS. ISBN: 1682-1750.
46. Chauvin, R.; Nou, J.; Thil, S.; Traoré, A.; Grieu, S. Cloud Detection Methodology Based on a Sky-imaging System. *69*, 1970–1980. doi:10.1016/j.egypro.2015.03.198.
47. Yang, J.; Min, Q.; Lu, W.; Ma, Y.; Yao, W.; Lu, T.; Du, J.; Liu, G. A total sky cloud detection method using real clear sky background. *9*, 587–597. doi:https://doi.org/10.5194/amt-9-587-2016.
48. Souza-Echer, M.P.; Pereira, E.B.; Bins, L.S.; Andrade, M.a.R. A Simple Method for the Assessment of the Cloud Cover State in High-Latitude Regions by a Ground-Based Digital Camera. *23*, 437–447. Publisher: American Meteorological Society, doi:10.1175/JTECH1833.1.
49. Fa, T.; Xie, W.; Wang, Y.; Xia, Y. Development of an all-sky imaging system for cloud cover assessment. *58*, 5516–5524. Publisher: Optical Society of America, doi:10.1364/AO.58.005516.
50. Liu, S.; Zhang, L.; Zhang, Z.; Wang, C.; Xiao, B. Automatic Cloud Detection for All-Sky Images Using Superpixel Segmentation. *12*, 354–358. doi:10.1109/LGRS.2014.2341291.
51. Krinitskiy, M. Application of machine learning methods to the solar disk state detection by all-sky images over the ocean. *57*, 265–269. doi:10.1134/S0001437017020126.
52. Dev, S.; Lee, Y.H.; Winkler, S. Color-Based Segmentation of Sky/Cloud Images From Ground-Based Cameras. *10*, 231–242. doi:10.1109/JSTARS.2016.2558474.
53. Norris, J.R. Low Cloud Type over the Ocean from Surface Observations. Part II: Geographical and Seasonal Variations. *11*, 383–403. Publisher: American Meteorological Society, doi:10.1175/1520-0442(1998)011<0383:LCTOTO>2.0.CO;2.
54. Klein, S.A.; Hartmann, D.L. The Seasonal Cycle of Low Stratiform Clouds. *6*, 1587–1606. Publisher: American Meteorological Society, doi:10.1175/1520-0442(1993)006<1587:TSCOLS>2.0.CO;2.
55. Houze Jr, R.A. *Cloud dynamics*; Academic press.
56. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, Cham, Lecture Notes in Computer Science, pp. 234–241. doi:10.1007/978-3-319-24574-4\_28.
57. Gutiérrez, P.A.; Pérez-Ortiz, M.; Sánchez-Monedero, J.; Fernández-Navarro, F.; Hervás-Martínez, C. Ordinal Regression Methods: Survey and Experimental Study. *28*, 127–146. Conference Name: IEEE Transactions on Knowledge and Data Engineering, doi:10.1109/TKDE.2015.2457911.
58. Kassianov, E.; Long, C.N.; Ovtchinnikov, M. Cloud Sky Cover versus Cloud Fraction: Whole-Sky Simulations and Observations. *44*, 86–98. doi:10.1175/JAM-2184.1.
59. Allmen, M.C.; Kegelmeyer, W.P. The Computation of Cloud-Base Height from Paired Whole-Sky Imaging Cameras. *13*, 97–113. Publisher: American Meteorological Society, doi:10.1175/1520-0426(1996)013<0097:TCOCBH>2.0.CO;2.
60. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. [1312.6114].
61. Simard, P.; Steinkraus, D.; Platt, J. Best practices for convolutional neural networks applied to visual document analysis. Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 958–963. doi:10.1109/ICDAR.2003.1227801.
62. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F.; Burges, C.J.C.; Bottou, L.; Weinberger, K.Q., Eds.; Curran Associates, Inc.; pp. 1097–1105.

63. Krizhevsky, A.; Hinton, G. Learning multiple layers of features from tiny images. Publisher: Citeseer.
64. Mash, R.; Borghetti, B.; Pecarina, J. Improved Aircraft Recognition for Aerial Refueling Through Data Augmentation in Convolutional Neural Networks. *Advances in Visual Computing*; Bebis, G.; Boyle, R.; Parvin, B.; Koracin, D.; Porikli, F.; Skaff, S.; Entezari, A.; Min, J.; Iwai, D.; Sadagic, A.; Scheidegger, C.; Isenberg, T., Eds. Springer International Publishing, Lecture Notes in Computer Science, pp. 113–122. doi:10.1007/978-3-319-50835-1\_11.
65. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop (IIPHDW), pp. 117–122. doi:10.1109/IIPHDW.2018.8388338.
66. Paschali, M.; Simson, W.; Roy, A.G.; Naeem, M.F.; Göbl, R.; Wachinger, C.; Navab, N. Data Augmentation with Manifold Exploring Geometric Transformations for Increased Performance and Robustness. [1901.04420].
67. Shorten, C.; Khoshgouftaar, T.M. A survey on Image Data Augmentation for Deep Learning. 6, 60. doi:10.1186/s40537-019-0197-0.
68. Wang, X.; Wang, K.; Lian, S. A Survey on Face Data Augmentation. 32, 15503–15531, [1904.11685]. doi:10.1007/s00521-020-04748-3.
69. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. [1708.04896]. version: 2.
70. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. AutoAugment: Learning Augmentation Policies from Data. [1805.09501]. version: 3.
71. Lim, S.; Kim, I.; Kim, T.; Kim, C.; Kim, S. Fast AutoAugment. [1905.00397]. version: 2.
72. Zoph, B.; Cubuk, E.D.; Ghiasi, G.; Lin, T.Y.; Shlens, J.; Le, Q.V. Learning Data Augmentation Strategies for Object Detection. [1906.11172]. version: 1.
73. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H.; Larochelle, H.; Beygelzimer, A.; dAlché-Buc, F.; Fox, E.; Garnett, R., Eds.; Curran Associates, Inc., 2019; pp. 8024–8035.
74. Sola, J.; Sevilla, J. Importance of input data normalization for the application of neural networks to complex industrial problems. 44, 1464–1468. Conference Name: IEEE Transactions on Nuclear Science, doi:10.1109/23.589532.
75. Deng, L. A tutorial survey of architectures, algorithms, and applications for deep learning. 3. doi:10.1017/atsip.2013.9.
76. Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; Alsaadi, F.E. A survey of deep neural network architectures and their applications. 234, 11–26. doi:10.1016/j.neucom.2016.12.038.
77. Krinitskiy, M.; Zyulyaeva, Y.; Gulev, S. Clustering of polar vortex states using convolutional autoencoders. Vol. 2426, pp. 52–61.
78. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. [1412.6980].
79. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. [1608.03983].
80. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. [1409.1556].
81. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. [1511.00561].
82. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. [1606.00915].
83. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. [1706.05587].
84. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. [1802.02611]. version: 3.
85. LeCun, Y.; Cortes, C.; Burges, C. MNIST handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist> 2010, 2.
86. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, pp. 248–255.

87. Deep Pyramidal Residual Networks.
88. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807. doi:10.1109/CVPR.2017.195.
89. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. pp. 770–778.
90. Costa, M. Probabilistic interpretation of feedforward network outputs, with relationships to statistical prediction of ordinal quantities. 7, 627–637. doi:10.1142/s0129065796000610.
91. da Costa, J.P.; Cardoso, J.S. Classification of Ordinal Data Using Neural Networks. Machine Learning: ECML 2005; Gama, J.; Camacho, R.; Brazdil, P.B.; Jorge, A.M.; Torgo, L., Eds. Springer, Lecture Notes in Computer Science, pp. 690–697. doi:10.1007/11564096\_70.
92. Fernández-Navarro, F.; Riccardi, A.; Carloni, S. Ordinal neural networks without iterative tuning. 25, 2075–2085. doi:10.1109/TNNLS.2014.2304976.
93. Hamsici, O.C.; Martinez, A.M. Multiple Ordinal Regression by Maximizing the Sum of Margins. 27, 2072–2083. doi:10.1109/TNNLS.2015.2477321.
94. Cheng, J.; Wang, Z.; Pollastri, G. A neural network approach to ordinal regression. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1279–1284. ISSN: 2161-4407, doi:10.1109/IJCNN.2008.4633963.
95. Pinto da Costa, J.F.; Alonso, H.; Cardoso, J.S. The unimodal model for the classification of ordinal data. 21, 78–91. doi:10.1016/j.neunet.2007.10.003.
96. Pinto da Costa, J.; Alonso, H.; Cardoso, J.S. Corrigendum to “The unimodal model for the classification of ordinal data” [Neural Netw. 21 (2008) 78–79]. 59, 73–75. doi:10.1016/j.neunet.2014.06.003.