Deep learning to estimate model biases in an operational NWP assimilation system

Patrick Laloyaux[1], Thorsten Kurth[2], Peter Dominik Dueben[1]and David Hall[2]

[1]European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom
[2]Nvidia Corporation, Santa Clara, California, United States

Key Points:

- Temperature retrievals from radio occultation measurements can be used as ground truth to measure stratospheric model biases
- 3D convolutional neural networks are suitable for model bias estimation but do not outperform weak-constraint 4D-Var
- Transfer learning can help to mitigate data limitations when the atmospheric model is upgraded

Corresponding author: Patrick Laloyaux, `European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, United Kingdom, Email: patrick.laloyaux@ecmwf.int`

Abstract

Model error is one of the main obstacles to improved accuracy and reliability in numerical weather prediction (NWP) conducted with state-of-the-art atmospheric models. To deal with model biases, a modification of the standard 4D-Var algorithm, called weak-constraint 4D-Var, has been developed where a forcing term is introduced into the model to correct for the bias that accumulates along the model trajectory. This approach reduced the temperature bias in the stratosphere by up to 50% and is implemented in the ECMWF operational forecasting system.

Despite different origins and applications, Data Assimilation and Deep Learning are both able to learn about the Earth system from observations. In this paper, a deep learning approach for model bias correction is developed using temperature retrievals from Radio Occultation (RO) measurements. Neural Networks require a large number of samples to properly capture the relationship between the temperature first-guess trajectory and the model bias. As running the IFS data assimilation system for extended periods of time with a fixed model version and at realistic resolutions is computationally very expensive, we have chosen to train, the initial Neural Networks are trained using the ERA5 reanalysis before using transfer learning on one year of the current IFS model. Preliminary results show that convolutional neural networks are adequate to estimate model bias from RO temperature retrievals. The different strengths and weaknesses of both deep learning and weak constraint 4D-Var are discussed, highlighting the potential for each method to learn model biases effectively and adaptively.

Plain Language Summary

The state of the Earth system is estimated via a combination of information from both previous weather predictions and Earth system observations. This complex, mathematical procedure is called data assimilation. Weather predictions could be improved if the error of the numerical models that are used could be reduced. Recent advances in data assimilation at the European Centre for Medium-Range Weather Forecasts (ECMWF) indicate that it is possible to estimate and correct for a large fraction of systematic model errors of those models. During data assimilation, the forecast model and Earth system observations are representing the same situation of the global atmosphere. A direct comparison between models and observations during the short time interval of the data assimilation process can be used to diagnose model errors.

Deep learning is a comparably new method from machine learning that can be used to learn complex mapping procedures. The question we address in this paper is whether deep learning techniques can be used to predict model errors when they are trained to predict the mapping between the global temperature and the model error that was diagnosed during data assimilation.

1 Introduction

Machine learning (ML) has made rapid progress in many domains including natural language processing, computer vision, autonomous driving, healthcare and finance (Goodfellow et al., 2016). Machine learning applications can be very complex, and neural networks (NN) can consist of millions to billions of trainable parameters, large numbers of layers, and specialised architectures. In recent years, the weather and climate modelling community has started to explore machine learning techniques with many applications in Numerical Weather Predictions (NWP) (Dueben et al., 2021). In general, these applications can be divided into three groups: methods that improve computational efficiency, methods that improve the quality of the prediction system,

and methods that help improve our understanding of the Earth system, for example via unsupervised learning and causal discovery. This paper belongs to the group that aims to improve prediction quality. In particular, we will use deep learning to learn the systematic error of weather forecast models. Attempts to use DL techniques to estimate and correct for model errors have already been documented in the geophysical literature. For example, Watson (2019) uses an Artificial Neural Network (ANN) to estimate model error tendencies in the Lorenz-96 system. Predicting the error via deep learning is appealing, as errors can often be measured but are typically not easily described by a formula or theory, which makes them difficult to approach using conventional methods.

While there are several papers that learn the error during post-processing of the model output (Rasp & Lerch (2018); Groenquist et al. (2021)), this paper will investigate learning model error within the data assimilation (DA) framework of the European Centre for Medium-Range Weather Forecasts (ECMWF). DA is the process that involves merging information from observations with previous model predictions to generate initial conditions for weather forecasts that are both close to the observations and consistent with the state of the forecast models, in order to avoid shocks during model initialisation.

Weather observations make a crucial contribution to the quality of today's numerical weather forecasts. Satellites carry passive instruments (e.g. infrared or microwave) to measure natural radiation, while active instruments (e.g. scatterometer or lidar) probe the surface, clouds, and winds by sending out signals and measuring the backscatter (Saunders, 2021). Radio occultation observations evaluate signals sent from one satellite to another (Kursinski et al. (1997)). This array of satellite observations is complemented by a network of in-situ measurements coming from various platforms (e.g. surface stations, aircraft or radiosondes) with a rather inhomogeneous distribution compared to satellite data (Haiden et al., 2018). However, observations are inadequate to provide a complete and accurate picture of the state of the Earth system across the globe at a given point in time. The current model used in operations at the ECMWF contains almost one billion grid points that are updated several times per hour, while only 40 million observations are processed every 12 hours. For this reason, the DA community came up with methods to estimate the most likely state of a system by combining different imperfect sources of information. On the one hand, most observations are unevenly distributed in space and time. They come with errors, and they do not measure the prognostic model variables directly. Instead, they measure quantities linked to these variables, such as radiances or radar echoes. On the other hand, NWP models include the dynamics of the atmosphere and the physical processes that occur. DA combines observations and models in a way that accounts for the uncertainties in each. A popular DA algorithm is the four-dimensional variational (4D-Var) method that iteratively adjusts the initial conditions of a short-range forecast to bring it into closer agreement with meteorological observations in space and time (Rabier et al., 2000).

4D-Var is particularly well-suited to satellite data assimilation as it include a radiative transfer model that simulates the top of atmosphere radiances, which are compared to the observed radiances from a specific instrument. This enables the direct application of satellite measurements and extracts the maximum amount of information in clear-sky or all-sky conditions (A. J. Geer et al., 2018). Dealing with random and systematic errors in observations and models is critical for computing an accurate and unbiased estimate. For this reason, an observation error covariance matrix is introduced in the 4D-Var formulation to take into account stochastic observation errors arising from the instruments and from the observation operator (Janjic et al., 2018). The error covariance matrix can also represent spatial and inter-channel cross-correlations between observation errors (Waller et al., 2014). Similarly, a background

error covariance matrix is implemented to represent flow-dependent, spatially-random errors in the short-range forecast used in 4D-Var (Bonavita et al., 2016). This matrix weights the importance of the a-priori state and distributes information horizontally and vertically in space as well as between model variables (Bannister, 2008a,b). To deal with systematic observation errors, ECMWF played a pioneering role in the development of the Variational Bias Correction (VarBC) schemem, which is embedded in 4D-Var and automatically removes biases coming from observations and radiative transfer models. Similarly, the short-range forecast used in 4D-Var also contains systematic errors which grow over time. A weak-constraint 4D-Var formulation has been proposed to estimate these model biases within the assimilation process and to correct the dynamical model accordingly (Laloyaux, Bonavita, Dahoui, et al., 2020).

There are strong mathematical similarities between the 4D-Var formulation in data assimilation and the training of NNs. Both use gradient descent techniques, and the adjoint method for calculating gradients in 4D-Var is mathematically identical to the standard backpropagation method used in NN training. From a broad enough viewpoint, DA and ML may be viewed as two flavours of inverse methods that can be united under Bayesian statistics (A. Geer, 2020). Brajard et al. (2020) demonstrated a way to combine ML with DA when observations are noisy and partial. In their scheme, DA and ML alternate and compute progressively more accurate estimates of the state and of the surrogate predictive model. Following this idea, (Farchi, Laloyaux, et al., 2021) used a dataset of analysis increments to train a ML statistical/empirical model that complements the original dynamical model. The resulting hybrid surrogate model significantly improves the accuracy of the analysis and produces better short- and mid-range forecasts in a two-layer, two-dimensional, quasi-geostrophic channel model. These encouraging results with a simplified system have been confirmed to a certain extent in the operational atmospheric Integrated Forecasting System (IFS) model developed at ECMWF (Bonavita & Laloyaux, 2020). The idea of using time series of analysis increments fields to estimate the predictable component of model error is not new in the meteorological literature. For example, one of the algorithms proposed in (Dee, 2005) for the correction of model bias in a cycled data assimilation framework explicitly involves using an online model error estimate based on a running mean over past analysis increments. The increments have global, homogeneous coverage and are already available in the space of the dynamical model variables which makes the method easy to implement. However, this approach also has some limitations, as increments can contain signals that are not induced by model biases but by other error sources that have not been properly accounted for in the DA system. A well-known illustration is the positive temperature increment in the ERA-Interim reanalysis coming from aircraft temperature biases that have not been corrected properly by VarBC (Dee & Uppala, 2009).

This paper will focus on the estimation and correction of temperature systematic errors (bias) in the stratosphere using satellite temperature retrievals as ground truth. But how important are these biases for NWP? In a global NWP model, the troposphere may be viewed as a turbulent boundary layer for the atmosphere, and the stratosphere as being comparatively isolated from the surface of the Earth. To a first approximation, the global-mean stratosphere is in radiative equilibrium, with long-wave cooling balancing solar heating through ozone absorption (Fomichev et al., 2002). The latitudinal temperature structure is affected by the meridional circulation which is driven by breaking and dissipating planetary and gravity waves in the stratosphere. To quantify how stratospheric biases influence the troposphere, we ran a denial experiment and blacklisted observations that are important for stratospheric variables. This includes stratospheric observations from radiosondes, aircraft, RO bending angles above 100hPa, as well as the microwave and infrared stratospheric channels (see details in Table1). It is not possible to remove all observations that are sensitive to the stratospheric conditions, as microwave instruments measure radiances that reflect con-

Type	Pressure/Altitude/Channels
Radiosonde	above 100hPa
Aircraft	above 100hPa
RO	above 17km
AMSU-A	9,10,11,12,13,14
ATMS	10,11,12,13,14,15
AIRS	7, 15, 20, 21, 22, 27, 28, 40, 52, 69, 72, 92, 93, 98, 99, 104, 105, 110, 111, 116, 117, 123, 128, 129
IASI	16, 38, 49, 51, 55, 57, 59, 61, 63, 66, 70, 72, 74, 79, 81, 83, 85, 87, 89, 101, 104, 106, 109, 111, 113, 116, 119, 122, 125, 128, 131, 133,138, 135, 141, 144, 146, 148, 151, 154, 157, 159, 161, 163, 165, 167, 170, 176, 178, 183, 189, 191, 195, 197, 201, 203, 301, 303
CrIS	20, 23, 26, 33, 36, 39, 42, 45, 48, 51, 54, 57, 60, 61, 62, 63, 64, 65, 66, 68, 69, 70, 71, 73, 74, 113, 114

Table 1. List of all the observations considered as sensitive to stratospheric conditions and withheld in the denial experiments

ditions in a deep layer of the atmosphere. This means that some tropospheric-peaking channels could still have a slight impact on the stratosphere.

The data-denial experiment runs over two months, between the 25th of January 2020 and the 25th of March 2020. The top panel of Figure1 shows the impact on the analysis mean error when stratospheric observations are withheld. The large biases developed in the stratosphere over these two months are transferred to the troposphere, especially over the Southern pole. The bottom panel of Figure1 shows how these biases present in the analysis evolution during forecasts. The impact of the missing stratospheric observations can still be observed after 48 hours. The impact shrinks with the forecast lead time as the model drifts towards its climatology and forgets about the information present in the initial conditions. This experiment shows the importance of tackling residual stratospheric temperature biases as they can descend into the troposphere.

It is important to note, that the model bias changes when the IFS model is upgraded, on a regular basis. The most recent improvements to the stratospheric physics are the implementation of a new radiation scheme and ozone climatology in cycle 46r1 (Hogan et al., 2017; Shepherd et al., 2018). Furthermore, a quintic vertical interpolation has been implemented in the semi-Lagrangian advection in cycle 47r1 (Polichtchouk et al., 2019) to resolve a larger fraction of gravity waves in the vertical direction. These changes reduced the temperature bias in the stratosphere, but the residual bias is still significant. It consists of a global-mean cold bias in the lower/mid stratosphere of -0.5C and a global-mean warm bias in the upper stratosphere of 0.5C that accumulate over a 12-hour data assimilation window.

This article develops a deep learning solution for estimating the three-dimensional stratospheric temperature bias in the IFS. State-of-the-art NNs are trained to learn the mapping from three-dimensional fields of stratospheric temperature to the three-dimensional bias diagnosed via Radio Occultation (RO) temperature retrievals. As a first step, we use information from ERA5 reanalysis to show that deep learning can indeed learn to predict the three-dimensional temperature bias of short-term forecasts when using a large training data set spanning several years. In a second step, we study the use of transfer learning to adjust the trained model when only one year of training data is available for a new model cycle. Finally, we perform tests that apply

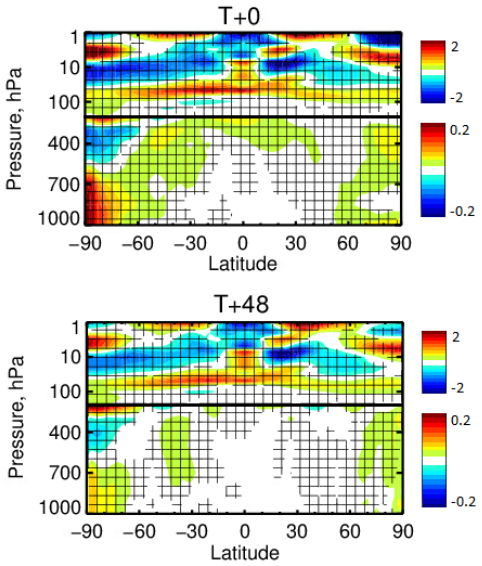

Figure 1. Difference in the forecast mean error across zonal bands at lead time +0h (top) and +48h (bottom) when all the stratospheric observations valid above 100hPa are withheld. Scores have been computed against the operational analysis between January 25th, 2020 and March 25th, 2020. Different colorbars are used for the stratosphere and for the troposphere.

the NN bias correction within 4D-Var DA experiments and compare results against weak-constraint 4D-Var which serves as a benchmark.

A description and assessment of the RO temperature retrieval dataset are presented in Section 2. The design and the training of various NN solutions are summarised in Section 3. Section 4 describes results obtained when the NN temperature correction is applied to the model in an assimilation experiment. This NN approach is then compared with the weak-constraint formulation used in operations at ECMWF in Section 5. We finally discuss various aspects of weak-constraint 4D-Var that are also essential for ML such as learning rate and NN retraining, in Section 6. We summarize the paper in Section 7 and provide a perspective for future developments.

2 Temperature retrieval datasets

It is very challenging to produce a ground-truth database for Numerical Weather Prediction (NWP) as all weather measurements and weather simulations contain errors that cannot be ignored. However, some types of observations are more accurate than others and can therefore serve as a reasonable proxy for the true atmospheric state. This is the case for the GNSS Radio Occultation (RO) measurements in the stratosphere, which offer a spatially homogeneous observing system. These measurements consist of high-quality bending-angle profiles that are sensitive to the stratospheric temperature. It has been shown that RO profiles reduce NWP analysis and forecast temperature biases in the lower and middle stratosphere for most NWP centres (Healy & Thépaut, 2006; Poli et al., 2009; Rennie, 2010; Cucurull et al., 2013).

The RO measurement technique is described in Kursinski et al. (1997). The GPS signals propagating between the GPS transmitter and a receiver on a low earth orbiting (LEO) satellite are bent by gradients of the refractive index in the ionosphere and neutral atmosphere, as they pass through the limb of the Earth. The ionospheric

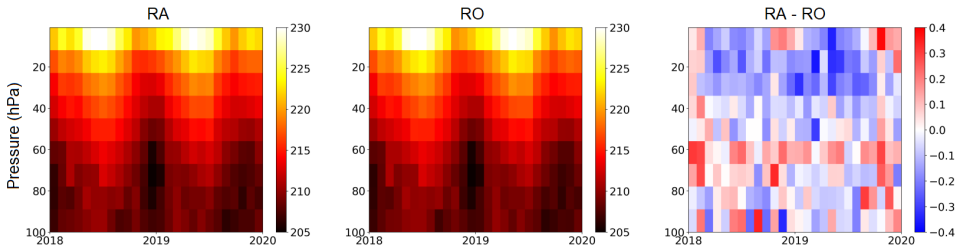

Figure 2. Timeseries of collocated radiosondes (left), RO temperature retrievals (middle) and the difference between the two (right). Observations are collocated on a 5-degree grid every hour within a 1hPa pressure difference between 2018 and 2020

bending can be removed with a simple correction (Vorobev & Krasilnikova, 1994). The ray bending as function of "impact parameter" can be determined, as a result of the motion of the LEO satellite. The impact parameter defines the height of the tangent point of the ray path above the surface. The ray-bending angle values as a function of impact parameter can be inverted to provide information about the atmospheric state, such as temperature. RO measurements are distributed globally, have good vertical resolution, and RO bending angles can be assimilated without bias correction into the NWP model (Healy & Thépaut, 2006). However, in the context of this work, profiles of mean bending angle departures can be difficult to interpret since a given bending angle can have both positive and negative sensitivity to temperature biases in the vertical profile (see section 5.3 Eyre, 1994). We have therefore mapped the bending angle profiles to temperature using a simple implementation of the widely used temperature retrieval algorithm described by Kursinski et al. (1997). Refractive index profiles are derived from bending angles with an Abel transform. There is no measurement information to enable the separation of the effects of temperature and water vapor, and therefore these quantities can be retrieved only using prior information (ERA5 reanalysis in our case). Although this retrieval method provides temperature values up to the top of the atmosphere, the retrieval noise increases with height and most of the information comes from the prior above 3hPa. Therefore, we only use the retrieved temperature values between model level 20 (3hPa) and model level 65 (125hPa) out of a total of 137 vertical levels in the IFS.

It is important to evaluate the quality and accuracy of the temperature retrievals, as they are used as ground truth in our study. RO temperature retrievals can be collocated with conventional temperature observations from radiosondes (RA) to quantify the error characteristics of the observing system (Sun et al., 2010). RO and RA profiles are not available at exactly the same vertical and horizontal location. For comparison, profiles have been collocated on a 5-degree grid, every hour, and within a 1hPa pressure difference. Figure 2 shows a timeseries of the collocated observations from RA (left) from RO (middle) and the difference RA-RO (right). This has been averaged over pressure levels and for every month between 2018 and 2020 to reduce the collocation errors introduced through spatial and temporal mismatch between RA and RO that could influence the accuracy of the obtained statistics. The RA and RO observations present a very similar seasonal signal when the stratosphere is warming up during the Northern hemisphere summer, or cooling down during the Northern hemisphere winter. This pattern arises from the inhomogeneous distribution of radiosondes, mainly sampling the Northern hemisphere. The difference between RA and RO (right panel of Figure 2) shows that the average discrepancies between the two types of observations in the mid/lower stratosphere are smaller than 0.2C and confirms what has been found in other collocation studies (Sun et al., 2010, 2019). In the upper stratosphere (above

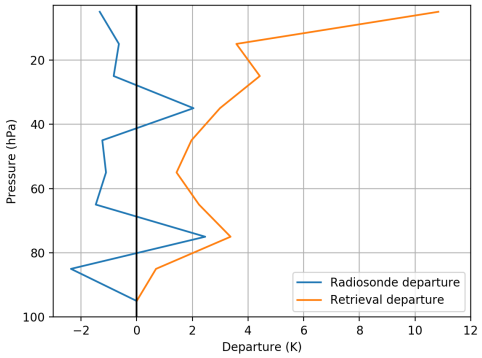

Figure 3. Vertical profile of the ERA5 first-guess departure from a collocated radiosonde (blue) and RO temperature retrieval (orange). Both profiles are measured over the USA (36N, 93W) on 16/10/2020.

30hPa), there is a systematic difference where RO observations are warmer than RA by approximately 0.3C, especially in summer. This shows the intrinsic challenge of finding the ground truth in NWP as each observing system will be sensitive to different sources of error (e.g. solar elevation angle, dry temperature adjustement, ...).

During the collocation study, a small fraction (less than 1%) of profiles showed very large discrepancies. One example is illustrated in Figure 3 for a collocated profile over the USA (36N, 93W) on 16/10/2020. The RA profile agrees roughly with the ERA5 first-guess trajectory, presenting a small first-guess departure. However, the RO profile shows very large differences with respect to the trajectory of ERA5 (over 5 degrees in the upper stratosphere). Future work could include an improved quality control procedure to detect and automatically remove outlier RO profiles with lower quality. The current QC is based on the parameters used in the bending angle assimilation, but the bending angle assimilation is more robust to measurement noise than the RO temperature retrievals used here.

The purpose of the NN is to learn a function representing the model bias that develops in the data assimilation system over the 12-hour assimilation window. A natural choice for the input of the NN is the temperature first-guess trajectory, as it contains the state of the model. The output of the NN is the model bias estimated as the difference between the temperature first-guess trajectory and the RO retrievals. The spatial and temporal structure of the stratospheric temperature bias has been studied in Laloyaux, Bonavita, Dahoui, et al. (2020) and presents large scale patterns that evolve slowly over time. For this reason, the first-guess trajectory and first-guess departure are averaged over a 10-degree regular grid for all the model levels between 130hPa (level 65) and 3hPa (level 20). This means that we have 31,635 inputs and the same number of outputs (19 latitude grid points x 37 longitude grid points x 45 vertical levels). Unfortunately, the observations are not available at every point in space and time. To reduce the number of missing data points, we average the input/output samples over 10 days. The averaging also helps capture slowly varying signals. Linear interpolation is used to fill the remaining observational gaps (representing 5% missing values when using the 10-day average).

Machine Learning requires a large number of samples to properly capture the relationship between input and output variables. To run a dedicated assimilation system with the current IFS model for a long time period is computationally expensive and serial in time, and therefore very slow. It is thus prohibitive to train the networks within the assimilation framework. Therefore, to obtain training data for a long time

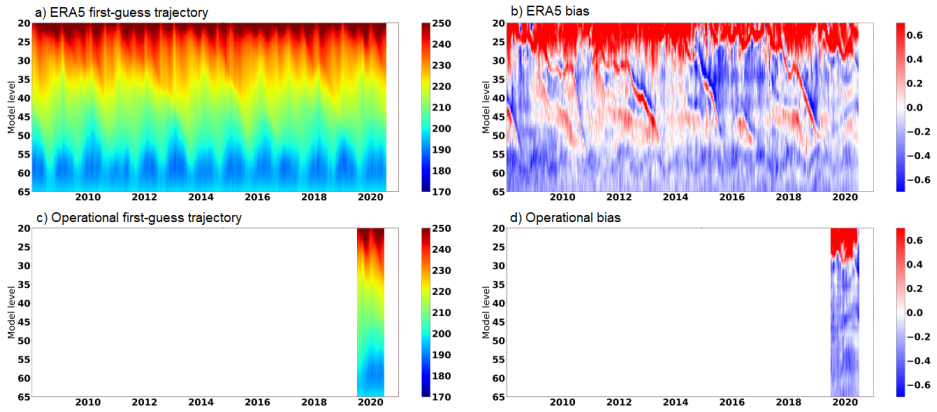

Figure 4. Timeseries of ERA5 temperature first-guess (top left) and departure with RO temperature retrievals (top right) for the different stratospheric model levels (level 20 is 3hPa and level 65 is 130hPa) averaged between 5N and 5S available between 2008 and 2020. The bottom panels show a similar timeseries from the operational dataset that is available only between June 2019 and June 2020.

period, we use data from the ERA5 reanalysis as the first-guess trajectories, and departures have been archived over the entire period for which good RO coverage is available (from 1st of January 2008 until 1st of June 2020). ERA5 is based on an IFS model version (cycle 41r2) implemented in 2015. As we also want to study how a trained bias correction tool can be adjusted to a new model cycle, we also estimate the bias of the model used in operations between June 2019 and June 2020. We will refer to this dataset as the "operational dataset". It consists of one year of first-guess trajectories from cycle 46r1, which improves several aspects of the dynamics and the physics of the model, compared with the ERA5 dataset. The spatial resolution of the two datasets is identical and equal to 18km (the control member of the Ensemble Data Assimilation system is used for the operational dataset, instead of the high-resolution system). Figure 4 shows a timeseries of inputs and outputs produced from the ERA5 (top) and the operational (bottom) dataset, averaged over the Tropics between 5S and 5N. The ERA5 model exhibits a cold bias in the mid/lower stratosphere and a warm bias in the upper stratosphere that propagates down during QBO events. The operational dataset has a similar vertical structure, although the amplitude is much larger.

ML studies generally divide the available data into three different datasets to train, develop, and evaluate an ML model. The training set is the largest and is used to learn the relationship between input and output vairables. The second set, referred to as the validation set, is used exclusively for tuning model hyper-parameters set manually by the model developer (e.g. activation function, learning rate). A key goal of the hyper-parameter tuning process is the optimization of the network's generalization capabilities, in order to avoid overfitting and ensure that the network will function well on previously unseen data. The third dataset is the test set, a collection of previously unseen data, which is used to evaluate the network. The three datasets should be independent of each other, but at the same time they should reflect the same statistical distribution. Several strategies are discussed by Schultz et al. (2021) to achieve this with meteorological time series that are usually auto-correlated. A block sampling strategy is used for our application in order to mitigate this issue. For the ERA5 datasets, we assign the first 10 days of every month and the days after the 20th of each month in 2019 to the validation set and the remaining samples to the

Dataset	training	validation	test
ERA5	2008-2019 (412)	2019 (26)	2020-2021 (42)
operational	2019 (15)	2019 (5)	2020 (18)

Table 2. Partition details for ERA5 and operational data sets. Shown are only the years and total number of samples, in parenthesis. For overlapping years, the data is split into date ranges for each month, as described in the text, in order to create disjoint sets.

training set. The test set is comprised of the entire year 2020 and the first half of 2021. For the operational dataset, we use a similar strategy for splitting it into training and validation sets. Since only data for half of 2019 is available, we assign June 21 to July 1st, July 21st to 31 and August 10 to the validation set, and the rest of the 2019 data to the training set. We assign the full set of 2020 operational data to the test set. Table 2 summarizes the various splits for the two datasets.

3 Design and training of neural networks

3.1 Data representation

The machine learning problem at hand is a multi-dimensional regression problem: the state of the IFS model is used as the input for our network, and the bias computed from the departures between the temperature retrievals and the IFS model is our prediction target. This means, that we are aiming to learn the departure values and not the RO ground-truth data itself, which typically produces a more stable learning process.

The raw data is available as tuples in the form $(\text{longitude}, \text{latitude}, \text{level}, \text{T})$, where T represents the temperature at these coordinates. In this paper, we make use of data regression on structured grids using convolutional neural networks, after converting the data into multi-channel images using suitable projections and interpolations, with the vertical model level mapped to the feature/channel dimension. We examine two possible interpretations of this data: they can be treated as three-dimensional fields, consisting of (projected longitude, latitude and level) with a single feature (temperature)[1], or as two-dimensional fields of (projected longitude, latitude) with a vector of features (temperatures at different altitudes). We will discuss the implication of these two different views below.

In order to stabilize training, we rescale the data using mean-variance normalization. For the 2D case we perform a seperate normalization per altitude/level, whereas for the 3D case we perform a single normalization across all levels. This is important in order to preserve vertical gradients in the data. In each case, we normalize the input and target datasets separately.

3.2 Network architecture

Image regression is similar to image segmentation, and NNs which are designed for segmentation can be adjusted to work for regression simply by dropping the final per-pixel softmax function. Therefore, we can choose a suitable network architecture from a plethora of available image segmentation networks, such as the UNet

[1] Note that model levels can be transformed into physical altitudes by applying an exponential mapping. However, in the 3D approach we treat them as equidistant and rely on the network to learn a reasonable set of filters.

(Ronneberger et al. (2015)) or DeepLab (Chen et al. (2018)) architectures. Both of these architectures employ an encoder, which is responsible for extracting features at multiple length scales. The encoder typically consists of convolutional blocks, with skip connections added to improve training stability. The output of the encoder is passed to a decoder which combines the extracted features to generate a prediction. DeepLab architectures employ an additional step between the encoder and decoder, the so-called atrous spatial pyramid pooling (ASPP) (Chen et al. (2018)) designed to improve feature combination at different scales. The DeepCAM (Kurth et al. (2018)) NN architecture has been successfully applied to the identification of extreme weather phenomena in climate simulations. Therefore, we use a modified variant of the original architecture in which the ResNet-50 (He et al. (2016)) backbone is replaced by an XCeption (Chollet (2017)) backbone. Also, instead of relying on interpolated upsampling, we employ a fully convolutional decoder. These two improvements lead to the network architecture which forms the basis of the MLPerf HPC DeepCAM benchmark (see *MLPerf HPC DeepCAM website* (2021)). In order to reduce checkerboard artifacts produced by convolutional upsampling, we furthermore insert average pooling layers after the convolutions with a pooling kernel size equal to the convolutional upsampling stride. It has been shown by Kinoshita & Kiya (2020) that this is an effective technique for reducing such artifacts in the generated images.

For the 2D representations of the data, we simply adjust the number of input channels in the previously described architecture. For the 3D representation, we convert all 2D operations (convolutions, batch-normalizations, pooling) into their respective 3D counterparts. A significant difference between these two approaches is that in the 2D case, all altitude levels are combined in an all-to-all fashion through the matrix multiplication along the feature dimension in the 2D convolutional kernel. In contrast, the 3D convolutions only correlate neighbouring levels. Therefore, they have are better suited to capturing temperature gradients between levels, whereas 2D convolutions might be better at capturing long distance correlations spanning multiple levels. Both architectures are reasonable choices for solving the bias prediction problem at hand, and thus we pursued both approaches.

3.3 Training process, R2-score and hyper parameter optimization

We employ the AdamW optimizer (Loshchilov & Hutter (2019)) and apply weight decay regularization in order to reduce overfitting, which is particularly important when training on the smaller, operational dataset. For the loss function, we use either the L2 distance or a smooth version of the L1 distance between network output and prediction target.

The R2 score is used as a validation metric for hyper parameter tuning. It is defined as

$$R2 = 1 - \frac{\sum_{i=1}^{m}(y^{(i)} - f^{(i)})^2}{\sum_{i=1}^{m}(y^{(i)} - \bar{y})^2}, \quad \text{where } \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)} \tag{1}$$

Here, $y^{(i)}$ is the NN prediction for sample i and $f^{(i)}$ is the corresponding ground truth, i.e. in our case the model bias. The R2 score compares the prediction accuracy with the intrinsic variance of the data: if prediction accuracy is high, then the numerator in equation (1) is small, which leads to $R2 \approx 1$. If the prediction accuracy does not outperform the intrinsic noise, then the numerator and denominator in equation (1) will be of similar magnitude and we find that $R2 \approx 0$. For predictions of even lower accuracy we have $R2 < 0$ which signals a failure of the model. In order to obtain a scalar score, we perform a summation over all the pixels and levels in the output image. However, a more detailed qualitative analysis is possible by computing the R2 score per level and/or per longitude/latitude coordinate.

We tune hyper parameters (HPO) using the ray.tune package (*Ray Tune website* (2021)) with HyperOpt (*Hyperopt website* (2021)), running 128 instances for both the 2d and 3d models. Tuneable hyper parameters in our model include learning rate, weight decay, learning rate schedules (selection of multi-step with different milestones, cosine annealing with different choices for decay frequency), loss definition (smooth L1 vs. L2) and batch size. Our hyperparameter optimization target is the maximisation of the R2 value, as described above. Each model is trained for about 150 epochs.

3.4 Computational Performance

We use a single NVIDIA DGX-2 system for training and run a single instance on each GPU concurrently. This means, we can train 16 instances in parallel. Training a single instance on an NVIDIA V100 GPU takes takes about 30 minutes for the 3D model. Therefore, training 128 instances does not take longer than 4 hours in total. It is unlikely that training more instances would lead to the discovery of better hyperparameters, because many good hyperparameter choices[2] perform equally well and it is hard to define a quantitative criterion which configuration to prefer over the others.

3.5 Training, using a small operational dataset

In order to produce the most accurate weather forecast possible, we would like to construct a NN bias-correction model based on data from the latest IFS model cycle. While there is plenty of ERA5 data available (based on the 2015 cycle), the dataset for the current cycle is much smaller. In our case, we had only 15 training samples and 5 validation samples available. We examined several approaches in an attempt to build the most useful tool for this scenario.

1. Finetuning: in this approach, the model does not begin using random initial weights. Rather, it begins with weights that have been pre-trained on a related dataset. Specfically, we pre-trained the model using the ERA5 dataset and then fine-tuned the entire model using only the operational dataset, but with a much smaller learning rate. Using this approach, we found that the NN quickly overfit to the operational data. Therefore, we did not pursue this approach further.
2. Training from scratch: In this case we trained the model using only data from the latest IFS cycle. While it appeared more promising than finetuning for the first few epochs, this approach broke down rapidly as well, heavily overfitting the training dataset for all hyper parameter configurations tried. Hence we abandoned this approach as well.
3. No retraining: in this, simplest approach. we used only the existing model, trained exclusively on the ERA5 dataset, with no fine-tuning. This model was then applied directly to the shorter, operational dataset. This approach is promising if the underlying intrinsic features of both datasets are similar. It turns out that this approach yields reasonable results, producing R2 values *only* about $\sim 20\%$ lower than the original ERA5 test dataset.
4. Training on both datasets simultaneously: for this transfer learning strategy, we implemented a data loader which can feed the NNs samples from either dataset. The two datasets have a relative sample imbalance of about 27:1 (ERA5:operational). In order to help the NN learn the features of the operational dataset, the dataloader selects samples from the both datasets, but with inverted frequencies. This means that the NN is presented samples from both datasets with almost equal probability. In practice, we chose a final ratio of

[2] *good* means that they deliver a high R2 score on the validation set

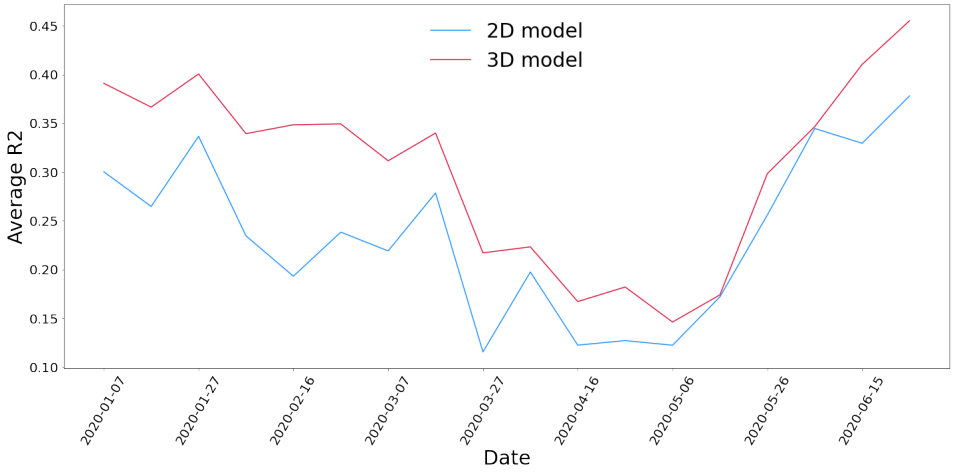

Figure 5. Timeseries of R2 values averaged over all levels for 2D + features (blue) and 3D NN (red) architectures on the 41r2 test dataset.

27:29 (ERA5 : opertional data) in order to provide a small emphasis on the importance of the operational data. (This is a tunable hyperparameter). For the validation dataset, we use only samples from the operational set, as we are interested only in the operational model performance. We also use only operational data to compute the R2 score when performing hyper-parameter optimisation. It turns out that training the network with this approach is very stable.

4 Results for bias correction with deep learning

In this section, we present results for temperature bias correction based on deep learning. The first and second subsections present results for offline bias correction, for the ERA5 and operational datasets respectively. The third subsection discusses the use of bias correction within data assimilation experiments.

4.1 Performance comparison of 2D and 3D models

We trained the 2D and 3D models with their respective best known hyperparameters on the ERA5 training dataset and compared their performance on the ERA5 test dataset. Figure 5 displays the R2 value, averaged over all levels, for the test set . The plot demonstrates that the 3D model outperforms the 2D model consistently. This is likely due to the importance of gradients and other local co-variances in the vertical direction, and the inherent advantage convolutions provide for learning such relationships in a data-efficient fashion. Therefore, we decided to conduct subsequent studies exclusively using 3D network architectures. The variability of the first-guess trajectory and of the model bias is larger between March and May, as the Northern hemisphere warms up and the Southern hemisphere cools down. There is a drop in the R2 value for both models as they struggle to accurately capture the model bias over that period.

Figure 6 shows the target bias (left) and the prediction of the 3D model (right) for two different vertical levels on February 6 2020, from the ERA5 test set. The NN clearly learns to reproduce important features, such as the negative bias correction around the equator for level 40. It also learns to reproduce the region of stronger

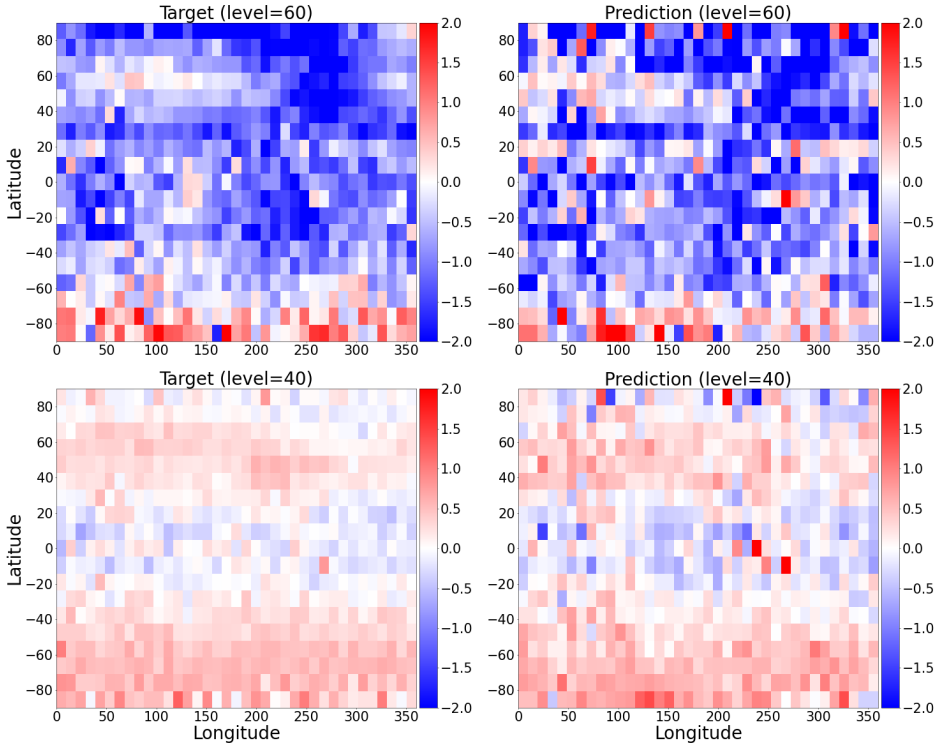

Figure 6. Target bias (left) and bias prediction from the NN model (right) for two levels on February 6 2020.

negative bias near (60° lat, 275° lon) as well as in the latitude band between 20° and 40°.

4.2 Performance of 3D models on ERA5 and operational datasets

In this section, we compare three test cases, each using the 3d convolutional architecture: (i) the original ERA5 model evaluated on the ERA5 test set, (ii) the original ERA5 model evaluated on the operational test set, and (iii) the model retrained on both ERA5 and operational training data, and evaluated on the operational test set. Note that test cases (ii) and (iii) correspond to training scenarios 3 and 4 from section 3.5.

Figure 7 shows a vertical profile of the globally-averaged R2 scores for each of the three test cases. We observe a steep drop in prediction quality when testing on the operational dataset. The retrained model produces a better prediction than the original ERA5-only model on the operational dataset, except for the top-most levels. Comparing the model biases from the ERA5 and operational datasets (panel b and d in Figure4), we see that the retrained model struggles to capture the larger warm bias in the top levels of the operational dataset.

Figure 8 shows the target and predicted biases for the original ERA5 model on the ERA5 targets and the retrained model on the operation targets. Before April 2020, both the original and retrained predictions correctly capture the patterns observed in the ERA5 targets, although the predictions have a somewhat larger amplitude than the target values. After April 2020, we see that the operational target values differ significantly, where the Northern hemisphere is nearly bias free and the Southern hemi-

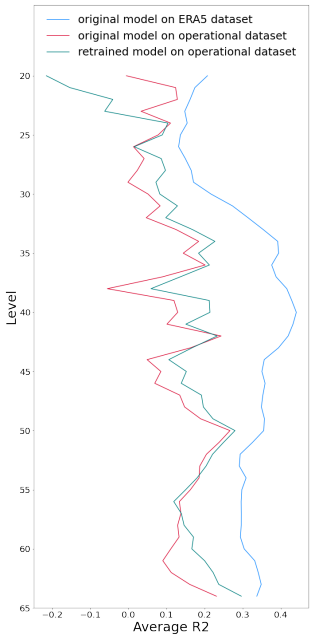

Figure 7. Vertical profile of globally averaged R2 values for the original model on the ERA5 target (blue), for the same model on the operational target (red) and for a retrained model using the sample balancing technique described above, on the operational target (green).

sphere exhibits a cold bias that was not present for the ERA5 target set. This feature is not capture by the NN prediction and may explain the much of the performance drop for these models in this time window.

4.3 NN bias correction in 4D-Var data assimilation

The bias predicted by our NNs can, in principle, be used to correct the model tendencies of the IFS within data assimilation experiments, in order to produce a better analysis. Unfortunately, it is technically challengng to introduce the bias correction tools into the workflow of the 4D-Var data assimilation experiments. Not only is it difficult to couple the machine learning tools with the IFS workflow, using our NN bias correction models to correct the IFS tendencies also requires one to re-gridding the model fields from the reduced-gaussian model grid of the IFS to the regular gaussian grid at the coarse resolution used to predict the bias. It is therefore beyond the scope of this paper to perform "online" simulations that calculate and correct the bias within 4D-Var experiments.

However, we are able to predict the model bias "offline" using the retrained NN on the first-guess trajectories contained in the operational test dataset. We can run a 4D-Var experiment where the model is corrected with the respective offline correction valid for the same date. The correction is applied as an integrated term between each model timestep. Using this framework, the machine learning approach is evaluated in 4D-Var over the test period between 1st January 2020 and 1st March 2020. Figure 9 shows the first-guess mean error with respect to RO temperature retrievals for different 4D-Var experiments. The red line is the control experiment, where the dynamical model is not corrected. The dotted blue line shows the first-guess mean error, where the dynamical model is corrected using the actual target from the RO datasets. This provides a estimate of how much the bias could actually be reduced if

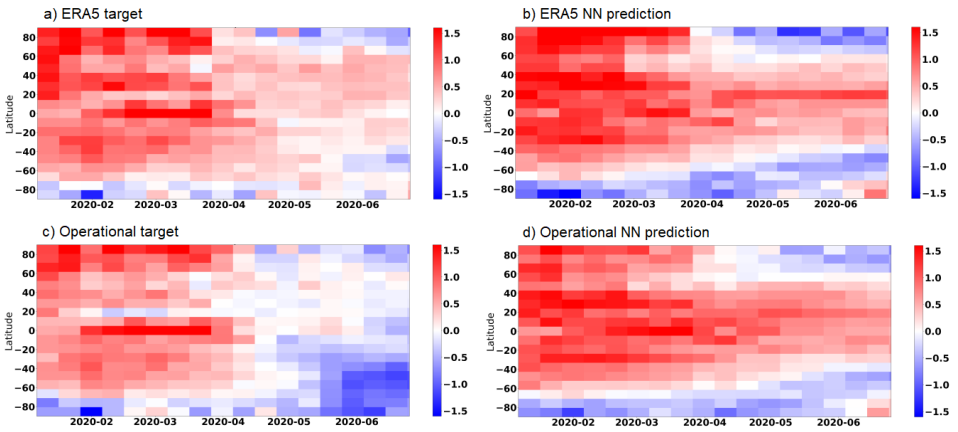

Figure 8. Zonal timeseries of the ERA5 targets (a) and the bias predictions from the original NN for model level 25 (6hPa). The bottom panels show similar timeseries for the operational targets (c) and the bias predictions from the retrained NN (d)

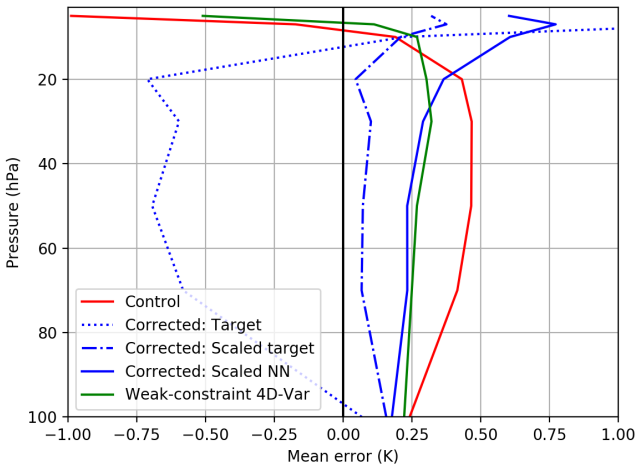

Figure 9. First-guess mean error with respect to RO temperature retrievals for the control (red), for weak-constraint 4D-Var (green), for the model corrected with the target (dotted blue), with the scaled target (dash-dot blue) and with the scaled prediction of the NN (solid blue). Statistics are averaged over the globe between 01/01/2020 and 01/03/2020.

the NNs provided a perfect fit to the training data. One can see that the model bias is over-corrected, for example, the original 0.4C cold bias at 60hPa became a -0.6C warm bias. This over-correction is due to the intrinsic cycling principle of data assimilation where the analysis valid at the beginning of the assimilation window is integrated forward in time to produce the background at the beginning of the next assimilation window. The first-guess departures used to diagnose the model bias contain not only the bias that develops over a single assimilation window but also includes the bias accumulated over the previous assimilation cycles contained in the background. A study of the background and analysis departures shows that only one quarter of the total bias comes from the current assimilation cycle while the other three quarters are carried forward in time from the previous cycle (e.g. at 50hPa, the global-mean analysis departure is equal to 0.38 and the global-mean background departure is equal to 0.5). Following these findings, another 4D-Var experiment was run where the model is corrected by a scaled target where all the values are reduced by a factor of 4. This approach, plotted in dash-dot blue, is able to efficiently correct the model bias for the entire stratosphere. The last experiment plotted in solid blue shows the results when the model is corrected by the offline NN predictions with the same 1/4 scaling. The NN is able to capture and correct a large fraction of the actual model bias. The first-guess mean error is reduced by almost 0.2C in the mid/lower stratosphere. The poor performance around 5hPa where the model is over-corrected is likely due to the small size of the operational dataset. The ERA5 warm bias at 5hPa is well captured by the initial NN (comparing left and right top plots in Figure 8). However, the operational model presents a smaller bias that is not well represented in the NN, which retains too much of the structure learned from ERA5. This means that the NN will cool the top of the atmosphere too aggressively, over-correcting the model warm bias.

5 Comparison with weak-constraint 4D-Var

Weak-constraint 4D-Var has been introduced by several authors to denote a family of algorithms which relax the perfect model assumption (Wergen, 1992; Zupanski, 1993; Bennett et al., 1996; Vidard et al., 2004; Dee, 2005; Trémolet, 2006). In the forcing formulation of weak-constraint 4D-Var (Trémolet, 2006) a forcing is estimated and then applied in the model's equations to represent the error which gradually enters into the model trajectory. The model is then treated in the same manner as other sources of information, taking into account that there is a degree of uncertainty about the information it can provide on the evolution of the atmospheric state over the analysis cycle. Mathematically, the weak-constraint 4D-Var formulation that has been implemented at ECMWF introduces a forcing η to represent the error which gradually enters into the model trajectory

$$\mathbf{x}_k = \mathcal{M}_{k,k-1}(\mathbf{x}_{k-1}) + \eta \qquad \text{for} \qquad k = 1, ..., N. \tag{2}$$

The model error forcing is assumed to be additive and constant within the 12-hour assimilation window (Laloyaux, Bonavita, Dahoui, et al., 2020; Laloyaux, Bonavita, Chrust, & Gürol, 2020). It contains temperature, vorticity and divergence. We also assume that the model error η follows a Gaussian distribution with no cross-correlation with the background error. This is justified if we assume that the model error that we want to estimate and the background errors act on different spatial and temporal scales. This set of assumptions allows one to write the weak-constraint 4D-Var cost function as

$$J_{WC}(\mathbf{x}_0, \eta) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^{\mathrm{T}} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \frac{1}{2}\left(\eta - \eta^b\right)^{\mathrm{T}} \mathbf{Q}^{-1}\left(\eta - \eta^b\right) + \frac{1}{2}\sum_{k=0}^{N}\left(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k\right)^{\mathrm{T}} \mathbf{R}_k^{-1}\left(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k\right) \tag{3}$$

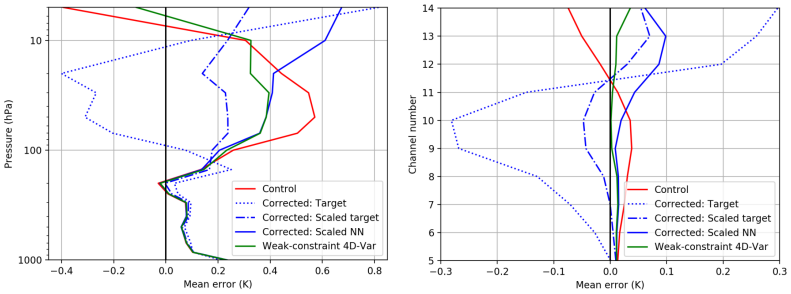

Figure 10. Same as Figure 9 but respect to radiosondes (left) and AMSU-A (right).

where η^b is the prior estimate of the model forcing estimated from the previous assimilation cycle. This forcing formulation of weak-constraint 4D-Var simultaneously estimates the initial state $\mathbf{x}_0$ and model forcing η that best fit the observations and the background information with respect to their error covariance matrices.

A weak-constraint 4D-Var experiment was run from the 1st of January 2020 and is presented in green in Figure 9. In this experiment, the model error estimate is set to zero initially as no a priori knowledge of the model error is assumed. After processing only one month of data, weak-constraint 4D-Var is able to correct for a significant fraction of the model bias without requiring the computation of a large dataset for offline training.

Weak-constraint 4D-Var can be seen as a specific machine learning algorithms that learns the model error by estimating the parameters in the forcing vector (Farchi, Bocquet, et al., 2021). However, there are several conceptual differences with the machine learning approach described in Section 4. Weak-constraint 4D-Var is an online learning algorithm which simultaneously estimates the model state and the model error while the NN approach is estimating the model error offline before estimating the model state. An online NN is feasible but it would require a stronger interaction between NN tools and the IFS model to exchange data at each assimilation cycle. This work would require a substantial effort, given the current software infrastructure. Another difference is the amount of information used to estimate the model bias. Weak-constraint 4D-Var uses the information from all observations (conventional and satellites) as all of these are actively assimilated thanks to the radiative transfer scheme included in the 4D-Var cost function. The NN has learned the model bias using only the RO temperature retrievals which represents a small subset of the whole observing system. It is therefore interesting to study how the two approaches will fit other conventional and satellite instruments. Figure 10 shows the first-guess mean error with respect to radiosondes (left) and AMSU-A (right). In the weak-constraint 4D-Var experiment, these observations have been actively used in the observation term of the cost function (see Equation 3). This means that the data assimilation algorithm finds the optimal state that fits all of the observations with respect to their uncertainties. In the NN approach, only RO retrievals have been used to estimate the model bias as radiosondes and AMSU-A observations have not been introduced during the training. The scaled NN shows a similar improvement than weak-constraint 4D-Var in the lower and mid stratosphere (i.e. radiosondes below 30hPa and AMSU-A channel numbers below 11). This is an excellent news that can possibly be explained by the fact that RO, radiosonde and AMSU-A observing systems are consistent between each others, highlighting a similar model bias. We chose to illustrate this point using AMSU-A observations, but a similar conclusion can be drawn for other microwave instruments (e.g. ATMS) or infrared instruments (e.g. AIRS or Cris). The performance of the NN approach is not as good in the upper stratosphere (i.e. radiosondes above 30hPa and

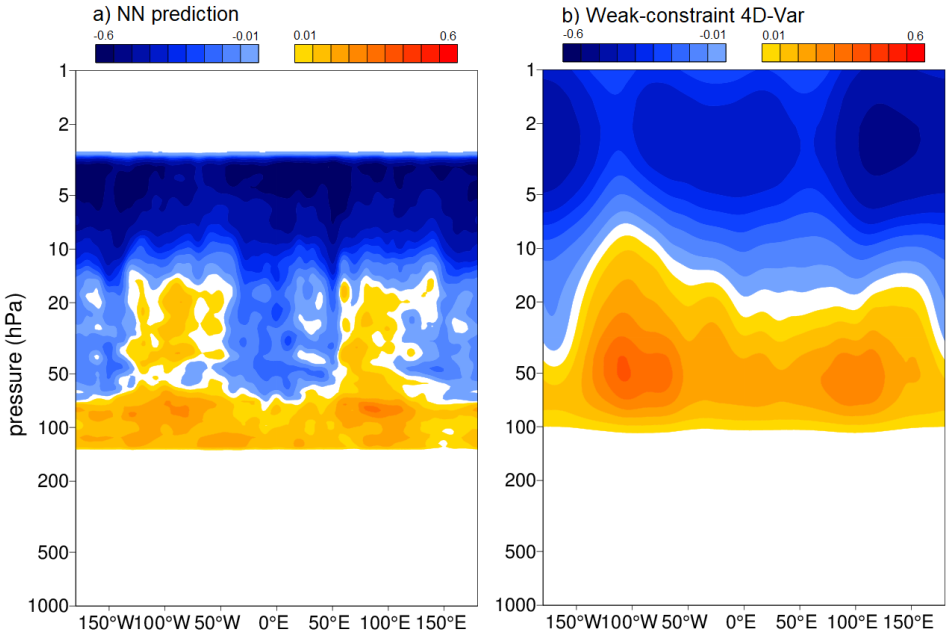

Figure 11. Meridional cross-section temperature error correction from the from the NN prediction (a) and from weak-constraint 4D-Var (b) averaged over the tropics (10N-10S) between 1st January 2020 and 1st March 2020.

AMSU-A channel numbers above 11) which confirms what has been noticed in Figure 8 against RO retrievals. We have run 10-day forecasts initialised with weak-constraint 4D-Var and NN analyses to study the impact on medium-range weather forecasting. We found that signals in the analysis are retained throughout the forecast and are still present after five days which confirms the results presented in Laloyaux, Bonavita, Dahoui, et al. (2020). In the lower and mid stratosphere, weak-constraint 4D-Var and NN forecasts show similar improvements at day five. The only difference happens in the upper stratosphere where the NN forecasts are degraded due to the poorer quality of the NN analysis above 20hPa (see Figure 9 and Figure 10).

Developing methods that estimate model biases should eventually help modellers improve their models by providing more complete knowledge of the bias structure. This will fulfil the synergies between better observations, sophisticated DA algorithms and improved physical models. It is therefore informative to study the model biases highlighted by both approaches. Figure 11 shows a meridional cross-section temperature error correction from the NN prediction (left) and from weak-constraint 4D-Var (right) averaged over the tropics (10N-10S) between 1st January 2020 and 1st March 2020. Both approaches warm up the atmosphere over areas of strong convection (e.g. Indonesia and Southern America). The weak-constraint 4D-Var model error estimate is smoother, due to the specification of the model error covariance matrix Q which retains only large-scale patterns. This could be linked to an insufficient representation of the effects of sub-gridscale gravity wave activity, which leads to missing momentum from the troposphere to the stratosphere (Polichtchouk et al., 2019). The NN prediction is also larger for the top of the stratosphere compared the the weak-constraint 4D-Var correction. This larger NN correction is the reason for the degradation observed in Figure 9 and 10 for the top of the stratosphere.

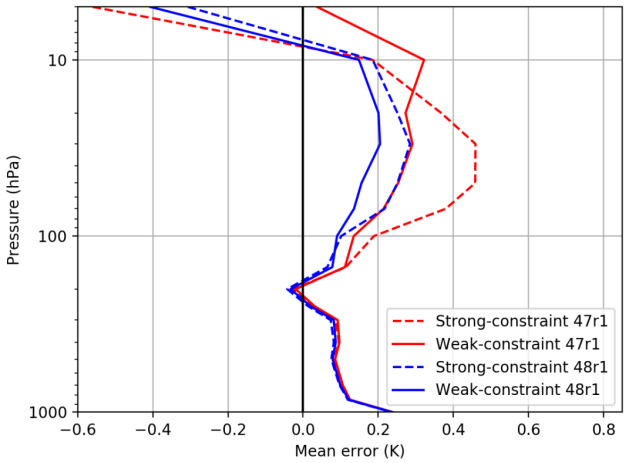

Figure 12. Vertical profile of first-guess departure with respect to radiosondes for 47r1 strong-constraint (dashed red), for 48r1 strong-constraint (dashed blue), for 47r1 weak-constraint (solid red) and for 48r1 strong-constraint (solid blue). Statistics are averaged over the globe between 20/01/2020 and 20/02/2020.

6 Weak-constraint 4D-Var learning rate

We already discussed the question of retraining the NN in Section 4 as new IFS models are made available on a regular basis with improved dynamical and physical processes of the atmosphere. We have shown that this is a challenge for the NN as the training dataset with the new model is usually relatively small (less than a year) as it is expensive to run the assimilation system for a longer period. We illustrate here how weak-constraint 4D-Var handles model upgrades using, as an example, the package of changes that is currently being tested as a possibility for the implementation of the next cycle (tentative 48r1). It contains the hybrid linear ozone, the semi-lagrangian vertical filter and a new solar spectrum. The impact of these model changes is assessed in the strong-constraint 4D-Var formulation where no model bias correction is computed. This allows one to accurately quantify how much the model upgrade reduces the model bias. Figure 12 shows the vertical profile of first-guess departure with respect to radiosondes for strong-constraint experiments with 47r1 (in dashed red) and tentative 48r1 model (in dashed blue). The improvements proposed for 48r1 significantly reduce the stratospheric model biases. At 50hPa, the original bias of 0.45 is brought down to 0.2. Weak-constraint 4D-Var aims to correct the residual model bias. The dashed red and blue lines in Figure 12 show the results of weak-constraint 4D-Var with the 47r1 and 48r1 model respectively. Although the structure of the bias is different for the two models, weak-constraint 4D-Var reduces the first-guess mean error in both situations. The weak-constraint 4D-Var cost function depends on a number of parameters that are estimated offline (e.g. standard deviation and correlation in Q). It is important to note that these parameters have not been retuned in the experiments. This shows the robustness of weak-constraint 4D-Var and its fast learning rate.

The initialisation of the model error correction at the beginning of an experiment can be compared to the challenge of initialising the weights of a NN. The middle panel in Figure 13 shows a timeseries of the model bias correction with the tentative 48r1 model when weak-constraint 4D-Var has been cold started (i.e. setting the model error correction to zero at the beginning of the experiment). It takes a couple of weeks for the model errors estimate to be properly spun-up. This is mainly because weak-constraint 4D-Var aims to correct model biases that are evolving slowly over time. To

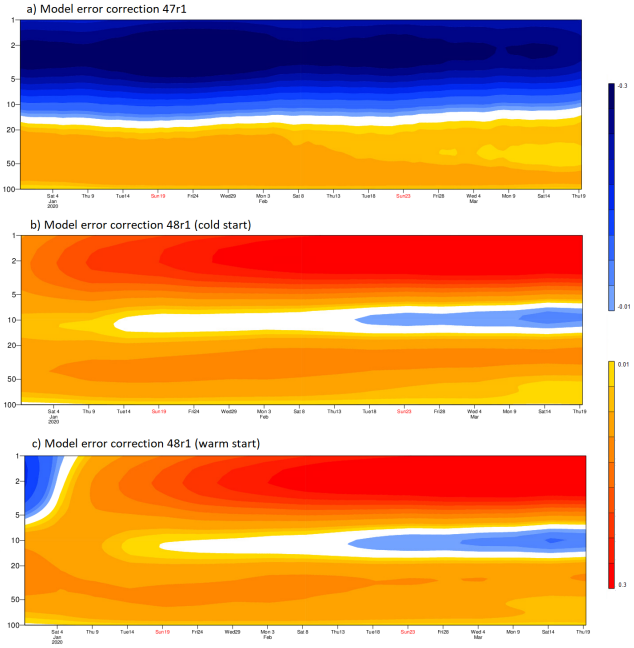

Figure 13. Timeseries of model error corrections estimated by weak-constraint 4D-Var for the 47r1 model (top), for the 48r1 model initialised from zero (middle) and for the 48r1 model initialised from the 47r1 bias estimate (bottom). Statistics are averaged between 70S and 30S.

study the sensitivity of the initialisation, a weak-constraint 4D-Var experiment was run where the model error correction is initialised from the previous 47r1 model error estimate. This timeseries is presented at the bottom panel in Figure 13 and shows that weak-constraint 4D-Var converges towards the same solution although the behaviour is different during the spin-up period. This is a reassuring result demonstrating that weak-constraint 4D-Var is not very sensitive to the way it has been initialised. This can be explained by having a number of observations assimilated in weak-constraint 4D-Var and the model error covariance Q, which are sufficient to constrain the model error correction.

Finally, it is important to understand how efficiently the model bias can be estimated during extreme events. The stratospheric sudden warming (SSW) is the most dramatic meteorological phenomenon to take place in the stratosphere, usually occuring over the north pole. As the temperature drops during winter, low-pressure (cyclonic) circulation begins to develop across the polar stratosphere. A strong polar vortex usually means strong polar circulation even at the lower levels. It can lock the cold air into the Polar regions, resulting in milder winters for most of the United States and Europe. If this vortex is disturbed, the winds can reverse and the temperature can rapidly increase by up to 50 degrees Celsius over a few days, in the vertical region between 1hPa and 10hPa. This can create a chain reaction, which can disrupt the jet stream, creating a high-pressure area over the Arctic circle. This, in turn, can release the cold arctic air into Europe and the United States (Polichtchouk et al., 2018; Mariotti et al., 2020). SSWs happen every-other year or so, with the most recently event recorded in January 2021. The top panel of Figure 14 shows a timeseries of first-guess departure with respect to RO temperature retrievals, averaged over the Northern pole (70N 90N) between September 24, 2020 and February 24, 2021. At the beginning of the SSW event (1st of January 2021), the structure of the model bias

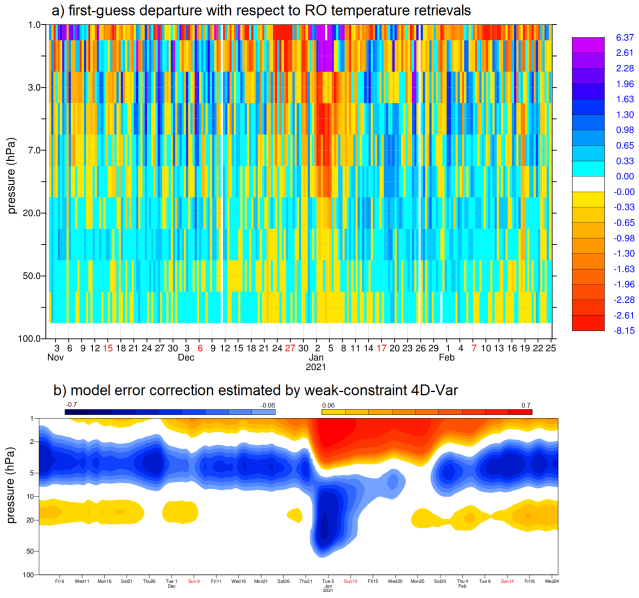

Figure 14. Timeseries of first-guess departure with respect to RO temperature retrievals (top) and timeseries of model error correction estimated by weak-constraint 4D-Var (bottom). Statistics are averaged over the Northern pole (70N 90N) between 24th September 2020 and 24th February 2021.

changes significantly as stratospheric dynamics are disrupted. There is a model cold bias above 3 hPa and a model warm bias between 50 hPa and 3 hPa. The bottom panel of Figure 14 shows the model error correction estimated by weak-constraint 4D-Var. The model bias change is captured quickly as weak-constraint 4D-Var warms up the stratosphere above 3hPa and cools down between 50hPa and 3hPa. This illustrates the efficient learning rate of weak-constraint 4D-Var when an extreme event occurs. A similar study could not be done for the NN approach as the test dataset (June 2019 to June 2020) does not contain such an event. This is however a critical aspect that will be studied in the future, as extreme events occur infrequently in the training dataset and it might be challenging for the NN to correctly represent the model error structure.

7 Summary and perspectives

Artificial intelligence and machine learning are entering the domain of Earth system predictions in parallel with the development of more heterogeneous High-Performance Computing (HPC) architectures. This changing context presents new development opportunities that ECMWF is considering, with the ambition of retaining leadership in global medium- and extended range weather forecasting. 4D-Var data assimilation and machine learning share a common theoretical foundation and use similar computational tools. This has driven the work presented in this paper, which compares how each method is able to estimate and correct systematic errors in the IFS atmospheric model developed at ECMWF model.

The results of this paper show that convolutional NNs are adequate to learn to estimate three-dimensional model bias from RO temperature retrievals. While large datasets containing several years of data are required for the training to achieve optimal results, transfer learning can help to mitigate data limitations if only a small quantity

of training data is available. Still, when used to perform bias correction in data assimilation experiments for a recent IFS model cycle and with a single year of training data for re-training, the deep learning tools of this paper were not able to outperform the current weak-constraint 4D-Var formulation that is in operational use at ECMWF.

However, direct comparison between the two methods has one main limitation. Weak-constraint 4D-Var can be seen as an "online" machine learning method, where observations over the last 12 hours are used to update the previous weather forecasts. The machine learning tool of this paper is based on an "offline" training. Furthermore, the deep learning bias correction was computed "offline" before the assimilation experiment was started. It is difficult to estimate how much results would change if an update of the bias correction was calculated during the assimilation experiment which is – for technical reasons – beyond the scope of this paper. Another difference between the two approaches lies in the physical variables that are corrected. Weak-constraint 4D-Var estimates a forcing field for temperature, vorticity and divergence. Although very few stratospheric wind observations are available, these variables are linked through the model's equation in the 4D-Var cost function. This means that wind corrections are made in conjunction with temperature adjustements. The NN approach corrected only temperature biases. The weak-constraint 4D-Var also includes a model error covariance matrix Q that represents separately the statistics of the model error for temperature, vorticity and divergence. Cross-correlation between variables are not taken into account at the moment. Diagnostics in the IFS model show that the stratospheric temperature model biases evolve on larger spatial scales and longer timescales than background errors (Laloyaux, Bonavita, Chrust, & Gürol, 2020). This information is contained in the Q matrix and helps weak-constraint 4D-Var to correctly attribute the different sources of errors. A similar approach could be investigated in the NN approach introducing a similar regularization term in the loss function. Finally, the jump from the model cycle used in ERA5 and in operations as performed in this paper represents a significant change in the temperature bias as it represents a transition over several years of model development.

The deep learning approach has room for improvement, for example by extending the dataset to encompass more observation types. However, this is challenging as most observations do not measure model prognostic variables on a given grid point but a radiance that is sensitive to a broad vertical level. The development of machine learned observation operators to project observations onto model fields would be mandatory. The use of deep learning methods could also be extended further to include estimates of background and observation error covariance matrices, and to represent uncertainties explicitly, for example via Generative Adversarial Networks (Leinonen et al. (2021)). The treatment for sparsity observations could also be improved further, for example via the use of graph-NNs, which could evaluate observations at the points in space and time when they are available, and even respect spherical symmetry of the globe (cf. e.g. Defferrard et al. (2020)). Graph-NNs would also allow for the use of unstructured grids potentially including the native grid of the IFS and could better exploit the sparsity of the data by replacing the interpolation step with a NN based extrapolation. An online NN could be implemented in the future to study the full potential of a ML solution in the 4D-Var framework. However, this is work in progress and will require further developments regarding software infrastructure and more research to find the best way to update NN weights in a 4D-Var cycling environment. One of the key aspects of ECMWF business is the Research-to-Operations (R2O) process, which is followed to upgrade the software used in forecast production (Buizza et al., 2017). R2O includes a series of actions that could be summarized in 6 activities: planning, development, testing, evaluation, communication and implementation. The IFS model is upgraded at every cycle to better represent physical processes or introduce new ones that were missing. This paper illustrated the strength of weak-constraint 4D-Var that is able to estimate the bias of a new model with no need to construct a new training dataset

or to retune parameters. Specific solutions are required to achieve a similar flexibility with a NN.

8 Acknowledgements

PD gratefully acknowledge funding from the Royal Society for his University Research Fellowship as well as the ESiWACE project funded under Horizon 2020 No. 823988 and the MAELSTROM EuroHPC-JU project (JU) funded under No 955513. The JU receives support from the European Union's Horizon research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, and Norway.

9 Open research

Data availability statement: The input and output data of the experiments described in the paper is freely available for research purposes from ECMWF and can be requested following the procedures described in https://www.ecmwf.int/en/forecasts/datasets

References

Bannister, R. N. (2008a). A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, *134*(637), 1951-1970.

Bannister, R. N. (2008b). A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. *Quarterly Journal of the Royal Meteorological Society*, *134*(637), 1971-1996.

Bennett, F., Chua, A., & Leslie, B. (1996, 03). Generalized inversion of a global numerical weather prediction model. *Meteorology and Atmospheric Physics*, *60*, 165-178. doi: 10.1007/BF01029793

Bonavita, M., Hólm, E., Isaksen, L., & Fisher, M. (2016). The evolution of the ECMWF hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 287–303.

Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. *Journal of Advances in Modeling Earth Systems*, *12*(12).

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2020). Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the lorenz 96 model. *Journal of Computational Science*, *44*.

Buizza, R., Andersson, E., Forbes, R., & Sleigh, M. (2017). *The ECMWF research to operations (R2O) process* (Technical Memorandum No. 806). Reading, UK: ECMWF.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834-848.

Chollet, F. (2017). *Xception: Deep learning with depthwise separable convolutions.*

Cucurull, L., Derber, J. C., & Purser, R. J. (2013). A bending angle forward operator for global positioning system radio occultation measurements. *Journal of Geophysical Research*, *118*.

Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3323–3343.

Dee, D. P., & Uppala, S. (2009). Variational bias correction of satellite radiance data

in the ERA-Interim reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *135*(644), 1830–1841.

Defferrard, M., Milani, M., Gusset, F., & Perraudin, N. (2020). *Deepsphere: a graph-based spherical cnn.*

Dueben, P., Modigliani, U., Geer, A., Siemen, S., Pappenberger, F., Bauer, P., ... Baousis, V. (2021). *Machine learning at ECMWF: A roadmap for the next 10 years* (Technical Memorandum No. 878). Reading, UK: ECMWF.

Eyre, J. R. (1994). *Assimilation of radio occultation measurements into a numerical weather prediction system* (Technical Memorandum No. 199). Reading, UK: ECMWF.

Farchi, A., Bocquet, M., Laloyaux, P., Bonavita, M., & Malartic, Q. (2021). *A comparison of combined data assimilation and machine learning methods for offline and online model error correction.*

Farchi, A., Laloyaux, P., Bonavita, M., & Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, *147*(739), 3067-3084.

Fomichev, V. I., Ward, W. E., Beagley, S. R., McLandress, C., McConnell, J. C., McFarlane, N. A., & Shepherd, T. G. (2002). Extended canadian middle atmosphere model: Zonal-mean climatology and physical parameterizations. *Journal of Geophysical Research: Atmospheres*, *107*(D10).

Geer, A. (2020). *Learning earth system models from observations: machine learning or data assimilation?* (Technical Memorandum No. 863). Reading, UK: ECMWF.

Geer, A. J., Lonitz, K., Weston, P., Kazumori, M., Okamoto, K., Zhu, Y., ... Schraff, C. (2018). All-sky satellite data assimilation at operational weather forecasting centres. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1191-1217.

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

Groenquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194).

Haiden, T., Dahoui, M., Ingleby, B., de Rosnay, P., Prates, C., Kuscu, E., ... Jones, L. (2018). *Use of in situ surface observations at ecmwf* (Technical Memorandum No. 834). Reading, UK: ECMWF.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 770-778). doi: 10.1109/CVPR.2016.90

Healy, S. B., & Thépaut, J.-N. (2006). Assimilation experiments with CHAMP GPS radio occultation measurements. *Quarterly Journal of the Royal Meteorological Society*, *132*, 605–623.

Hogan, R., Ahlgrimm, M., Balsamo, G., Beljaars, A., Berrisford, P., Bozzo, A., ... Wedi, N. (2017). *Radiation in numerical weather prediction* (Technical Memorandum No. 816). Reading, UK: ECMWF.

Hyperopt website. (2021, 11). Retrieved from `http://hyperopt.github.io/hyperopt/`

Janjic, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., ... Weston, P. (2018). On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1257-1278.

Kinoshita, Y., & Kiya, H. (2020). Fixed smooth convolutional layer for avoiding checkerboard artifacts in cnns. In *Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 3712-3716).

Kursinski, E., Hajj, G., Schofield, J., Linfield, R., & Hardy, K. (1997). Observing earth's atmosphere with radio occultation measurements using the Global

Positioning System. *Journal of Geophysical Research*, *102*, 23.429–23.465.

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E. H., ... Houston, M. (2018). Exascale deep learning for climate analytics. *CoRR*, *abs/1810.01993*.

Laloyaux, P., Bonavita, M., Chrust, M., & Gürol, S. (2020). Exploring the potential and limitations of weak-constraint 4d-var. *Quarterly Journal of the Royal Meteorological Society*, *146*(733), 4067-4082.

Laloyaux, P., Bonavita, M., Dahoui, M., Farnan, J., Healy, S., Holm, E., & Lang, S. T. K. (2020). Towards an unbiased stratospheric analysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 2392-2409.

Leinonen, J., Nerini, D., & Berne, A. (2021). Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7211-7223.

Loshchilov, I., & Hutter, F. (2019). *Decoupled weight decay regularization.*

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., ... Albers, J. (2020). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, *101*(5), E608 - E625.

Mlperf hpc deepcam website. (2021, 11). Retrieved from `https://github.com/mlcommons/hpc/tree/main/deepcam`

Poli, P., Moll, P., Puech, D., Rabier, F., & Healy, S. B. (2009). Quality control, error analysis, and impact assessment of FORMOSAT-3/COSMIC in numerical weather prediction. *Terrestrial, Atmospheric and Oceanic Sciences*, *20*, 101–113.

Polichtchouk, I., Shepherd, T. G., Hogan, R. J., & Bechtold, P. (2018). Sensitivity of the brewer–dobson circulation and polar vortex variability to parameterized nonorographic gravity wave drag in a high-resolution atmospheric model. *Journal of the Atmospheric Sciences*, *75*(5), 1525-1543.

Polichtchouk, I., Stockdale, T., Bechtold, P., Diamantakis, M., Malardel, S., Sandu, I., ... Wedi, N. (2019). *Control on stratospheric temperature in IFS: resolution and vertical advection* (Technical Memorenda No. 847). Shinfield Park, Reading RG2 9AX, United Kingdom: ECMWF.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F., & Simmons, A. (2000). The ECMWF operational implementation of four-dimensional variational assimilation. i: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, *126*(564), 1143-1170.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11).

Ray tune website. (2021, 11). Retrieved from `https://www.ray.io/ray-tune`

Rennie, M. P. (2010). The impact of GPS radio occultation assimilation at the Met Office. *Quarterly Journal of the Royal Meteorological Society*, *136*, 116–131.doi: 10.1002/qj.521.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation.*

Saunders, R. (2021). The use of satellite data in numerical weather prediction. *Weather*, *76*(3), 95-97.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., ... Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194).

Shepherd, T., Polichtchouk, I., Hogan, R., & Simmons, A. (2018). *Report on stratosphere task force* (Technical Memorenda No. 824). Shinfield Park, Reading RG2 9AX, United Kingdom: ECMWF.

Sun, B., Reale, A., Seidel, D. J., & Hunt, D. C. (2010). Comparing radiosonde and cosmic atmospheric profile data to quantify differences among radiosonde

types and the effects of imperfect collocation on comparison statistics. *Journal of Geophysical Research: Atmospheres*, *115*(D23).

Sun, B., Reale, T., Schroeder, S., Pettey, M., & Smith, R. (2019). On the accuracy of vaisala rs41 versus rs92 upper-air temperature observations. *Journal of Atmospheric and Oceanic Technology*, *36*(4), 635 - 653.

Trémolet, Y. (2006). Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, *132*(621), 2483-2504.

Vidard, P., Piacentini, A., & Dimet, F.-X. L. (2004). Variational data analysis with control of the forecast bias. *Tellus A*, *56*(3), 177-188.

Vorobev, V. V., & Krasilnikova, T. G. (1994). Estimation of the accuracy of the atmospheric refractive index recovery from doppler shift measurements at frequencies used in the NAVSTAR system. *USSR Phys. Atmos. Ocean, Engl. Transl.*, *29*, 602–609.

Waller, J. A., Dance, S. L., Lawless, A. S., & Nichols, N. K. (2014). Estimating correlated observation error statistics using an ensemble transform kalman filter. *Tellus A: Dynamic Meteorology and Oceanography*, *66*(1), 23294.

Watson, P. A. G. (2019). Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *Journal of Advances in Modeling Earth Systems*, *11*(5), 1402-1417.

Wergen, W. (1992). The effect of model errors in variational assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, *44*(4), 297-313.

Zupanski, M. (1993, 08). Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Monthly Weather Review*, *121*, 2396-2408.