ISPRS Journal of Photogrammetry and Remote Sensing 158 (2019) 231–240

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Statistical retrieval of atmospheric profiles with deep convolutional neural networks

David Malmgren-Hansen[a,*], Valero Laparra[b], Allan Aasbjerg Nielsen[a], Gustau Camps-Valls[b]

[a] Technical University of Denmark, Lyngby, Denmark
[b] Image Processing Laboratory (IPL), Universitat de València, València, Spain

ARTICLE INFO

Keywords:
Atmospheric measurements
Neural networks
Infrared measurements
Information retrieval

ABSTRACT

Infrared atmospheric sounders, such as IASI, provide an unprecedented source of information for atmosphere monitoring and weather forecasting. Sensors provide rich spectral information that allows retrieval of temperature and moisture profiles. From a statistical point of view, the challenge is immense: on the one hand, "underdetermination" is common place as regression needs to work on high dimensional input and output spaces; on the other hand, redundancy is present in all dimensions (spatial, spectral and temporal). On top of this, several noise sources are encountered in the data.

In this paper, we present for the first time the use of convolutional neural networks for the retrieval of atmospheric profiles from IASI sounding data. The first step of the proposed pipeline performs spectral dimensionality reduction taking into account the signal to noise characteristics. The second step encodes spatial and spectral information, and finally prediction of multidimensional profiles is done with deep convolutional networks. We give empirical evidence of the performance in a wide range of situations. Networks were trained on orbits of IASI radiances and tested out of sample with great accuracy over competing approximations, such as linear regression (+32%). We also observed an improvement in accuracy when predicting over clouds, thus increasing the yield by 34% over linear regression. The proposed scheme allows us to predict related variables from an already trained model, performing transfer learning in a very easy manner. We conclude that deep learning is an appropriate learning paradigm for statistical retrieval of atmospheric profiles.

1. Introduction

Temperature and water vapour atmospheric profiles are essential meteorological parameters for weather forecasting and atmospheric chemistry studies. Observations from high spectral resolution infrared sounding instruments on board satellites provide unprecedented accuracy and vertical resolution of temperature and water vapour profiles. However, it is not trivial to retrieve the full information content from radiation measurements. Accordingly, improved retrieval algorithms are desirable to achieve optimal performance for existing and future infrared sounding instrumentation.

The use of MetOp data observations has an important impact on several Numerical Weather prediction (NWP) forecasts. The Infrared Atmospheric Sounding Interferometer (IASI) sensor is implemented on the MetOp satellite series. In particular, IASI collects rich spectral information to derive temperature and moisture profiles, which are essential to the understanding of weather and to derive atmospheric forecasts. The sensor provides infrared spectra, from which temperature and humidity profiles with high vertical resolution and accuracy are derived. Additionally, it is used for the determination of trace gases such as ozone, nitrous oxide, carbon dioxide and methane, as well as land and sea surface temperature, emissivity, and cloud properties (EUMETSAT, 2014; Tournier et al., 2002).

EUMETSAT, NOAA, NASA and other operational agencies are continuously developing product processing facilities to obtain L2 atmospheric profile products from infrared hyperspectral radiance instruments, such as IASI, AIRS or the upcoming MTG-IRS. A standard approach relies on physical models in general and the optimal estimation method (OEM) approach, (August et al., 2012). The use of linear regression (LR) is widely adopted to provide a first guess estimate of the variable of interest to start the OEM run. Because of the strong input spectral co-linearity, the LR model is actually run on top of the data projected onto the first principal components or Empirical Orthogonal Functions (EOF) of the measured brightness temperature spectrum (or radiances). To further improve the results of this first guess estimate, nonlinear statistical retrieval methods can be applied. Inclusion of

* Corresponding author.
E-mail address: dmal@dtu.dk (D. Malmgren-Hansen).

https://doi.org/10.1016/j.isprsjprs.2019.10.002
Received 8 May 2019; Received in revised form 2 October 2019; Accepted 5 October 2019

nonlinear and nonparametric machine learning models here produce more accurate first guesses, hence faster convergence of the final (eventually OEM) retrieval approach. These methods have proven to be valid in retrieval of temperature, dew point temperature (humidity), and ozone atmospheric profiles (Camps-Valls et al., 2012).

From a statistical standpoint, the challenge is immense. On the one hand, the curse of dimensionality is often present when using infrared sounding data for atmospheric profile estimation, because of the high dimensional input and output spaces, and exponential increase computational resources needed. On the other hand, redundancy is present in all dimensions (spatial, spectral and temporal). Additionally, several noise sources and high noise levels are encountered in the data, which in many cases are correlated with the signal. The previous L2 processing scheme presented in Camps-Valls et al. (2012) consisted of first performing a *spectral* dimensionality reduction based on Principal Component Analysis (PCA) (Hotelling, 1933), and then a nonlinear regression based on kernel methods (Camps-Valls et al., 2011, 2012; Camps-Valls and Bruzzone, 2009). Despite being an effective approach, the scheme reveals some deficiencies. The PCA transformation accounts for most of the signal variance, but does not consider the correlation between the signal and the noise. On the other hand, the spatial information is discarded and the retrieval algorithm acts on a pixel (FOV) basis. Only very recent methods have included spatial-spectral feature relations in the retrieval algorithm, yet in an indirect way through either post-filtering of the product, or via data compression (García-Sobrino et al., 2017). In this paper, we propose a general scheme to cope with all these problems.

Three main motivations guide our proposal:

- *Accounting for noisy and spatial features.* Recently, in García-Sobrino et al. (2017, 2019), great improvement in the performance of retrieval methods was reported when applying standard compression algorithms to the images. Although this result may appear counterintuitive since compression implies reduction on the amount of information in the images, a certain level of compression is actually beneficial because: (1) compression removes information but also noise, and it could be useful to remove the components with low signal-to-noise ratio (SNR); and (2) spatial compression introduces information about the neighbouring pixels in an indirect yet simple way. Including the noise estimate in the design of PCA of infrared sounders has been considered before (Collard et al., 2010), and actually it is currently implemented in the IASI pipeline (EUMETSAT, 2017). However the inclusion of the spatial information, while it is important (Laparra and Santos-Rodríguez, 2016), has obtained less attention. The use of Minimum Noise Fractions (MNF) employed in this paper is a simple and mathematically elegant way to take advantage of both properties simultaneously. The way we apply MNF here enforces the inclusion of spatial information as noise is estimated by the residuals from fitting a quadratic surface locally. In this work, we compare the effect of using noise-free PCA and MNF when retrieving temperature profiles using IASI data. Moreover, since PCA and MNF are both linear and unsupervised transformations, using MNF does not introduce any critical modification in the data processing pipeline. One can simply replace the PCA principal components with MNF components. Also, replacing PCA with MNF could be advantageous in other retrieval schemes.
- *Accounting for smoothness in the spatial and vertical dimensions.* All previous machine learning based algorithms (Blackwell et al., 2008; Camps-Valls et al., 2012; Camps-Valls, 2016; Laparra et al., 2017; Rivera-Caicedo et al., 2017) used for statistical retrieval exploited the spectral information in the FOVs only, and discarded spatial information of the acquired scene. Including spatial information in classifiers and regression methods has been done traditionally via hand-crafted features (Plaza et al., 2002; Tuia et al., 2010; Camps-Valls et al., 2006, 2014). This, however, requires expert knowledge, it is time consuming and scenario dependent. In the last decade, convolutional neural networks (CNNs) have excelled in many classification problems in remote sensing (Aptoula et al., 2016; Geng et al., 2015; Zhang et al., 2016; Luus et al., 2015; Maggiori et al., 2017; Zhang et al., 2016; Romero et al., 2016). CNNs allow to easily *learn* the proper filters to process images and optimize a task (in our case, prediction of atmospheric profiles). Traditional artificial neural network architectures have been explored for this problem e.g. in Aires et al. (2002), Whitburn et al. (2016), but these architectures consider single point measurements in the retrieval and fail to incorporate spatial information and feature transformations. It is, however, quite striking that very few applications of CNNs are found in the field of regression, and none to our knowledge for bio-geophysical parameter retrieval. In this paper, we present for the first time the use of deep convolutional neural networks for the retrieval of atmospheric profiles from IASI sounding data. We should note that, neural networks offer an additional advantage to our multivariate regression problem: models are intrinsically multi-output and account for the cross-relations between the state vector at different altitudes. This allows us to attain smoothness, and hence consistency, across the atmospheric column in a very straightforward way.
- *Accounting for higher level feature representations.* The problem of translating radiances to state parameters is challenging because of its intrinsic high non-linearity and underdertermination. Deep learning offers a simple strategy to approach the problem of complex feature representations by stacking together several convolutional layers. In the last decade, deep networks have replaced shallow architectures in many recognition and detection tasks.

Capitalizing on these three motivations, in this paper we propose a chained scheme that exploits the MNF transformation and deep convolutional neural networks for atmospheric parameter retrieval. In summary, the proposed scheme performs multidimensional output nonlinear regression, accounts for noise features, and exploits correlations in all dimensions.

The remainder of the paper is organized as follows. Section 2 presents the processing scheme and analyses the building blocks (dimensionality reduction and retrieval) in detail. Section 2.1 describes the datasets used for the development of the algorithm. Section 3 illustrates the performance of the proposed method in terms of accuracy, bias and smoothness of the estimates (across the space and vertical dimensions), both over land and over ocean. We also pay attention to the algorithm performance when predicting over clouds as a function of the cloud fraction. The section ends with an exploratory analysis of the performance of the method to estimate other (yet related) variables with minimal retraining. We outline the conclusions of the work and the foreseeable future developments in Section 4.

2. Methodology

Rather than modeling atmospheric parameters from single point measurements, the purpose here is to investigate spatial dependencies in the retrieval. IASI data are collected as 30 point measurements in a swath scan using a 2×2 pixel grid simultaneously. The IASI instrument scans around 765 swaths per orbit which then becomes 1530 lines of 60 points per orbit. This fact can be used to structure the data in rectangular grids and treat them as images likewise (García-Sobrino et al., 2017). We use this approach in two steps of our prediction pipeline illustrated in Fig. 1. The pipeline consists of (1) removing irrelevant spectral channels and structuring the data as images of dimension of $1530 \times 60 \times 4699$ per orbit (cf. Camps-Valls et al., 2012), (2) applying the linear basis (learned using an MNF decomposition) on the spectral components, (3) extracting patches from data so that observations are local neighbourhoods around each pixel, (4) running either a CNN model or a linear regression for retrieval of atmospheric parameters at 90 different altitudes simultaneously. Let us describe in detail each of these steps.

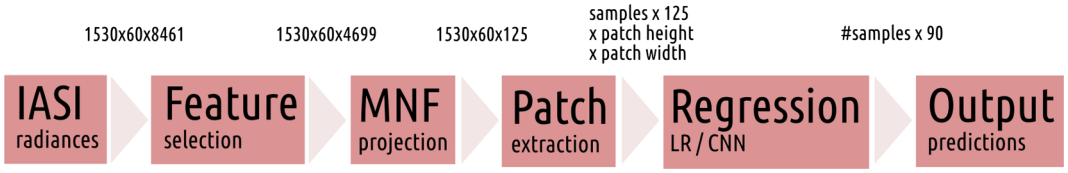

Fig. 1. Pipeline schematic: IASI spectra are first reduced from the original 8461 spectral channels by selecting a subset of 4699 channels according to noise specifications in Camps-Valls et al. (2012), which then pass through an MNF projection to reduce the dimensionality to 125 features. Subsequently, patch extraction is performed with varying sizes. Finally, either a linear regression or a CNN is used for prediction of the atmospheric profiles sampled at 90 ver.tical positions.

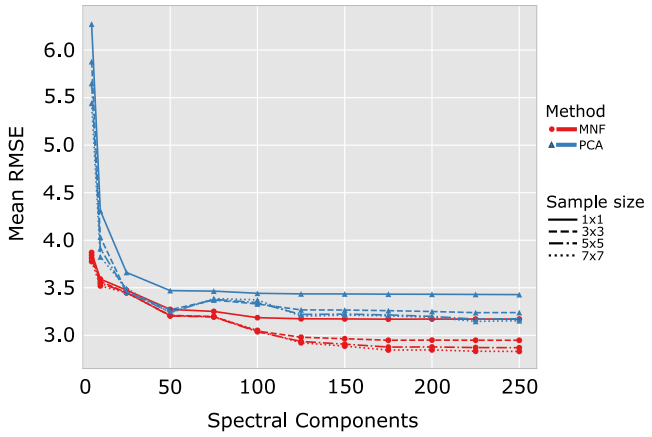

Fig. 2. Mean RMSE error for linear regression as a function of number of spectral components included, when predicting atmospheric temperatures. PCA and MNF signal decompositions of spectral channels are compared.

2.1. Data collection and preprocessing

The Infrared Atmospheric Sounding Interferometer (IASI) is an instrument implemented on the MetOp satellite series. From MetOp's polar orbit, the IASI instrument scans the Earth at an altitude of, approximately, 820 km. The instrument measures in the infrared part of the electromagnetic spectrum (between 645 cm^{-1} and 2760 cm^{-1}) at a horizontal resolution of 12 km over a swath width of, approximately, 2200 km. It obtains a global coverage of the Earth's surface every 12 h, representing 7 orbits in a sun-synchronous mid-morning orbit. This represents more than one million high dimensionality samples to be processed each day. Obtaining all the products provided by IASI with classical methods requires an enormous computational load. To process these products efficiently some works have focused on using machine learning methods (Camps-Valls et al., 2012; Laparra et al., 2015, 2017).

Each original sample has 8461 spectral channels, but following previous recommendations (Camps-Valls et al., 2012) we performed feature selection removing the most noisy channels and keeping 4699. Even with such drastic feature reduction, regression methods can suffer and easily overfit as many parameters need to be learned. In addition, even though some noise is removed by doing this channel selection, there still remains noise and spectral redundancy in the data. Actually it has been suggested that simple spatial smoothing techniques remove the noise and help improving the predictions quality (García-Sobrino et al., 2017). In the following subsection we pay attention to the feature extraction step to better pose the problem.

Products obtained from IASI data are used to for meteorological models. For instance humidity profiles reach an error of 10% at a vertical resolution of one kilometer. Temperature profiles can reach an accuracy of one Kelvin (EUMETSAT, 2017). To facilitate the training of our machine learning algorithms we matched each pixel with the temperature and dew point temperature profiles estimated using the European Center for Medium-Range Weather Forecasts (ECMWF) model. The ECMWF model provides estimations for 137 different pressure levels between $[10^{-2} \cdots 10^{3}]$ hPa in the atmosphere and spatial resolution of 0.5 degrees.

2.2. Dimensionality reduction

Traditionally, dimensionality reduction is done by means of PCA, or equivalently by means of Singular Value Decomposition (SVD) (Golub and Van Loan, 1996). In this context, PCA compresses the total variation of the original variables (i.e. radiance) into fewer uncorrelated variates termed 'principal components' which minimize the reconstruction error of the original variables. Alternatively, one could use a different feature extraction method, for instance Independent Component Analysis (ICA) (Hyvärinen et al., 2001) where the uncorrelated and statistically independent variates maximize a measure of non-Gaussianity such as negentropy in all original variables. Taking into account the noise estimate when designing PCA for infrared sounders has been shown to be important (Collard et al., 2010; EUMETSAT, 2017). In Malmgren-Hansen et al. (2017) we applied a minimum noise fraction (MNF) transformation that simultaneously minimizes the noise fraction or equivalently maximize the signal-to-noise ratio (given a noise model) in all original variables and takes into account the information contained in the spatial neighbors. The noise is estimated as the pixel-wise residual from a quadratic function fitted in a 3x3 window. It can be shown that the MNF variates can be considered as a form of independent components, (Larsen, 2002). Fig. 2 shows a result from Malmgren-Hansen et al. (2017) that compares MNF and PCA for analysis at the pixel level (1×1), as well as when local 3×3, 5×5, and 7×7 neighbourhoods are used as input to a linear regression. It is seen that the performance gain converges above 100 spectral components even for increasing spatial sample sizes in the experiments. We have chosen 125 spectral MNF components for the experiments presented Section 3 based on the studies in Malmgren-Hansen et al. (2017), Camps-Valls et al. (2012) as it was an optimal trade off between accuracy and low number of components.

2.3. Regression models

In this work we use ordinary least squares (OLS) linear regression as a benchmark method to compare the results using CNNs. This is chosen out of consideration that it is a widely adopted and used model, not only atmospheric parameter retrieval, but also in many other fields. Further, we were able to implement a version of OLS that could be trained on the same large scale dataset of up to 525,000 samples as we used for the CNN, which enables direct comparison between the models. With other popular regression models such as kernel methods the memory consumption often makes sub-sampling of large scale datasets necessary in order to run it in practice. The OLS model, here in its simplest version, is for every sample, n, giving an estimate, $\hat{\mathbf{t}}_n$, of a target vector $\mathbf{t}_n$ with K elements as,

$$\hat{\mathbf{t}}_n = \mathbf{f}(\mathbf{x}_n) = \mathbf{W}\mathbf{x}_n + \mathbf{b} \tag{1}$$

where $\mathbf{x}_n$ is the n'th observation of size I input variables, $\mathbf{W}$ is a matrix of coefficients and $\mathbf{b}$ the model intercept. In our regression I would equal 125 decomposed spectral radiances times the number of local neighbourhood pixels (e.g. $125 \times 3 \times 3 = 1125$). Given all N observations, a closed form solution can be found to the minimization of the

residuals,

$$\arg\min_{\mathbf{W},\mathbf{b}} \quad \|\mathbf{t}_n - (\mathbf{W}\mathbf{x}_n + \mathbf{b})\|^2 \quad for \quad n = 1, ..,N \tag{2}$$

This gives a set of independent predictions for all target variables $t_{k,n}$, with $k = (1, ...,K)$. In our regression we are predicting 90 atmospheric temperatures, hence $K = 90$.

If we keep the assumption of output independence and further assume $t_{k,n}$ to be Gaussian distributed and represented by a deterministic function with noise added, $\mathbf{t}_n = \mathbf{f}(\mathbf{x}_n) + \mathbf{e}_n$, we see that the likelihood,

$$p(\mathbf{t}_n|\mathbf{x}_n) = \prod_{k=1}^{K} p(t_{k,n}|\mathbf{x}_n), \tag{3}$$

$$p(t_{k,n}|\mathbf{x}_n) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(f_k(\mathbf{x}_n) - t_{k,n})^2}{2\sigma^2}\right) \tag{4}$$

reduces to the following error function for maximizing the likelihood over all N observations,

$$E = \sum_{n=1}^{N}\sum_{k=1}^{K} (f_k(\mathbf{x}_n) - t_{k,n})^2 = \|\mathbf{f}(\mathbf{x}_n) - \mathbf{t}_n\|^2 \tag{5}$$

when we take the negative logarithm and remove additive and multiplicative constants, with f_k being a single target of the 90 atmospheric temperatures. Minimizing this error function is as well the most commonly used approach to regression with neural networks. Using the nonlinear function $\mathbf{y}(\mathbf{x}_n;\mathbf{W})$ in the minimization functional in Eq. (5) gives rise to a non-convex problem which cannot be solved analytically (Bishop, 2006). Typically gradient descent techniques are deployed here to learn the network parameters collectively summarized in $\mathbf{W}$. Note that if linearity is kept in the last layer of the neural network, i.e. no non-linear activation function is applied on the output, our model can be written as,

$$\mathbf{y}(\mathbf{x}_n;\mathbf{W}) = \mathbf{W}_L\mathbf{g}(\mathbf{x}_n;\mathbf{W}_{1,..,L-1}) + \mathbf{b}, \tag{6}$$

and we see that the last layer, L, of the neural network is a linear regression on a set of non-linear feature extractions from the previous $L - 1$ layers. When the first layers' weight vectors $\mathbf{W}_{1,..,L-1}$ are given, the last layer weights $\mathbf{W}_L$ can be found with a closed form solution as with the linear regression. This can be used to ensure the optimal set of parameters for the last layer after CNN training (Bishop, 1995) or used in a hybrid training algorithm as suggested in Webb and Lowe (1988).

The error in Eq. (5) corresponds to minimizing the variances of our estimated target functions given that each K outputs are independent and can be modelled with one global parameter for the variance,

$$\sigma_{err}^2 = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K} (y_k(\mathbf{x}_n;\mathbf{W}) - t_{k,n})^2 = \|\mathbf{f}(\mathbf{x}_n) - \mathbf{t}_n\|^2 \tag{7}$$

σ_{err}^2 here refers to the variance in the prediction error which our objective is to minimize. The assumption of output variable independence is not always true. In our case one could assume that nearby variables in the vertical atmospheric profile will be correlated resulting in a more complex likelihood function. For simplicity we will keep the assumption of independence, but other approaches could be adopted in future studies.

The purpose of comparing a linear model with a CNN for estimating atmospheric temperatures is to study how the spatial information in the data helps determining the optimal prediction model. In a linear model we can model local input correlations by concatenating a neighbourhood of spectral pixel values when predicting the center pixel. In a CNN all spatial content in the given input patch is mapped to a latent representation through a series of stacked convolutions, for which the kernel coefficients are a part of the parameters we optimize. If, e.g. the proximity of a coastline has a high influence on the target variable for the IASI data, a kernel in the CNN can learn to represent this feature in the latent representation, no matter where in the patch that the coastline appears.

To find the optimal set of weights for the CNN we use an iterative stochastic gradient descent (SGD) based update scheme. It is well known that estimating the error for all training samples in each iteration leads to slower convergence. For this reason a mini batch approach is used for training deep learning models. This, though, leads to more noisy estimates of the error function and methods to cope with this stochastic noise have been proposed. We use the method called ADAM (Kingma and Ba, 2014) with batch size equal to 128 samples and adopt the suggested parameters for learning rate, etc. This method applies exponential moving averages of the gradients and squared gradients which are used to ensure a smooth convergence of the training. Since our initial target state vector (temperatures, dew-point temperatures, etc.) can have different variances across the atmosphere, one could choose to normalize the target variables. This might lead to significantly different solutions in an SGD based scheme, as opposed to e.g. an SVD factorization of the ordinary least squares problem. The fact that target values might change scale can have a big impact on some SGD schemes since the gradient term scales as well. Unless accounted for in the learning rate, this will change the convergence of a solution. The ADAM scheme chosen for optimization in our experiments is practically invariant to scaling of the gradients due to its update rule based on first and second order moment vectors. These vectors impose an individual stepsize for each parameter in the network during the iterative parameter updates. In this work all CNN configurations were trained for 400 epochs without an early stopping scheme that requires a independent validation dataset. Neither, was extensive hyper-parameter tuning a part of this study which focuses on the retrieval with spatial inclusion.

3. Experimental results

The goal of our experiments is to demonstrate the advantages of CNNs for the retrieval of atmospheric variables from infrared sounders. In particular, we will illustrate how the networks include spatial regularization in a natural way. This feature results in improved prediction in the case of cloud coverage or noisy settings. Another advantage of the method is that cross-relations between the different atmospheric states are captured, so smoothness in the vertical profile is also achieved. Finally, we explore a very interesting possibility of the network to perform transfer learning, by which a network trained for example for temperature profile estimation can be re-used for moisture estimation with minor retraining.

3.1. Experimental setup

We will employ the data collected in 13 consecutive orbits within the same day, 17-08-2013, by the IASI sensor. Each orbit consists of approximately 92,000 samples. We use the first 7 orbits (which cover most of the Earth) for training and the last 6 for testing (which also cover most of the Earth). Fig. 3 shows the coverage of the two different sets of data taken on the same day.

3.2. Models

In order to investigate different levels of spatial information, four CNN models have been designed. Essential CNN features of the four proposed models including the full architecture description of each model is found in Table 1.

Our experiments consist of comparisons between CNN predictions with an ordinary least squares (OLS) linear regression model. We designed OLS models that also include spatial information from surrounding measurements. The linear model defined according to Eq. (1) can be extended to different spatial sample size by appending new variables to the columns of the data matrix $\mathbf{X}$ that holds N data sample vectors $\mathbf{x}_n$ for $n = 0, ...,N$ as rows. As the input dimensionality

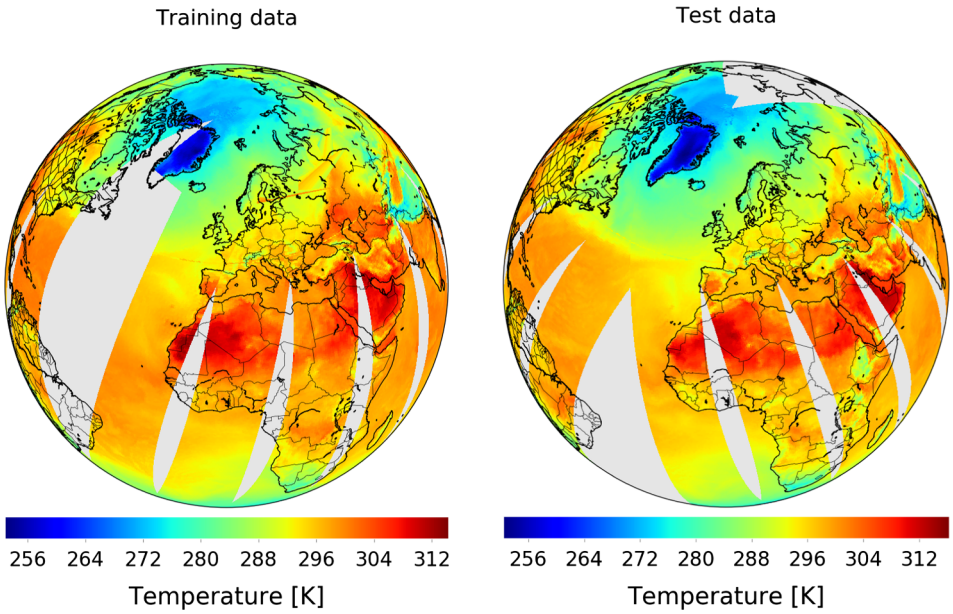

Fig. 3. Example of how we split the data: training (left) and test (right). Figures show surface temperatures for different orbits.

quadratically grows with increasing w, the size of the dataset limits the spatial extend that can be included in the regression. We have therefore limited the OLS regression to $w = 15$.

In particular, CNN A is trained on patches of $w = 3$. With a convolution kernel size, s, in the first layer of 3×3 coefficients this results in one valid convolution per patch. Practically this is equivalent to a multi layer perceptron network with 3×3 pixels concatenated as an input vector. CNN A therefore does not learn detection of the presence of nearby features in the same manner as the other CNN models, but we include it in our experiments to compare directly with the OLS for small patch sizes. In CNN B, C, D we keep s as 3×3 filters, while letting the patch size increase resulting in increased number of convolutions across the patch, i.e. we model local correlations (features) across an entire patch.

3.3. Retrieval performance and evaluation

In Table 2 the mean over the RMSE vertical profiles are given for our regression models with different sizes of w with the corresponding individual profiles shown in Fig. 4. In general, CNNs outperform linear regression models. CNN performance can be further improved by re-training the last layer once the regular training is done. According to Eq. (6) we can find the last CNN layer weights with the OLS algorithm when we fix the previous layers' weights. In our experiments, this procedure improved the CNN predictions with additionally 12–17% for all CNN architectures. It should be noted that the OLS generally performs worse than any CNN model except for at few very high altitudes. This is likely a consequence of optimizing the CNNs on the sum of squared errors over all altitudes. When optimizing over a sum of target errors we look for the best average solution rather than the best individual solution.

Let us now analyze some key aspects of the proposed CNN models: (1) the smoothness of the prediction profiles across space and vertical dimensions, and (2) the transferability of the models to be re-used in predicting other variables.

During a neural network optimization the average gradient of the error function is back propagated to update the weights. In this way we capture the best average solution to our regression. This is in contrast to a linear model where each output is an independent model of the input. In the case of atmospheric parameter retrieval where neighbouring targets are spatially dependent, the average gradient or the non-linearity in the CNN seems to smooth vertical predictions as well. Fig. 5 shows 4 transects of the mean error for a given path in an orbit of data. It can be seen that the linear model can obtain spatial smooth (horizontally) predictions by increasing the input patch size. The CNN ensures a smooth error profile both in the vertical and horizontal directions of the transect. The estimated cloud fraction is marked on each pixel of our dataset and can be seen as the white dashed line in Fig. 5. Though higher errors are generally expected in cloudy areas it seems that the correlation between the cloud fraction and the error are weak. We shall explore this further.

The water vapor profiles are often of interest as well for NWP models and so the prediction performance of these profiles is relevant too. Fig. 6 shows the profiles for the CNN and OLS models on 15x15 neighborhood pixel samples. Generally it is a harder problem to predict water vapor profiles and hence we see a higher error here than for atmospheric temperatures. The mean RMSE for the CNN profile is 3.34 K and 4.21 K for the OLS model.

3.3.1. Predictions over clouds

Predictions are inherently disturbed by strong attenuation or mixing of radiometric contributions from a large number of sources. It is commonly known that the presence of clouds attenuate the signal and hamper retrieval of parameters. Some approaches to temperature prediction typically act on cloud-free marked pixels only to be confident on the obtained predictions. Cloud masks can to some extent be estimated from optical sensors, but different approaches to generating cloud masks can have high influence on the final result. Since we are predicting the center pixel profile from a neighbourhood of pixels one could filter patches based on the amount of clouds. Nevertheless, in the results shown here, no such pre-filtering was performed, but CNNs figure out how to exploit the (possibly less cloudy) neighbouring radiances. Fig. 7 shows the error of a CNN on pixels with less than 50% clouds and pixels with more than 50% clouds. The cloud mask contains mostly 0% and 100% cloud fractions. For linear regression the difference between predicting over clouds or in cloud free areas is clear, around an increment of one degree of the error in lower atmospheric layers. In the case of CNNs this difference is less noticeable, around 0.25 degrees in the same area. An important thing to stress is that the CNNs model obtains less prediction error over cloudy areas than the linear model does over cloud free areas.

Table 1
Table of CNN architectures. B.R.D. is a concatenation of 3 layers, Batch Normalization, Rectified Linear Unit activation layer and Dropout. Dropout is performed with a probability $p = 0.5$ in all cases. The parameter column denote (number of channels in previous layer) x (filters in this layer) x (filter height) x (filter width).

CNN A		
Type	Parameters	Output
Input	–	125 × 3 × 3
Conv	125 × 60 × 3 × 3 + 60	60 × 1 × 1
B.R.D.	4 × 60	60 × 1 × 1
Conv	60 × 120 × 1 × 1 + 120	120 × 1 × 1
B.R.D.	4 × 120	120 × 1 × 1
Conv	120 × 240 × 1 × 1 + 240	240 × 1 × 1
B.R.D.	4 × 240	240 × 1 × 1
Conv	240 × 90 × 1 × 1 + 90	90 × 1 × 1

CNN B		
Type	Parameters	Output
Input	–	125 × 10 × 10
Conv	125 × 60 × 3 × 3 + 60	60 × 10 × 10
Conv	60 × 60 × 3 × 3 + 60	60 × 10 × 10
Pool	–	60 × 5 × 5
B.R.D.	4 × 60	60 × 5 × 5
Conv	60 × 120 × 3 × 3 + 120	120 × 5 × 5
Conv	120 × 120 × 3 × 3 + 120	120 × 3 × 3
Pool	–	120 × 1 × 1
B.R.D.	4 × 120	120 × 1 × 1
Conv	120 × 240 × 1 × 1 + 240	240 × 1 × 1
B.R.D.	4 × 240	240 × 1 × 1
Conv	240 × 90 × 1 × 1 + 90	90 × 1 × 1

Network	**CNN A**	**CNN B**
Number of parameters	127,290	347,070
Output dimension	90	
Optimizer	ADAM (Kingma and Ba, 2014)	
Approx. Training time	4 h	11 h
GPU Core Utilization	62%	82%
# train. samples	524,552	460,887
Mean test RMSE [K]	2.48	2.43

CNN C		
Type	Parameters	Output
Input	–	125 × 15 × 15
Conv	125 × 100 × 3 × 3 + 100	100 × 15 × 15
Conv	100 × 100 × 3 × 3 + 100	100 × 13 × 13
Pool	–	100 × 6 × 6
B.R.D.	4 × 100	100 × 6 × 6
Conv	100 × 160 × 3 × 3 + 160	160 × 4 × 4
Conv	160 × 160 × 3 × 3 + 160	160 × 2 × 2
Pool	–	160 × 1 × 1
B.R.D.	4 × 160	160 × 1 × 1
Conv	160 × 240 × 1 × 1 + 240	240 × 1 × 1
B.R.D.	4 × 240	240 × 1 × 1
Conv	240 × 90 × 1 × 1 + 90	90 × 1 × 1

CNN D		
Type	Parameters	Output
Input	–	125 × 25 × 25
Conv	125 × 100 × 3 × 3 + 100	100 × 23 × 23
Conv	100 × 100 × 3 × 3 + 100	100 × 21 × 21
Pool	–	100 × 10 × 10
B.R.D.	4 × 100	100 × 10 × 10
Conv	100 × 160 × 3 × 3 + 160	160 × 8 × 8
Conv	160 × 160 × 3 × 3 + 160	160 × 6 × 6
Pool	–	160 × 3 × 3
B.R.D.	4 × 160	160 × 3 × 3
Conv	160 × 200 × 3 × 3 + 200	200 × 1 × 1
B.R.D.	4 × 200	200 × 1 × 1
Conv	200 × 240 × 1 × 1 + 240	240 × 1 × 1
B.R.D.	4 × 240	240 × 1 × 1
Conv	240 × 90 × 1 × 1 + 90	90 × 1 × 1

Network	**CNN C**	**CNN D**
Number of parameters	639,750	938,350
Output dimension	90	
Optimizer	ADAM (Kingma and Ba, 2014)	
Approx. Training time	18 h	36 h
GPU Core Utilization	80%	85%
# train. samples	415,472	324,792
Mean test RMSE [K]	2.19	2.28

Table 2
Summary of the mean RMSE (across the atmosphere profile) on temperature prediction. Each row represents a different model and each column is the effect of chaning the patch size. The CNN + Opt. row is the same network as the first but where the last layer is optimized with the closed form least square solution after training.

Patch Size	1 × 1	3 × 3	5 × 5	7 × 7	10 × 10	15 × 15	25 × 25
CNN	–	2.48	–	–	2.43	2.20	2.28
CNN + Opt.	–	2.11	–	–	2.01	**1.94**	2.01
OLS	3.30	3.00	2.91	2.86	2.84	2.85	–

3.3.2. Predictions over land

Prediction over land is typically more challenging than over ocean, mainly due to the more varying conditions, landscape and land cover, and changes in bodies' emissivities. We aimed to study the performance of algorithms as a function of the land cover per pixel. Fig. 8 shows the error in predictions from a linear model and a CNN with 15 × 15 pixel input patch size, conditioned on the land fraction. The land fraction mask contains mostly 0% or 100% values, but some coastal areas are given as intermediate values due to the resolution cell covering both land and sea. On the other hand, the land fraction has a high influence on the predictions, and continues to be a challenge for precise predictions of atmospheric temperature profiles.

3.4. Transfer learning

The concept of transfer learning within deep learning has proven useful for a range of computer vision tasks. Deep CNNs trained on large databases of natural images can be transferred to smaller datasets for specific applications with high end performance.

There are two overall different approaches to transfer learning. One, as in Yosinski et al. (2014), where the training of a Network is repeated on a new dataset but starting with the weights found solving the first

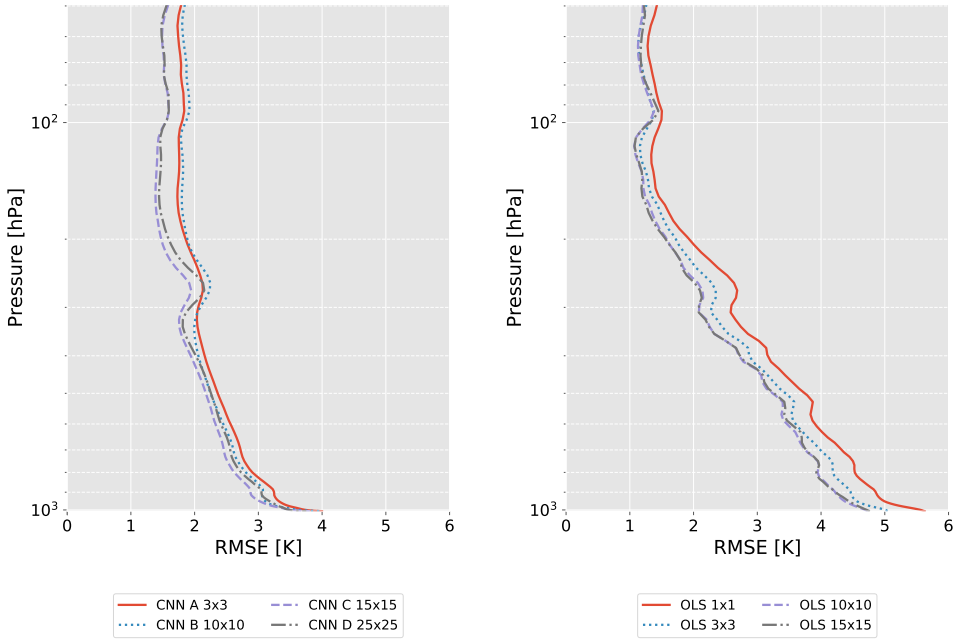

Fig. 4. RMSE error profiles of model prediction for both CNNs (left) and linear models (right) at different input spatial patch sizes. The CNN generally outperform OLS regression except at very high altitudes. The temperatures at lower altitudes (> 200 hPa) are the most important for meteorological models.

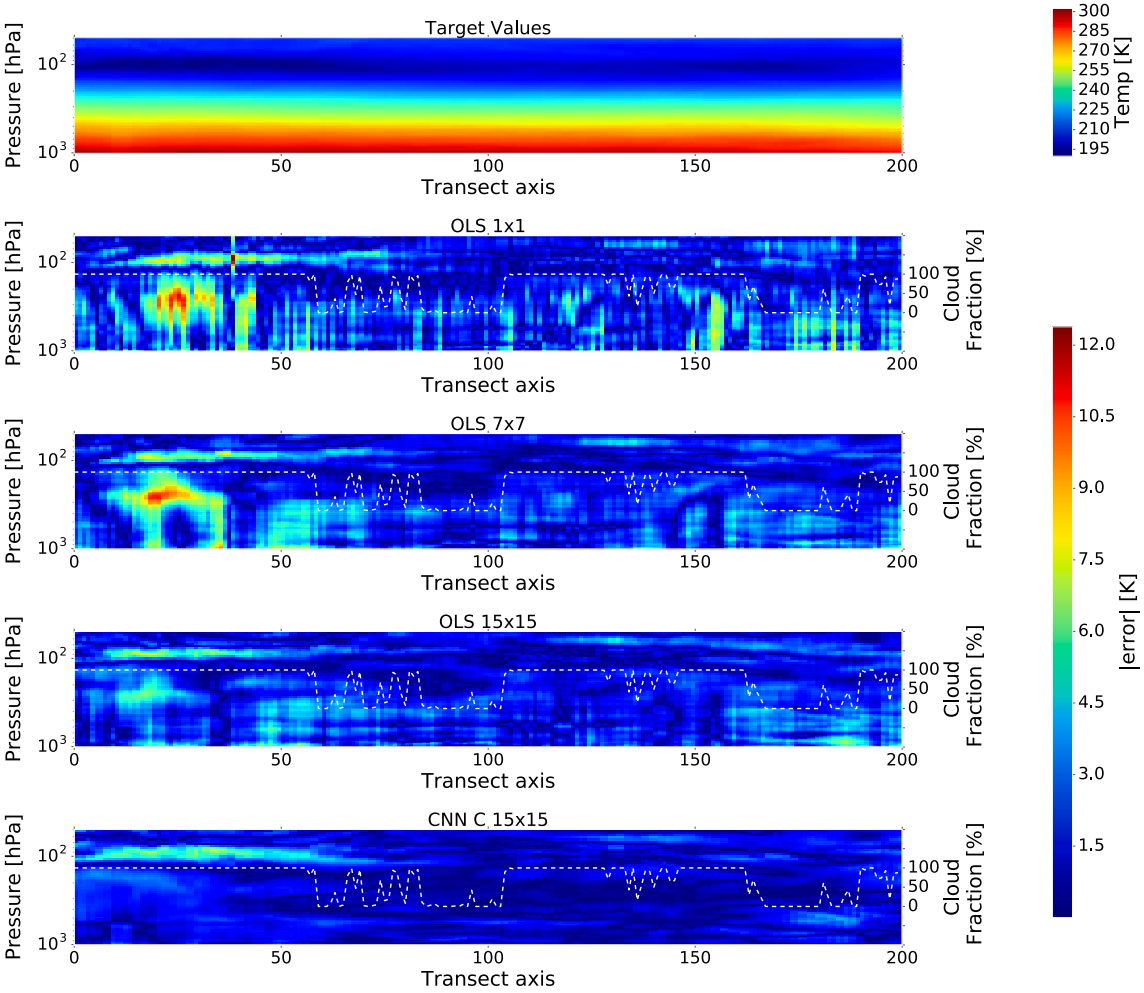

Fig. 5. Top plot shows the target temperature along a transect profile, lower four plots shows the transect profile of the prediction error from different regression models. White dashed line is the cloud fraction, i.e. the percentage on cloud each input sample is marked with. The y-axis is the altitude pressure level.

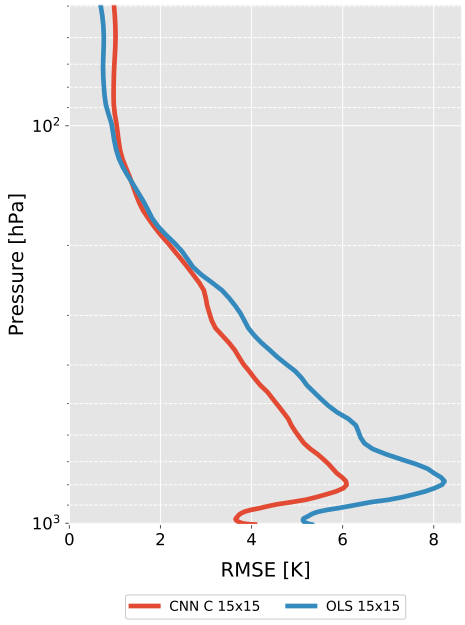

Fig. 6. Water vapor atmospheric profiles for the best performing CNN and Linear model versions. Unit is dew point temperature in Kelvin.

problem. The second is to consider a part of the network a feature extractor and access the latent representation learned from one dataset and classifier to solve a problem on a second related dataset (Sharif Razavian et al., 2014).

The purpose of exploring transfer learning in our setup is not to unveil whether cross domain features can be learned. Instead, we explore the ability of a model trained to predict atmospheric temperature profiles to be transferred to other output variables, such as moisture profiles. Possible benefits are shorter training time, as well as higher accuracy.

Fig. 9 shows faster training convergence on predicting dew point temperatures when considering initialization of a CNN with either weights from a network trained on atmospheric temperatures or a standard random initialization for training from scratch. The figure shows that the performance reached by CNN initialized from random weights can be reached in less than around $\frac{1}{8}$ (50 epochs instead of 400) of the training time if the weights are transferred from a model trained for another output variable.

Considering a model trained on atmospheric temperatures, a feature extractor for a linear regression to predict dewpoint temperatures can as well be done, and this approach is conceptually closer to the one proposed in Sharif Razavian et al. (2014). RMSE profiles from the transfer learning experiments are shown in Fig. 10.

The red and blue profiles in Fig. 10 show that we reach the same performance whether we start with a model trained on atmospheric temperatures or random weights, when predicting dewpoint temperatures. This is not surprising since it is the same dataset we fit the models on, all we change is the response variable. When considering the second transfer learning approach where a CNN trained on temperature prediction is used as a feature extractor with a linear regression to predict dewpoint we reach a less optimal solution. The grey profile in Fig. 10 shows the second transfer learning approach and training a linear regression directly on the input radiance is shown as the purple profile. At low altitudes we get better accuracy than the shallow linear regression model ($> 1^\circ$ RMSE terms). At higher altitudes though, the second transfer learning approach does it worse. Fine tuning for a specific output variable is necessary in order to achieve good predictions. The high RMSE at mid range altitudes is caused by the frequent presence of clouds in this range, i.e. higher absolute dew point values.

4. Conclusion and discussion

We present for the first time the use of deep convolutional neural networks for the retrieval of atmospheric profiles from infrared sounding data, particularly for IASI data. The proposed scheme performs multidimensional output nonlinear regression, accounts for noise features, and exploits correlations in all dimensions. Good experimental results were obtained compared widely adopted OLS approaches despite also adapting neighbourhood pixel in OLS regression. Networks were trained on full orbits and tested out of sample with great accuracy. We also observed a huge benefit in accuracy when predicting over clouds, increasing the yield by 34% over linear regression. The proposed scheme is modular and allows us to predict related variables from an already trained model. We also illustrated this by exploiting the learned network to predict temperature profile and retraining it to fit moisture (dew point temperature) profiles. Good results were obtained

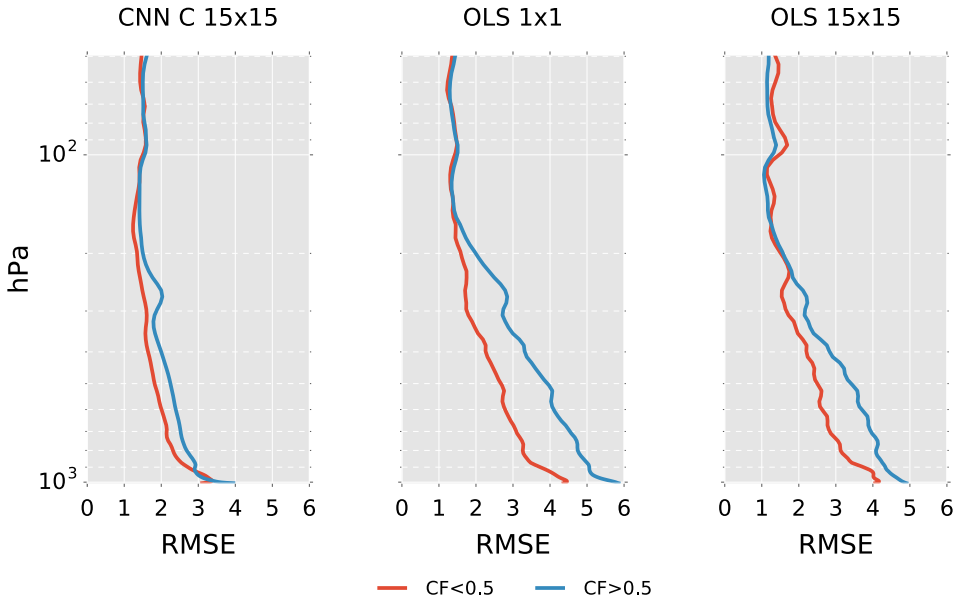

Fig. 7. RMSE profile when testing on cloudy samples (CF $>$ 0.5) versus samples marked cloud free (CF $<$ 0.5).

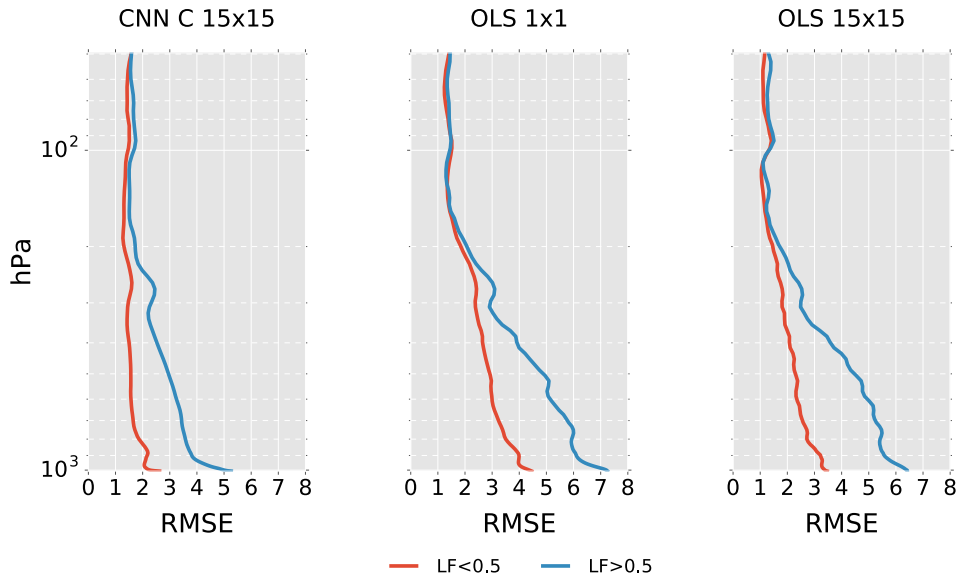

Fig. 8. RMSE profiles when predicting temperature profiles over land (LF $>$ 0.5) and over sea (LF $<$ 0.5).

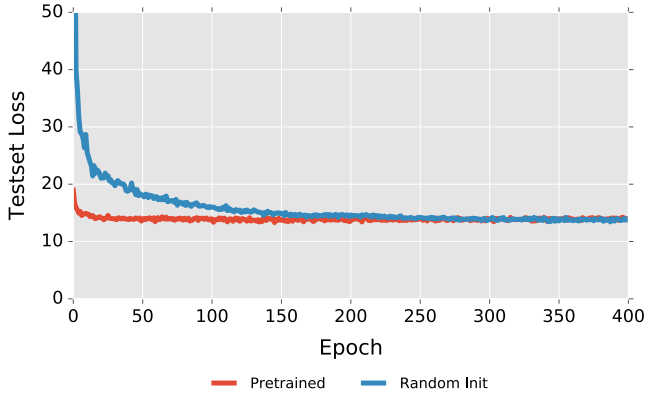

Fig. 9. Test error convergence during training for dewpoint temperature prediction. Blue curve is a CNN initialized with random weights and the red is a CNN initialized with the weights for a model that predicts air temperatures. Both models converge to a mean RMSE error of 3.34 K after 400 epochs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

too, which demonstrates that the learned features by the network impose a sort of spatial and vertical smoothness that can be exploited for other state variables that share these features, such as some trace gases as well. We conclude that a deep network is an appropriate learning paradigm for statistical retrieval of atmospheric profiles.

There are several aspects of the modeling to explore in the future to improve the statistical retrieval. It would be relevant to explore model architectures that directly model the output correlations. This could be done with the neural network by including the joint probabilities between neighbouring targets at the expense of a more complicated error function. Alternatively one can predict the difference between neighbouring target variables rather that their value and, in this way, incorporate neighbourhood correlation in the targets. We have shown that there is a high potential for models that incorporate feature extracting abilities as well as capabilities of modeling non-linear phenomena in statistical retrieval. Exploring a greater variety of dimensionality reduction techniques, such as the spatio-spectral supervised technique presented in Huang et al. (2019), or incorporating it into the CNN architecture is also an area to further explore. Finding optimal architectures for CNNs remains an open task in the deep learning literature, and due to the non-convexity of the problem,

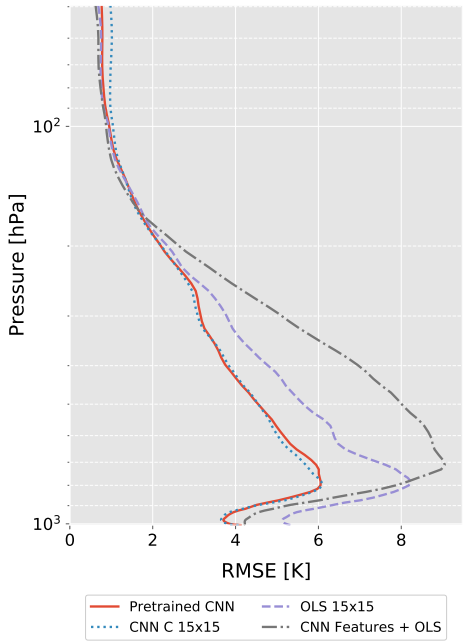

Fig. 10. Dew point temperature RMSE profiles of transfer learning regression models. The features learned on temperatures are poor for dewpoint prediction (grey profile) unless fine tuned (red profile). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

experiments are the only way to find optimal models. In this work, a few architectures have been explored but a larger analysis of this is highly relevant for the application. Further, results in this work were performed on a single day dataset, and larger datasets that capture more variances, such as (monthly, yearly) temporal variations are needed regardless of the chosen method. Recent alternatives on efficient training of convolutional nets could relieve the induced complexity (Giusti et al., 2013; Sermanet et al., 2013; Kampffmeyer et al., 2016). Further work could also includes deeper comparison of the results with other methods and radiosonde measurements as in Jiménez-Muñoz et al. (2010), Sobrino et al. (2015), Zhang et al. (2015), Julien et al. (2015) to get a better measure of in-application performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We want to thank Dr. Thomas August (EUMETSAT, Germany) and Xavier Calbet (AEMET, Spain) for the provided data and the fruitful discussions on atmospheric retrievals. This work was supported in part by the Spanish Ministry of Economy and Competitiveness and by the European Regional Development Fund under Grant TIN2015-71126-R and by the European Research Council under Consolidator Grant SEDAL ERC-2014-CoG 647423. The Ph.D. project of D. Malmgren-Hansen was jointly funded by the Terma A/S and the Innovation Fund Denmark Grant 4135-00041B under the Danish program for industrial Ph.D. projects.

References

Aires, F., Rossow, W., Scott, N., Chédin, A., 2002. Remote sensing from the infrared atmospheric sounding interferometer instrument 2. simultaneous retrieval of temperature, water vapor, and ozone atmospheric profiles. J. Geophys. Res.: Atmos. 107 (D22).

Aptoula, E., Ozdemir, M.C., Yanikoglu, B., 2016. Deep learning with attribute profiles for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 13 (12), 1970–1974.

August, T., Klaes, D., Schlüssel, P., Hultberg, T., Crapeau, M., Arriaga, A., O'Carroll, A., Coppens, D., Munro, R., Calbet, X., 2012. IASI on Metop-A: operational level 2 retrievals after five years in orbit. J. Quant. Spectrosc. Radiat. Transfer 113 (11), 1340–1371.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press.

Bishop, C.M., 2006. Pattern recognition. Mach. Learn. 128.

Blackwell, W.J., Pieper, M., Jairam, L., 2008. Neural network estimation of atmospheric profiles using AIRS/IASI/AMSU data in the presence of clouds. In: Larar, A.M., Lynch, M.J., Suzuki, M. (Eds.), Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques, and Applications II. Proceedings of SPIE, vol. 7149 Bellingham, WA.

Camps-Valls, G., Bruzzone, L. (Eds.), 2009. Kernel Methods for Remote Sensing Data Analysis. Wiley & Sons, UK.

Camps-Valls, G., Gomez-Chova, L., Munoz-Mari, J., Vila-Frances, J., Calpe-Maravilla, J., 2006. Composite kernels for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 3 (1), 93–97 cited By 341.

Camps-Valls, G., Tuia, D., Gómez-Chova, L., Jiménez, S., Malo, J., 2011. Remote Sensing Image Processing. Synthesis Lectures on Image, Video, and Multimedia Processing. Morgan & Claypool Publishers.

Camps-Valls, G., Muñoz-Marí, J., Gómez-Chova, L., Guanter, L., Calbet, X., 2012. Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data. IEEE Trans. Geosci. Remote Sens. 50 (5 PART 2), 1759–1769. https://doi.org/10.1109/TGRS.2011.2168963. http://www.scopus.com/inward/record.url?eid=2-s2.0-84860333725&partnerID=40&md5=79b8f967ca53944341da6a574d036f77.

Camps-Valls, G., Muñoz-Marí, J., Gomez-Chova, L., Guanter, L., Calbet, X., 2012. Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data. IEEE Trans. Geosci. Remote Sens. 50 (5), 1759–1769.

Camps-Valls, G., Tuia, D., Bruzzone, L., Benediktsson, J., 2014. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. IEEE Signal Process. Mag. 31 (1), 45–54.

Camps-Valls, G., Verrelst, J., Muñoz-Marí, 2016. A survey on gaussian processes for earth observation data analysis: A comprehensive investigation. IEEE Geosci. Remote Sensing Mag.(6).

Collard, A.D., McNally, A.P., Hilton, F.I., Healy, S.B., Atkinson, N.C., 2010. The use of principal component analysis for the assimilation of high-resolution infrared sounder observations for numerical weather prediction. Quart. J. Roy. Meteorol. Soc. 136 (653), 2038–2050. https://doi.org/10.1002/qj.701.

EUMETSAT, 2014. IASI Level 1: Product Guide, EUM/OPS-EPS/MAN/04/0032.

EUMETSAT, 2017. IASI Level 2: Product Guide, EUM/OPS-EPS/MAN/04/0033.

García-Sobrino, J., Serra-Sagristà, J., Laparra, V., Calbet, X., Camps-Valls, G., 2017. Statistical atmospheric parameter retrieval largely benefits from spatial–spectral image compression. IEEE Trans. Geosci. Remote Sens. 55 (4), 2213–2224.

Garcí a-Sobrino, J., Laparra, V., Serra-Sagristà, Calbet, X., Camps-Valls, G., 2019. Improved statistically-based retrievals via spatial-spectral data compression for IASI data. IEEE Trans. Geosci. Remote Sens. PP (99), 1–12.

Geng, J., Fan, J., Wang, H., Ma, X., Li, B., Chen, F., 2015. High-resolution sar image classification via deep convolutional autoencoders. IEEE Geosci. Remote Sens. Lett. 12 (11), 2351–2355.

Giusti, A., Ciresan, D.C., Masci, J., Gambardella, L.M., Schmidhuber, J., 2013. Fast image scanning with deep max-pooling convolutional neural networks. In: 2013 20th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 4034–4038.

Golub, G.H., Van Loan, C.F., 1996. Matrix Computations Vol. 3 JHU Press.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Edu. Psychol. 24 (6), 417.

Huang, H., Duan, Y., He, H., Shi, G., Luo, F., 2019. Spatial-spectral local discriminant projection for dimensionality reduction of hyperspectral image. ISPRS J. Photogramm. Remote Sens. 156, 77–93.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons.

Jiménez-Muñoz, J.C., Sobrino, J.A., Mattar, C., Franch, B., 2010. Atmospheric correction of optical imagery from modis and reanalysis atmospheric products. Remote Sens. Environ. 114 (10), 2195–2210. https://doi.org/10.1016/j.rse.2010.04.022. http://www.sciencedirect.com/science/article/pii/S0034425710001355.

Julien, Y., Sobrino, J.A., Mattar, C., Jiménez-Muñoz, J.C., 2015. Near-real-time estimation of water vapor column from msg-seviri thermal infrared bands: implications for land surface temperature retrieval. IEEE Trans. Geosci. Remote Sens. 53 (8), 4231–4237. https://doi.org/10.1109/TGRS.2015.2393378.

Kampffmeyer, M., Salberg, A.-B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

Laparra, V., Santos-Rodríguez, R., 2016. Spatial/spectral information trade-off in hyperspectral images. In: Proceedings IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1124–1127.

Laparra, V., Malo, J., Camps-Valls, G., 2015. Dimensionality reduction via regression in hyperspectral imagery. IEEE J. Sel. Top. Signal Process. 9 (6), 1026–1036.

Laparra, V., Muñoz-Marí, J., Gómez-Chova, L., Calbet, X., Camps-Valls, G., 2017. Nonlinear statistical retrieval of surface emissivity from iasi data. In: IEEE International and Remote Sensing Symposium (IGARSS).

Larsen, R., 2002. Decomposition using maximum autocorrelation factors. J. Chemom. 16 (8–10), 427–435. http://www2.imm.dtu.dk/pubdb/p.php?209.

Luus, F.P.S., Salmon, B.P., van den Bergh, F., Maharaj, B.T.J., 2015. Multiview deep learning for land-use classification. IEEE Geosci. Remote Sens. Lett. 12 (12), 2448–2452.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55 (2), 645–657.

Malmgren-Hansen, D., Laparra, V., Aasbjerg Nielsen, A., Camps-Valls, G., 2017. Spatial noise-aware temperature retrieval from infrared sounder data. IEEE Int. Geosci. Remote Sens. Symp.

Plaza, A., Martínez, P., Pérez, R., Plaza, J., 2002. Spatial/spectral endmember extraction by multidimensional morphological operations. IEEE Trans. Geosci. Remote Sens. 40 (9), 2025–2041.

Rivera-Caicedo, J.P., Verrelst, J., Muñoz-Marí, J., Camps-Valls, G., Moreno, J., 2017. Hyperspectral dimensionality reduction for biophysical variable statistical retrieval. ISPRS J. Photogramm. Remote Sens. 132, 88–101.

Romero, A., Gatta, C., Camps-Valls, G., 2016. Unsupervised deep feature extraction for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 54 (3), 1349–1362.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229.

Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806–813.

Sobrino, J.A., Jiménez-Muñoz, J.C., Mattar, C., Sòria, G., 2015. Evaluation of terra/modis atmospheric profiles product (mod07) over the iberian peninsula: a comparison with radiosonde stations. Int. J. Digital Earth 8 (10), 771–783. https://doi.org/10.1080/17538947.2014.936973.

Tournier, B., Blumstein, D., Cayla, F., Chalon, G., 2002. IASI level 0 and 1 processing algorithms description. In: Proc. of ISTCXII Conference.

Tuia, D., Ratle, F., Pozdnoukhov, A., Camps-Valls, G., 2010. Multisource composite kernels for urban-image classification. IEEE Geosci. Remote Sens. Lett. 7 (1), 88–92 cited By 42.

Webb, A., Lowe, D., 1988. A hybrid optimisation strategy for adaptive feed-forward layered networks, Tech. rep., DTIC Document.

Whitburn, S., Van Damme, M., Clarisse, L., Bauduin, S., Heald, C., Hadji-Lazaro, J., Hurtmans, D., Zondlo, M., Clerbaux, C., Coheur, P.-F., 2016. A flexible and robust neural network iasi-nh3 retrieval algorithm. J. Geophys. Res.: Atmos. 121 (11), 6581–6599.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 3320–3328.

Zhang, X., Pang, J., 2015. A comparison between atmospheric water vapour content retrieval methods using msg2-seviri thermal-ir data. Int. J. Remote Sens. 36 (19–20), 5075–5086. https://doi.org/10.1080/01431161.2015.1041180.

Zhang, F., Du, B., Zhang, L., 2016. Scene classification via a gradient boosting random convolutional network framework. IEEE Trans. Geosci. Remote Sens. 54 (3), 1793–1802.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci.d Remote Sens. Mag. 4 (2), 22–40.