AGU ADVANCING EARTH AND SPACE SCIENCE

JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2022MS003016

Special Section:
Data assimilation for Earth system models

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, UK, [2]Nvidia Corporation, Santa Clara, CA, USA

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 four-dimensional variational
- Transfer learning can help to mitigate data limitations when the atmospheric model is upgraded

Correspondence to:

P. Laloyaux,
patrick.laloyaux@ecmwf.int

Citation:

Laloyaux, P., Kurth, T., Dueben, P. D., & Hall, D. (2022). Deep learning to estimate model biases in an operational NWP assimilation system. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS003016. https://doi.org/10.1029/2022MS003016

Received 25 JAN 2022
Accepted 11 MAY 2022

Abstract Model bias is one of the main obstacles to improved accuracy and reliability in numerical weather prediction conducted with state-of-the-art atmospheric models. To deal with model bias, a modification of the standard four-dimensional variational (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 European Centre for Medium-Range Weather Forecasts operational forecasting system. Despite different origins and applications, data assimilation (DA) 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 (NNs) require a large number of samples to properly capture the relationship between the temperature first-guess trajectory and the model bias. As running the Integrated Forecasting System (IFS) DA 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 NNs are trained using the ERA5 reanalysis before using transfer learning on 1 year of the current IFS model. Preliminary results show that convolutional NNs 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 In the practice of the numerical weather prediction, 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 (DA). Weather predictions could be improved if the error of the numerical models that are used could be reduced. Recent advances in DA at the European Centre for Medium-Range Weather Forecasts indicate that it is possible to estimate and correct for a large fraction of systematic model errors of those models. During DA, 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 DA 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 DA.

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). ML applications can be very complex, and neural networks (NN) can consist of millions to billions of trainable parameters, large numbers of layers, and specialized architectures. In recent years, the weather and climate modeling community has started to explore ML 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 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 (Groenquist et al., 2021; Rasp & Lerch, 2018), 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, to avoid shocks during model initialization.

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 (RO) 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 1 billion grid points that are updated several times per hour, while only 40 million observations are processed every 12 hr. 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 DA 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, 2008b). 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 DA 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 flavors 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 data set 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 DA 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 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 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 1 year of training data is available for a new model cycle. Finally, we perform tests that apply the NN bias correction within 4D-Var DA experiments and compare results against weak-constraint 4D-Var which serves as a benchmark.

The importance of tackling stratospheric model biases is described in Section 2. A description and assessment of the RO temperature retrieval data set are presented in Section 3. The design and the training of various NN solutions are summarized in Section 4. Section 5 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 6. Various aspects of weak-constraint 4D-Var that are also essential for ML such as learning rate and NN retraining are discussed in that section. We summarize the paper in Section 7 and provide a perspective for future developments.

2. Stratospheric Temperature Bias

This paper 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 DA experiment where observations sensitive to stratospheric model variables are withheld (denial experiment). This includes stratospheric observations from radiosondes, aircraft, RO bending angles above 100 hPa, as well as the microwave and infrared stratospheric channels (see details in Table 1). It is not possible to remove all observations that are sensitive to the stratospheric conditions, as microwave instruments measure radiances that reflect conditions 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 2 months, between 25 January 2020 and 25 March 2020. The top panel of Figure 1 shows the impact on the mean analysis error when stratospheric observations are withheld. The large biases developed in the stratosphere over these 2 months are transferred to the troposphere, especially over the Southern pole. The bottom panel of Figure 1 shows how these biases present in the analysis evolve during forecasts. The impact of the missing stratospheric observations can still be observed after 48 hr. The impact shrinks as the forecast time increases as the model drifts toward 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.

Table 1
List of all the Observations Considered as Sensitive to Stratospheric Conditions and Withheld in the Denial Experiments

Type	Pressure/altitude/channels
Radiosonde	Above 100 hPa
Aircraft	Above 100 hPa
RO	Above 17 km
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

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 −0.5°C and a global-mean warm bias in the upper stratosphere of 0.5°C that accumulate over a 12-hr DA window.

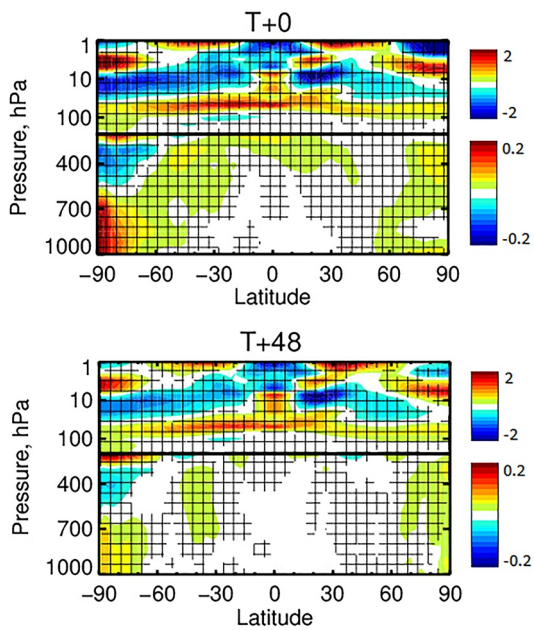

Figure 1. Difference in the mean forecast error across zonal bands at lead time +0 hr (top) and +48 hr (bottom) when all the stratospheric observations valid above 100 hPa are withheld. Scores have been computed against the operational analysis between 25 January 2020 and 25 March 2020. Different colorbars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

3. Temperature Retrieval Data Sets

It is very challenging to produce a ground-truth database for 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 Global Navigation Satellite System Radio Occultation 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 centers (Cucurull et al., 2013; Healy & Thépaut, 2006; Poli et al., 2009; Rennie, 2010).

The RO measurement technique is described in Kursinski et al. (1997). The Global Positioning System (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 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 with a good vertical resolution (several 1,000 bending angle measurements are available and spread homogeneously for each kilometer between the altitudes of 2 and 50 km). These RO bending angles can be assimilated without bias correction into the NWP model (Healy &

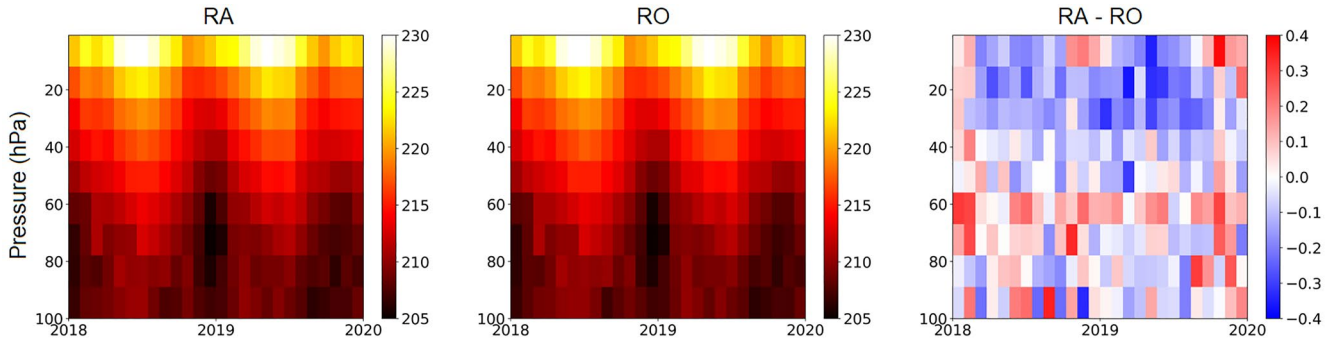

Figure 2. Timeseries of collocated radiosondes (left), radio occultation temperature retrievals (middle) and the difference between the two (right). Observations are collocated on a 5° grid every hour within a 1 hPa pressure difference between 2018 and 2020.

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 in 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 3 hPa. Therefore, we only use the retrieved temperature values between model level 20 (3 hPa) and model level 65 (125 hPa) 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 location, horizontal location and time. For comparison, profiles have been collocated on a 5° grid within a 1 hPa pressure difference and valid at the same hour. 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 0.2°C and confirms what has been found in other collocation studies (Sun et al., 2010, 2019). In the upper stratosphere (above 30 hPa), there is a systematic difference where RO observations are warmer than RA by approximately 0.3°C, 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 adjustment, …).

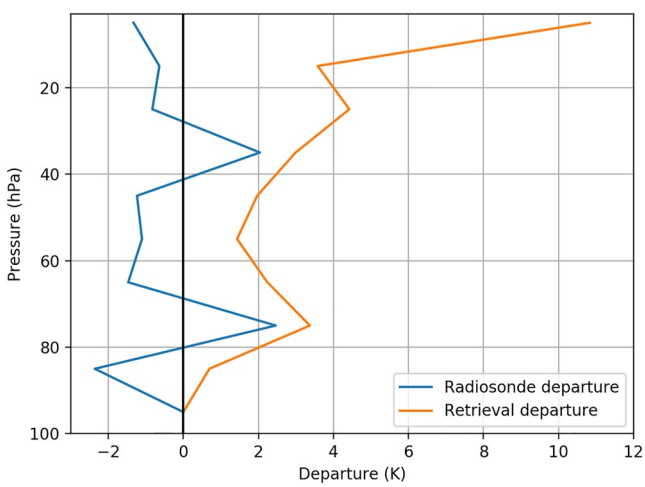

Figure 3. Vertical profile of the ERA5 first-guess departure (OBS-ERA5) from a collocated radiosonde (blue) and radio occultation temperature retrieval (orange). Both profiles are measured over the USA (36°N, 93°W) on 16 October 2020.

During the collocation study, a small fraction (less than 1%) of profiles showed very large discrepancies (difference of several degrees). One example is illustrated in Figure 3 for a collocated profile over the USA (36°N and 93°W) on 16 October 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

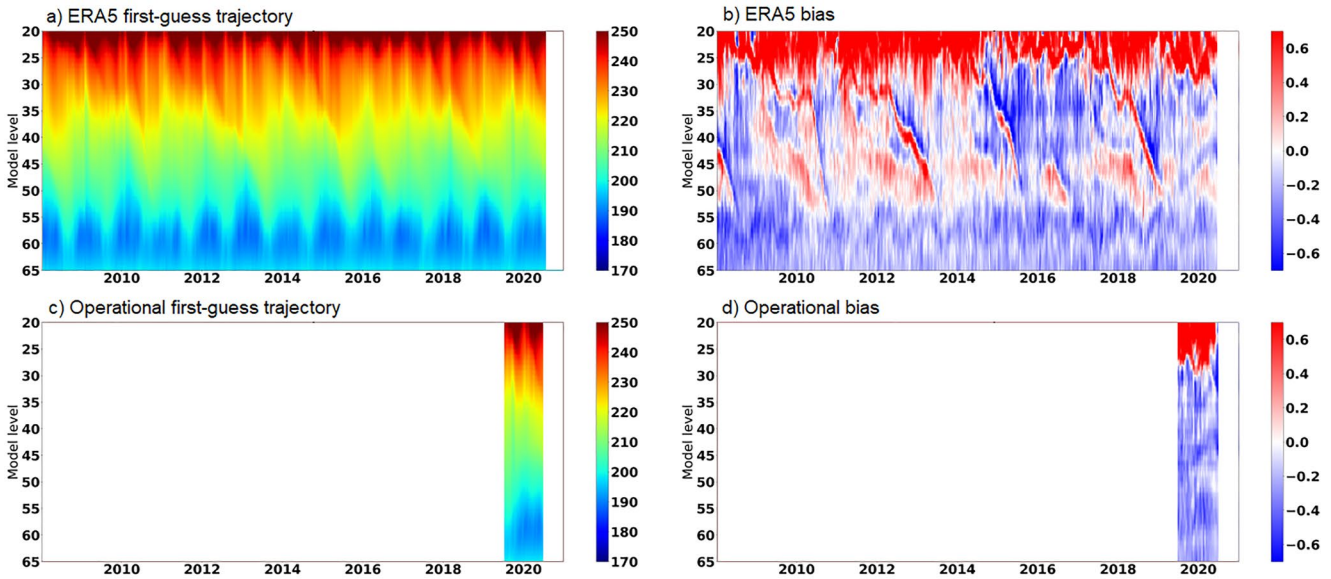

Figure 4. Timeseries of ERA5 temperature first-guess (top left) and departure with radio occultation temperature retrievals (top right) for the different stratospheric model levels (level 20 is 3 hPa and level 65 is 130 hPa) averaged between 5°N and 5°S available between 2008 and 2020. The bottom panels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

ERA5 (over 5° 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 DA system over the 12-hr 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 departures are averaged over a 10° regular grid for all the model levels between 130 hPa (level 65) and 3 hPa (level 20). This means that we have 31,635 inputs and the same number of outputs (19 latitude grid points × 37 longitude grid points × 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).

ML 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 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 1 January 2008 until 1 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 data set as the "operational data set." It consists of 1 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 data set. The spatial resolution of the two data sets is identical and equal to 18 km (the control member of the Ensemble Data Assimilation system is used for the operational data set, instead of the high-resolution system). Figure 4 shows a timeseries of inputs and outputs produced from the ERA5 (top) and the operational (bottom) data set, averaged over the Tropics between 5°S and 5°N. The ERA5 model exhibits a cold bias in the

Table 2
Partition Details for ERA5 and Operational Data Sets

Data set	Training	Validation	Test	Sizes (MB)
ERA5	2008–2019 (412)	2019 (26)	2020–2021 (42)	150/10/16
Operational	2019 (15)	2019 (5)	2020 (18)	6/2/7

Note. 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, to create disjoint sets. The numbers in the last columns represent the sizes of the individual splits in MB for training, validation, and test, respectively.

mid/lower stratosphere and a warm bias in the upper stratosphere that propagates down during Quasi-Biennial Oscillation events. The operational data set has a similar vertical structure, although the amplitude is much larger. This larger amplitude is mainly due to the higher resolution used in the operational model that introduces discretization errors in the vertical advection, associated with inadequate representation of resolved gravity waves in the vertical direction (Polichtchouk et al., 2019; Sandu et al., 2019).

ML studies generally divide the available data into three different data sets 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 variables. 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, to avoid overfitting and ensure that the network will function well on previously unseen data. The third data set is the test set, a collection of previously unseen data, which is used to evaluate the network. The three data sets 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 to mitigate this issue. For the ERA5 data sets, the test set is comprised of the entire year 2020 and the first half of 2021. The validation set contains samples from the year 2019, discarding every third. The remaining samples from 2019 and the previous years are used for the training set. For the operational data set, we use a similar strategy for splitting the samples. We assign the full set of 2020 operational data to the test set. Since only data for half of 2019 is available, the validation set contains only samples from July and August 2019, discarding every third. The rest of the 2019 data are used for the training set. Table 2 summarizes the various splits for the two data sets.

4. Design and Training of Neural Networks

4.1. Data Representation

The ML 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. This procedure typically results in a more stable learning process.

The raw data is available as tuples in the form (longitude, latitude, level, T), where T represents the temperature at these coordinates. In this paper, we make use of data regression on structured grids using convolutional NNs, after converting the data into multi-channel images using suitable projections and interpolations, with the vertical model level mapped to the feature/channel dimension. The advantage of this approach is that traditional convolutional NNs can be applied to the data. The disadvantage is that this data representation does not take into account that not all neighboring grid points are equidistant, that is, two neighboring grid points near the poles are separated by a smaller physical distance than those near the equator (One could use the approximate surface area around these points as a weight to mitigate this effect, however we did not find significant improvement of our results when applying this method).

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) (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), or as two-dimensional fields of (projected longitude and latitude) with a vector of features (temperatures at different altitudes). We will discuss the implication of these two different views below.

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

4.2. Network Architecture

Using the grid representation for the data, we have mapped the problem to an image regression problem. Image regression means that for every pixel of an input image with a fixed number of input channels/features, a fixed number of output channels/features is predicted for each pixel of another image. In our case, the input and output images have the same height, width and depth (or features, depending on whether we regard the data as 3D or 2D). Image segmentation is very similar to image regression, with the only difference that the output features are categorical, that is, for each pixel of the output image we predict a class value. Therefore, every image segmentation network can be transformed into an image regression network by replacing the classifier with a small regression network: for example, by removing the final per-pixel softmax layer and replacing it with a small convolutional NN or a multilayer perceptron (MLP). In this context, we can choose a suitable network architecture from a plethora of available image segmentation networks, such as the UNet (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 (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). 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 and 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 neighboring levels. Therefore, they 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 contain around 90 millions of parameters and are reasonable choices for solving the bias prediction problem at hand, therefore we pursued both approaches.

4.3. Training Process, R^2-Score and Hyper Parameter Optimization

We employ the AdamW optimizer (Loshchilov & Hutter, 2019) and apply weight decay regularization to reduce overfitting, which is particularly important when training on the smaller, operational data set. 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 R^2 score is used as a validation metric for hyper parameter tuning. It is defined as

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

Here, $y^{(i)}$ is the NN prediction for sample i and $f^{(i)}$ is the corresponding ground truth, that is, in our case the model bias. The R^2 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 $R^2 \approx 1$. 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 $R^2 \approx 0$. For predictions of even lower accuracy we have $R^2 < 0$ which signals a failure of the model. 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 R^2 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 maximization of the R^2 value, as described above. Each model is trained for about 150 epochs.

4.4. Computational Performance

We use a single NVIDIA DGX-2 system (comprised of 16 V100-16 GB GPUs, 2 Intel Xeon E5-2689 CPUs with 40 logical cores each and 512 GB DRAM in total) 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 min for the 3D model. Therefore, training 128 instances does not take longer than 4 hr in total. It is unlikely that training more instances would lead to the discovery of better hyperparameters, because many hyperparameter choices with a high R^2 score on the validation set perform equally well and it is hard to define a quantitative criterion which configuration to prefer over the others.

4.5. Training, Using a Small Operational Data Set

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 data set for the current cycle is much smaller. In our case, we had only 15 training samples and 5 validation samples available (remember that each individual sample is averaged over 10 consecutive days). We examined several approaches in an attempt to build the most useful tool for this scenario.

1. *Finetuning*: In this approach, the model training does not start from randomly initialized weights. Rather, training starts from weights that have been pre-trained on a related data set. Specifically, we pre-trained the model using the ERA5 data set and then fine-tuned the entire model using only the operational data set. We experimented with a factor of 10–100 smaller learning rates compared to the pre-training to prevent the model from overfitting. Nevertheless, Using this approach, we found that the NN quickly overfit to the operational data, likely due to the very small size of the training data set. 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 data set 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 data set, with no fine-tuning. This model was then applied directly to the shorter, operational data set. This approach is promising if the underlying intrinsic features of both data sets are similar. It turns out that this approach yields reasonable results, producing R^2 values *only* about ∼20% lower than the original ERA5 test data set.
4. *Training on both data sets simultaneously*: For this transfer learning strategy, we implemented a data loader which can feed the NNs samples from either data set. The two data sets have a relative sample imbalance of about 27:1 (ERA5:operational). To help the NN learn the features of the operational data set, the dataloader selects samples from the both data sets, but with inverted frequencies. This means that the NN is presented samples from both data sets with almost equal probability. In practice, we chose a final ratio of 27:29 (ERA5:-operational data) to provide a small emphasis on the importance of the operational data. (This is a tunable hyperparameter.) For the validation data set, 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 R^2 score when performing hyper-parameter optimization. It turns out that this approach leads to stable training without significant overfitting and further delivered slightly better results than the previous approach.

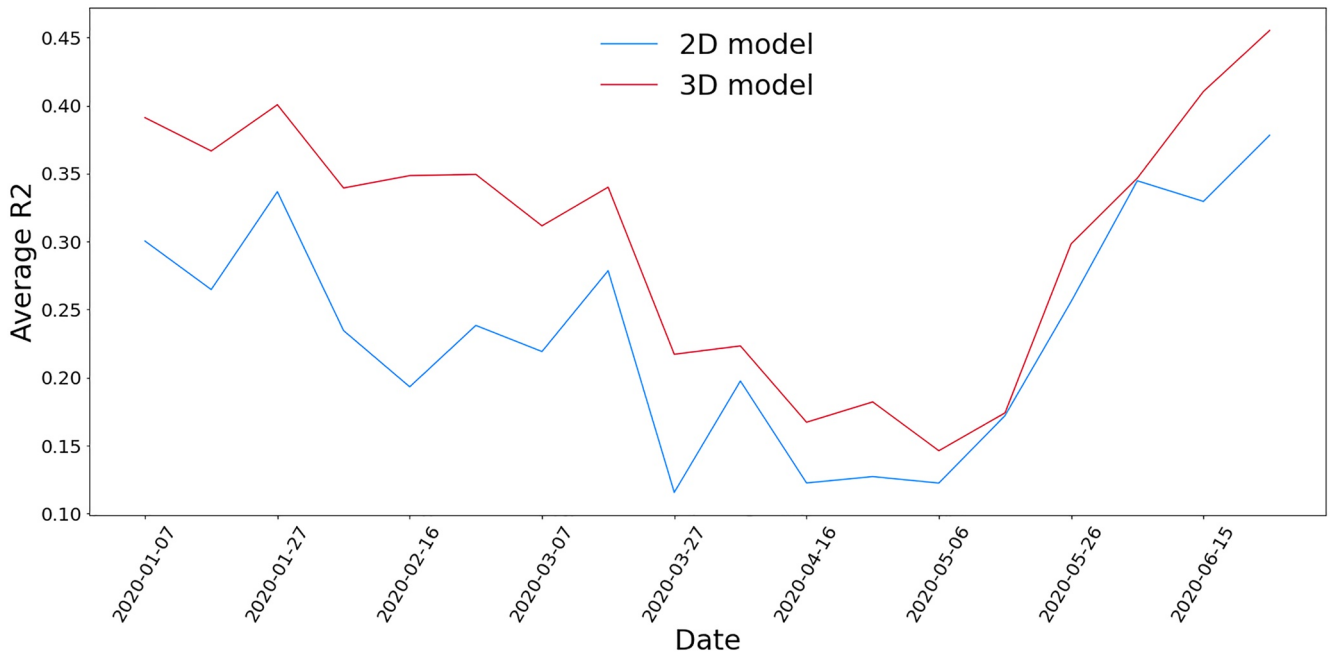

Figure 5. Timeseries of R^2 values averaged over all levels for 2D + features (blue) and 3D NN (red) architectures on the 41r2 test data set.

5. 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 data sets respectively. The third subsection discusses the use of bias correction within DA experiments.

5.1. Performance Comparison of 2D and 3D Models

We trained the 2D and 3D models with their respective best known hyperparameters on the ERA5 training data set and compared their performance on the ERA5 test data set. Figure 5 displays the R^2 value, averaged over all levels, for the test set. The plot demonstrates that the 3D model outperforms the 2D model consistently. It also outperforms the R^2 values obtained by the 1D column approach described in Bonavita and Laloyaux (2020), although it is not exactly applied to the same data set. The good performance of the 3D model is likely due to the importance of gradients and other local covariances 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 R^2 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 6 February 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 negative bias near (60° lat, 275° lon) as well as in the latitude band between 20° and 40°.

5.2. Performance of 3D Models on ERA5 and Operational Data Sets

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 4.5.

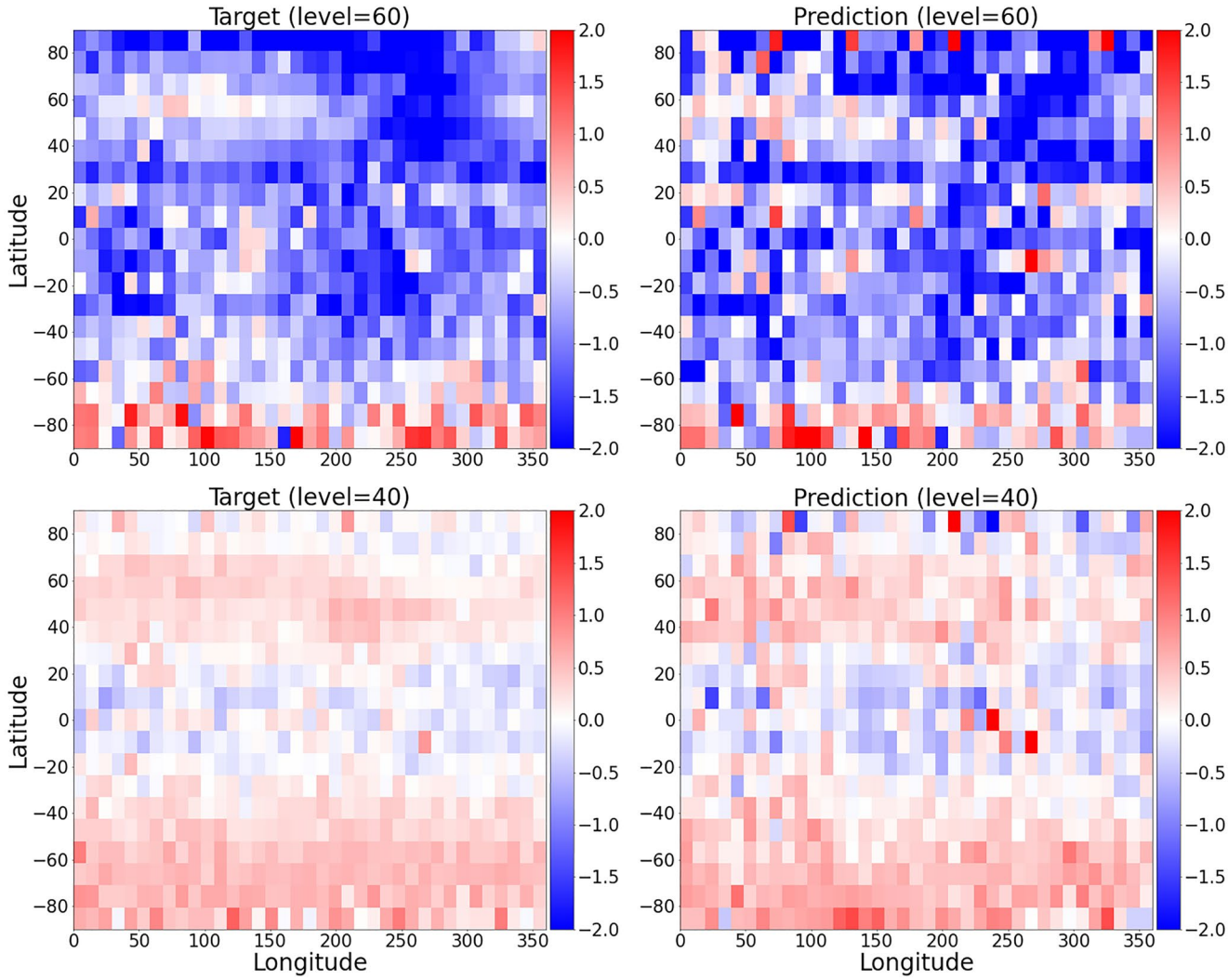

Figure 6. Target bias (left) and bias prediction from the NN model (right) for two levels on 6 February 2020.

Figure 7 shows a vertical profile of the globally averaged R^2 scores for each of the three test cases. We observe a steep drop in prediction quality when testing on the operational data set. The retrained model produces a better prediction than the original ERA5-only model on the operational data set, except for the top-most levels. Comparing the model biases from the ERA5 and operational data sets (panels b and d in Figure 4), we see that the retrained model struggles to capture the larger warm bias in the top levels of the operational data set.

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 hemisphere 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.

5.3. 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 DA experiments, to produce a better analysis. Unfortunately, it is technically challenging to introduce the bias correction tools into the workflow of the 4D-Var DA experiments. Not only is it difficult to couple the ML tools

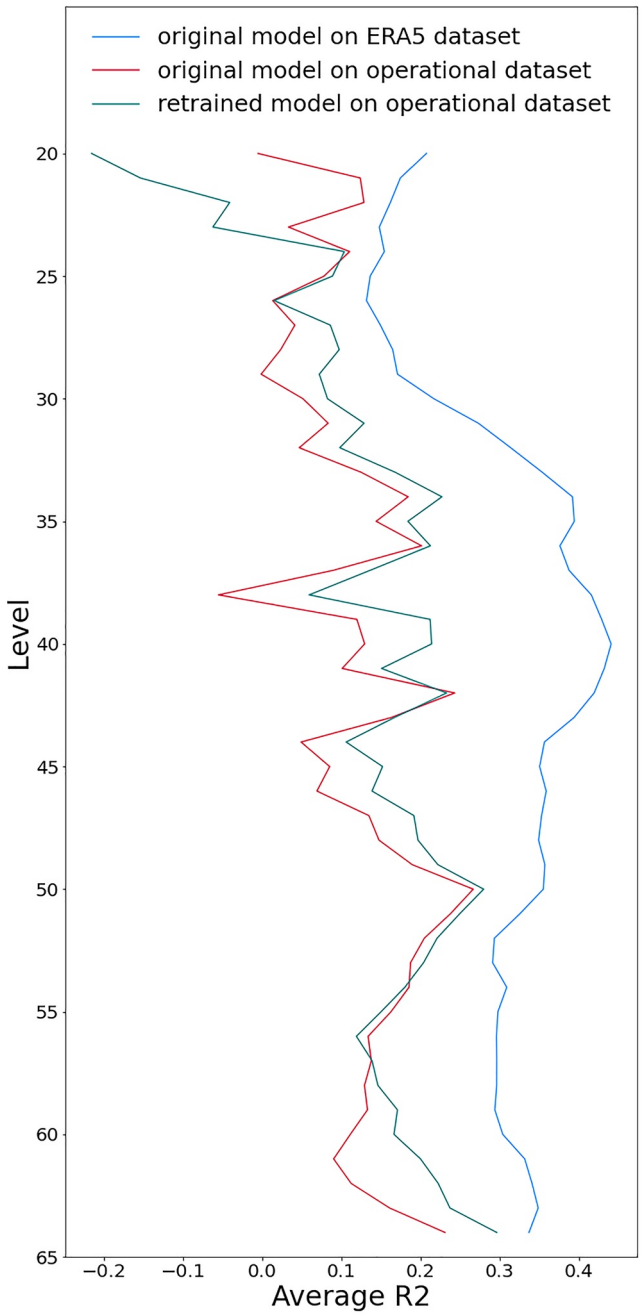

Figure 7. Vertical profile of globally averaged R^2 values for the original model on the ERA5 target (blue), for the same model on the operational target (red) and for a retrained model using the sample balancing technique described above, on the operational target (green).

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 data set. 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. This is practically done by assuming a linear growth of errors in time, spreading the offline correction throughout the assimilation window. Using this framework, the ML approach is evaluated in 4D-Var over the test period between 1 January 2020 and 1 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 data sets. This provides a estimate of how much the bias could actually be reduced if the NNs provided a perfect fit to the training data. One can see that the model bias is almost everywhere over-corrected, for example, the original 0.4°C cold bias at 60 hPa became a −0.6°C warm bias. This over-correction is due to the intrinsic cycling principle of DA 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 with respect to radiosondes shows that only one quarter of the global mean departure comes from the current assimilation cycle while the other three quarters are carried forward in time from the previous cycle (e.g., at 50 hPa, the global-mean analysis departure is equal to 0.38 and the global-mean background departure is equal to 0.5). As the NN aims at correcting the model bias developing only over the 12-hr assimilation window, only a fraction of the NN correction should be applied. This approach where the target is divided by four is plotted in dash-dot blue and is able to efficiently correct the model bias for the entire stratosphere. The rescaling parameter has been determined empirically and future work could investigate how to provide an optimal forcing to the model. 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 0.2°C in the mid/lower stratosphere. The poor performance around 5 hPa where the model is over-corrected is likely due to the small size of the operational data set. The ERA5 warm bias at 5 hPa 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.

It is important to note that these results have been obtained with an offline NN approach where the corrections for the whole period are computed before running the 4D-Var experiments. This is potentially suboptimal as we are not feeding the NN with the actual first-guess trajectory computed every 12-hr by the 4D-Var system. An online

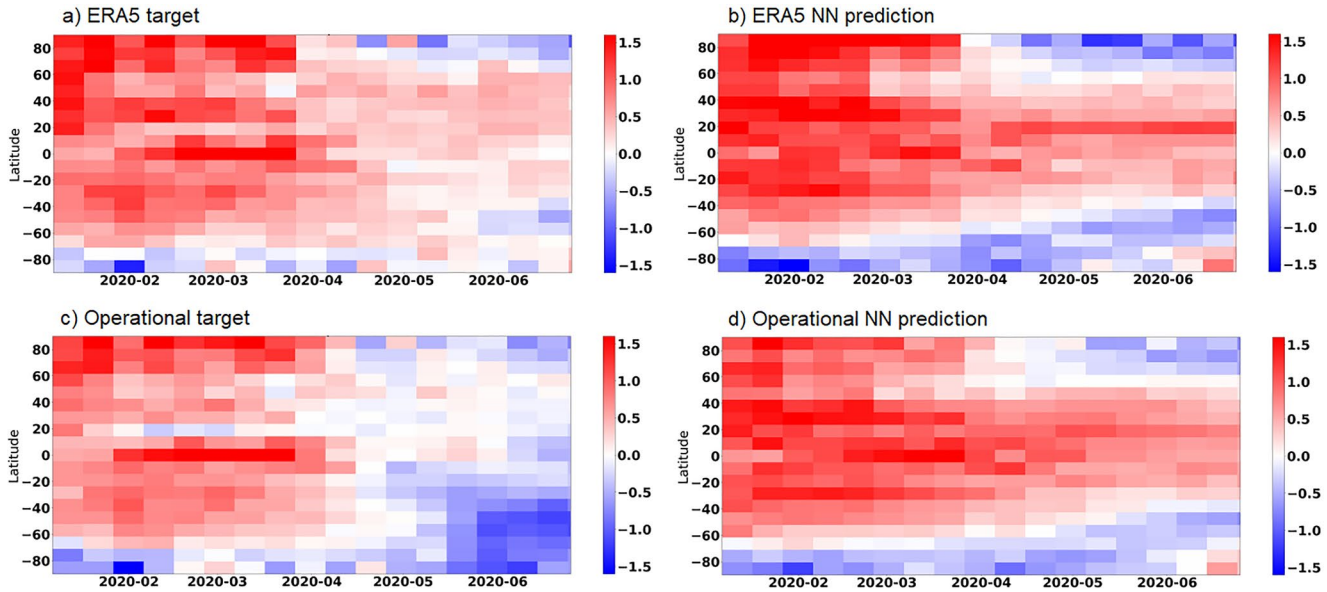

Figure 8. Zonal timeseries of the ERA5 targets (a) and the bias predictions from the original NN for model level 25 (6 hPa). The bottom panels show similar timeseries for the operational targets (c) and the bias predictions from the retrained NN (d).

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 is outside the scope of this paper and would require a substantial effort, given the current software infrastructure.

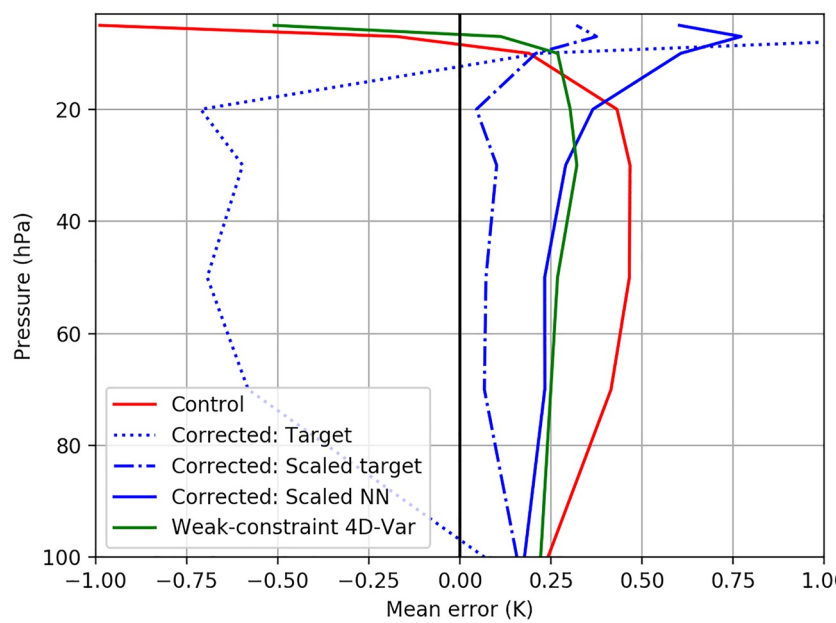

Figure 9. First-guess mean error with respect to radio occultation temperature retrievals for the control (red), for weak-constraint 4D-Var (green), for the model corrected with the target (dotted blue), with the scaled target (dash-dot blue) and with the scaled prediction of the NN (solid blue). Statistics are averaged over the globe between 1 January 2020 and 1 March 2020.

6. 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 (Bennett et al., 1996; Dee, 2005; Trémolet, 2006; Vidard et al., 2004; Wergen, 1992; Zupanski, 1993). 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 $\boldsymbol{\eta}$ to represent the error which gradually enters into the model trajectory

$$\mathbf{x}_k = \mathcal{M}_{k,k-1}\left(\mathbf{x}_{k-1}\right) + \boldsymbol{\eta} \quad \text{for} \quad k = 1, \ldots, N \tag{2}$$

where N is the number of the model time steps. The model error forcing is assumed to be additive and constant within the 12-hr assimilation window (Laloyaux, Bonavita, Chrust, & Gürol, 2020; Laloyaux, Bonavita, Dahoui, et al., 2020). It contains temperature, vorticity and divergence. We also assume that the model error $\boldsymbol{\eta}$ 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

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

which depends upon the value of the model initial condition $\mathbf{x}_0$ and the model error forcing $\boldsymbol{\eta}$. This information is carried forward in time between assimilation cycles with the background of the model state $\mathbf{x}_0^b$ and of the model forcing $\boldsymbol{\eta}^b$. The error statistics of the model state, model error forcing and observations $\mathbf{y}_k$ are specified in $\mathbf{B}$, $\mathbf{Q}$, and $\mathbf{R}_k$, respectively. This forcing formulation of weak-constraint 4D-Var simultaneously estimates the initial state $\mathbf{x}_0$ and model forcing $\boldsymbol{\eta}$ that best fit the observations and the background information with respect to their error covariance matrices.

6.1. Comparison With the NN Approach

A weak-constraint 4D-Var experiment was run from 1 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 1 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 data set for offline training.

Weak-constraint 4D-Var can be seen as a specific ML algorithm 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 ML 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. A last difference is the use of the model error covariance matrix $\mathbf{Q}$ in weak-constraint 4D-Var that specifies the error statistics of the model bias (Laloyaux, Bonavita, Dahoui, et al., 2020). 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 DA 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.

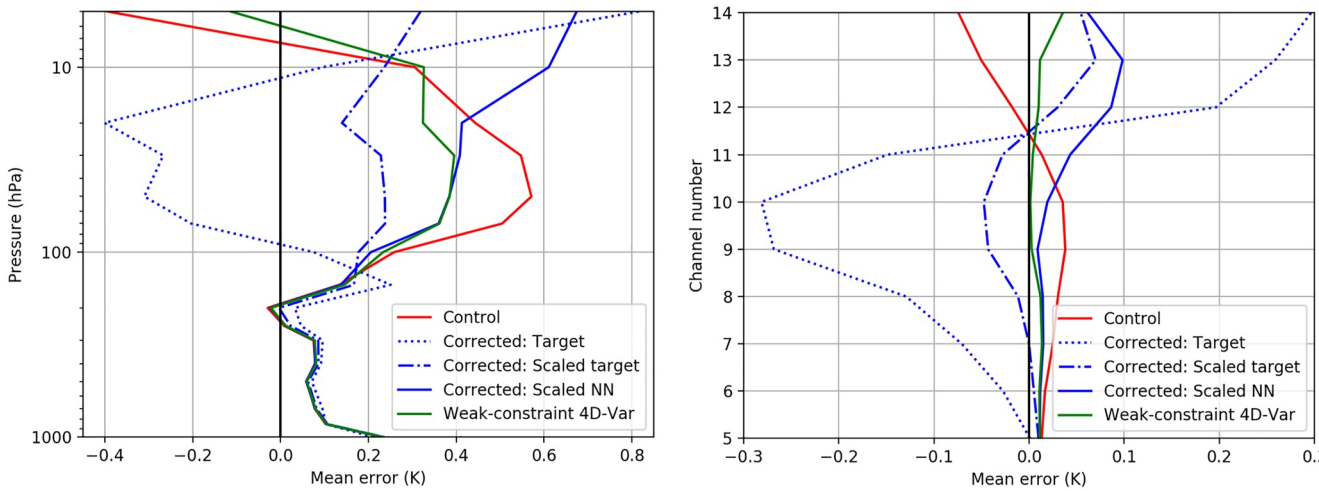

Figure 10. Same as Figure 9, but respect to radiosondes (left) and AMSU-A (right).

The scaled NN shows a similar improvement than weak-constraint 4D-Var in the lower and mid stratosphere (i.e., radiosondes below 30 hPa and AMSU-A channel numbers below 10). This is an excellent news that can possibly be explained by the fact that RO, radiosonde and AMSU-A observing systems are consistent with 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 30 hPa and AMSU-A channel numbers above 11) which confirms what has been noticed in Figure 8 against RO retrievals. We have run 10-day forecasts initialized 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 5 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 20 hPa (see Figures 9 and 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 fulfill 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 1 January 2020 and 1 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 $\mathbf{Q}$ 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 weak-constraint 4D-Var correction. This larger NN correction is the reason for the degradation observed in Figures 9 and 10 for the top of the stratosphere.

6.2. Weak-Constraint 4D-Var Learning Rate

We already discussed the question of retraining the NN in Section 5 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 data set 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

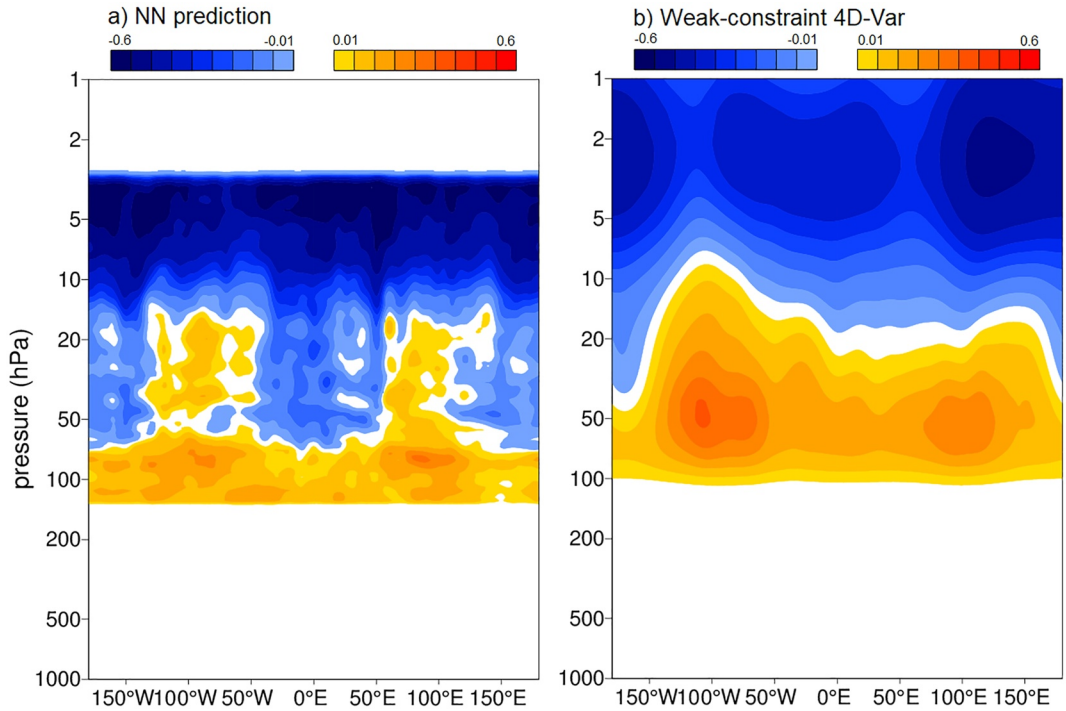

Figure 11. Meridional cross-section temperature error correction from the from the NN prediction (a) and from weak-constraint 4D-Var (b) averaged over the tropics (10°N–10°S) between 1 January 2020 and 1 March 2020.

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

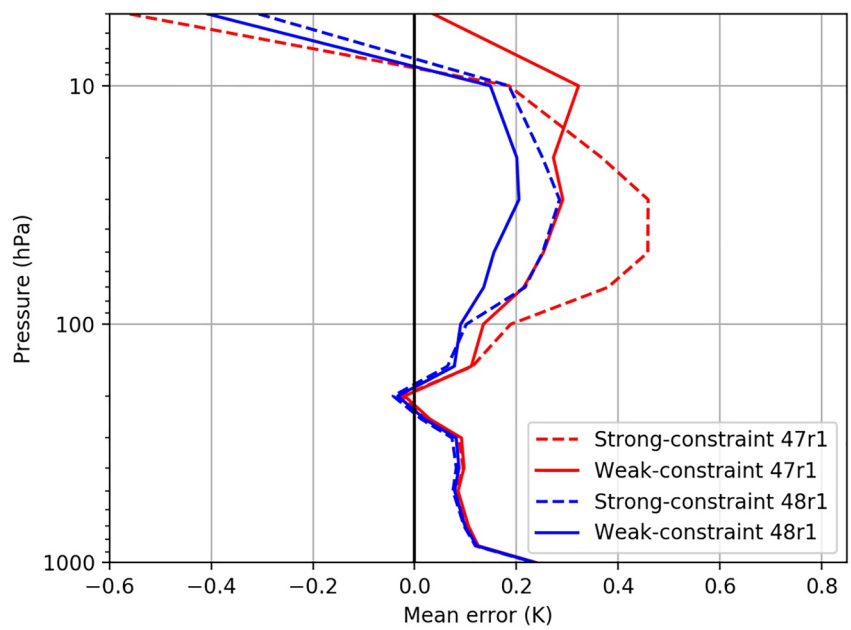

Figure 12. Vertical profile of first-guess departure with respect to radiosondes for 47r1 strong-constraint (dashed red), for 48r1 strong-constraint (dashed blue), for 47r1 weak-constraint (solid red), and for 48r1 strong-constraint (solid blue). Statistics are averaged over the globe between 20 January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

biases. At 50 hPa, 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.

The initialization of the model error correction at the beginning of an experiment can be compared to the challenge of initializing 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 study the sensitivity of the initialization, a weak-constraint 4D-Var experiment was run where the model error correction is initialized 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 toward the same solution although the behavior 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 initialized and its fast learning rate. 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 occurring 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°C over a few days, in the vertical region between 1 and 10 hPa. 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 (Mariotti et al., 2020; Polichtchouk et al., 2018). 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 (70 N 90 N) between 24 September 2020 and 24 February 2021. At the beginning of the SSW event (1 January 2021), the structure of the model bias changes significantly as stratospheric dynamics are disrupted. There is a model cold bias above 3 hPa and a model warm bias between 50 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 3 hPa and cools down between 50 and 3 hPa. 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 data set (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 data set and it might be challenging for the NN to correctly represent the model error structure.

7. Summary and Perspectives

Artificial intelligence and ML 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 DA and ML 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 the three-dimensional model bias computed from RO temperature retrievals. While large data sets containing several years of data are

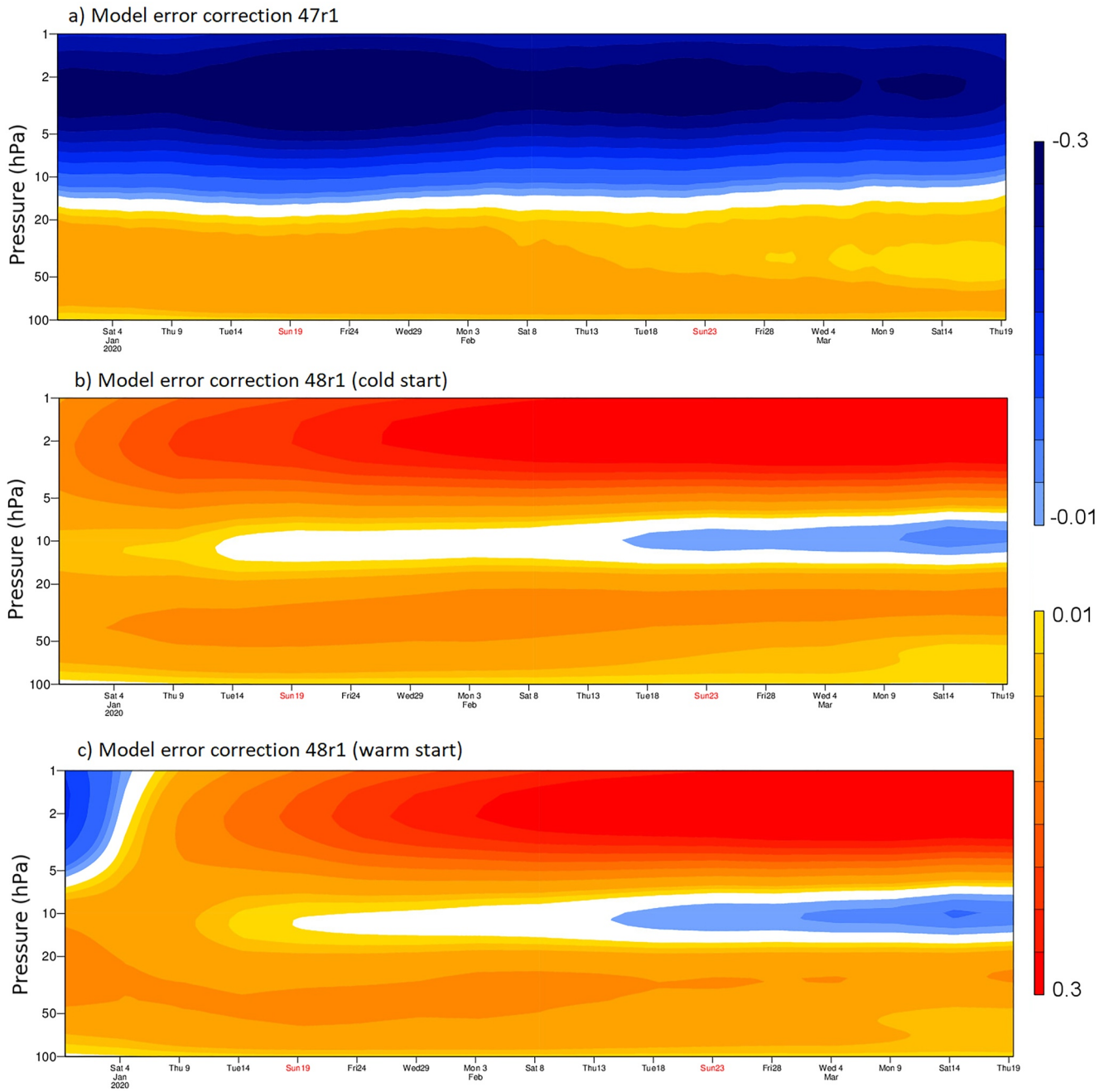

Figure 13. Timeseries of model error corrections estimated by weak-constraint 4D-Var for the 47r1 model (top), for the 48r1 model initialized from zero (middle), and for the 48r1 model initialized from the 47r1 bias estimate (bottom) between 1 January 2020 and 20 March 2020. Statistics are averaged between 70° and 30°S.

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 DA 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" ML method, where observations over the last 12 hr are used to update the previous weather forecasts. The ML 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

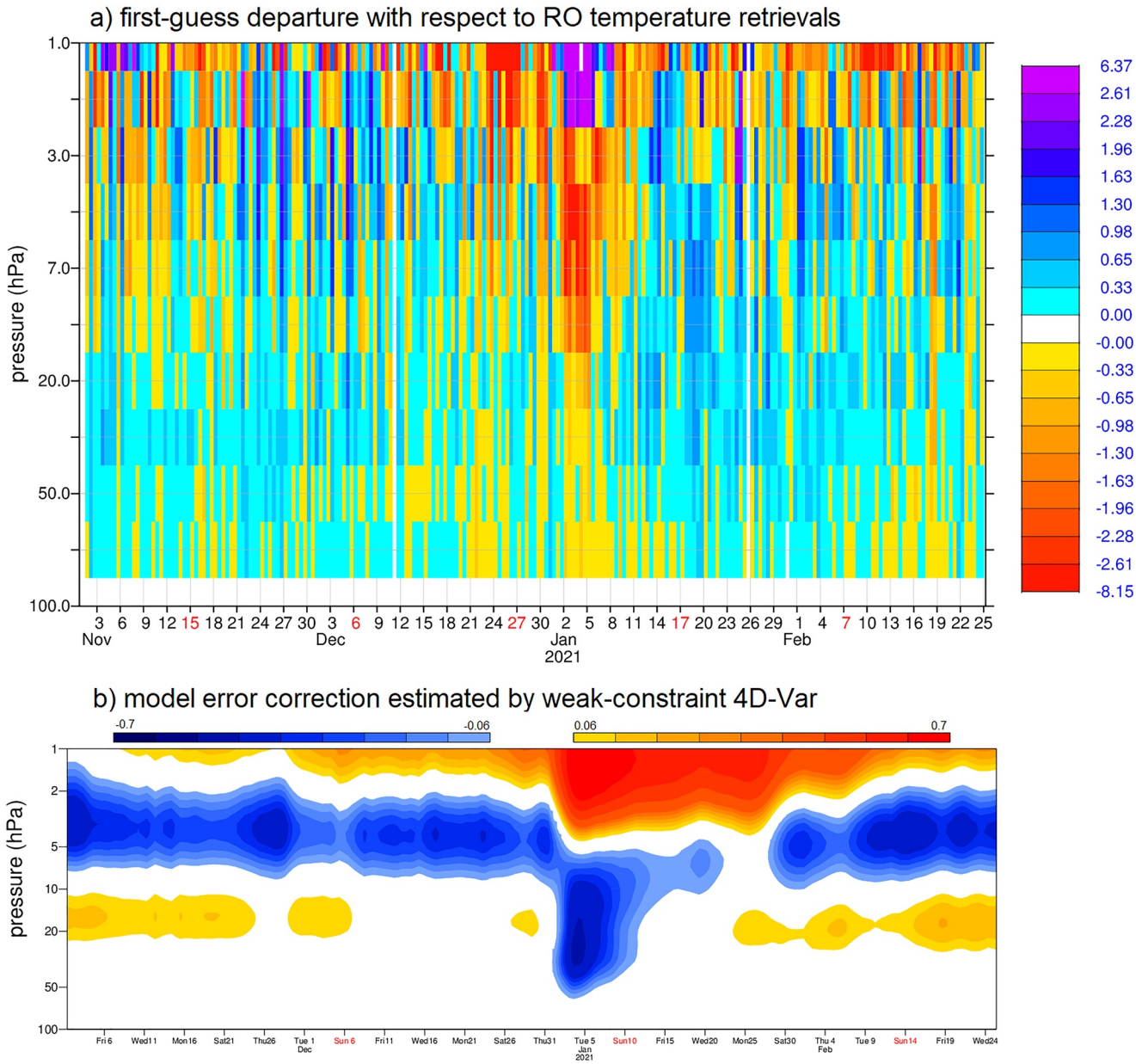

Figure 14. Timeseries of first-guess departure with respect to radio occultation temperature retrievals (top) and timeseries of model error correction estimated by weak-constraint 4D-Var (bottom). Statistics are averaged over the Northern pole (70°–90°N) between 24 September 2020 and 24 February 2021.

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 adjustments. 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 data set 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 sparse 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 (c.f., 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. The IFS model is upgraded at every cycle to better represent physical processes or introduce new ones that were missing. This Research-to-Operations (R2O) process, which is followed to upgrade the software used in forecast production is one of the key aspects of ECMWF business (Buizza et al., 2017). R2O includes a series of actions that could be summarized in six activities: planning, development, testing, evaluation, communication, and implementation. This procedure has already a very tight schedule and the best time to train a NN will have to be decided to be as least disruptive as possible if a NN solution wants to be implemented in operations. Finally, 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 data set or to retune parameters. Specific solutions are required to achieve a similar flexibility with a NN.

Data Availability Statement

The input and output data of the experiments described in the paper is freely available for research purposes from European Centre for Medium-Range Weather Forecasts and can be requested following the procedures described in https://www.ecmwf.int/en/forecasts/datasets.

Acknowledgments

The authors acknowledge the work done by the two reviewers who have provided a critical reading with helpful comments on an earlier version of the manuscript. P. D. Dueben gratefully acknowledges 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. This project receives support from the European Union's Horizon Research and Innovation Program and United Kingdom, Germany, Italy, Luxembourg, Switzerland, and Norway.

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. https://doi.org/10.1002/qj.339

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. https://doi.org/10.1002/qj.340

Bennett, F., Chua, A., & Leslie, B. (1996). Generalized inversion of a global numerical weather prediction model. *Meteorology and Atmospheric Physics*, *60*(1–3), 165–178. https://doi.org/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. https://doi.org/10.1002/qj.2652

Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. *Journal of Advances in Modeling Earth Systems*, *12*(12). https://doi.org/10.1029/2020ms002232

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*, 101171. https://doi.org/10.1016/j.jocs.2020.101171

Buizza, R., Andersson, E., Forbes, R., & Sleigh, M. (2017). *The ECMWF research to operations (R2O) process* (Technical Memorandum No. 806). 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. https://doi.org/10.1109/tpami.2017.2699184

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *arXiv:1706.03059v2* [cs.CL].

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*(1), 14–28. https://doi.org/10.1029/2012jd017782

Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3323–3343. https://doi.org/10.1256/qj.05.137

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. https://doi.org/10.1002/qj.493

Defferrard, M., Milani, M., Gusset, F., & Perraudin, N. (2020). DeepSphere: a graph-based spherical CNN. *arXiv preprint arXiv:2012.15000*. [cs. LG]. https://doi.org/10.48550/arXiv.2012.15000

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

Eyre, J. R. (1994). *Assimilation of radio occultation measurements into a numerical weather prediction system* (Technical Memorandum No. 199). 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, *Journal of computational science*, *55*, 101468. https://arxiv.org/abs/2107.11114

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. https://doi.org/10.1002/qj.4116

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*, *107*(D10). https://doi.org/10.1029/2001jd000479

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

Geer, A. J., Lonitz, K., Weston, P., Kazumori, M., Okamoto, K., Zhu, Y., et al. (2018). All-sky satellite data assimilation at operational weather forecasting centres. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1191–1217. https://doi.org/10.1002/qj.3202

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. 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 & Engineering Sciences*, *379*(2194), 20200092. https://doi.org/10.1098/rsta.2020.0092

Haiden, T., Dahoui, M., Ingleby, B., de Rosnay, P., Prates, C., Kuscu, E., et al. (2018). *Use of in situ surface observations at ECMWF* (Technical Memorandum No. 834). 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)* (pp. 770–778). https://doi.org/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*(615), 605–623. https://doi.org/10.1256/qj.04.182

Hogan, R., Ahlgrimm, M., Balsamo, G., Beljaars, A., Berrisford, P., Bozzo, A., et al. (2017). *Radiation in numerical weather prediction* (Technical Memorandum No. 816). ECMWF.

Hyperopt Website. (2021). Retrieved from http://hyperopt.github.io/hyperopt/

Janjic, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., et al. (2018). On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1257–1278. https://doi.org/10.1002/qj.3130

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)* (pp. 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*(D19), 23429–23465. https://doi.org/10.1029/97jd01569

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E. H., et al. (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. https://doi.org/10.1002/qj.3891

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. https://doi.org/10.1002/qj.3798

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. https://doi.org/10.1109/tgrs.2020.3032790

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *arXiv:2107.11114* [stat.ML].

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., et al. (2020). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, *101*(5), E608–E625. https://doi.org/10.1175/bams-d-18-0326.1

MLPerf HPC DeepCAM Website. (2021). 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*(1), 101–113. https://doi.org/10.3319/tao.2008.01.21.02(f3c)

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. https://doi.org/10.1175/jas-d-17-0304.1

Polichtchouk, I., Stockdale, T., Bechtold, P., Diamantakis, M., Malardel, S., Sandu, I., et al. (2019). *Control on stratospheric temperature in IFS: Resolution and vertical advection* (Technical Memoranda No. 847). 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. https://doi.org/10.1002/qj.49712656415

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/mwr-d-18-0187.1

Ray Tune Website. (2021). 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*(646), 116–131. https://doi.org/10.1002/qj.521

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Sandu, I., van Niekerk, A., Shepherd, T., Vosper, S., Zadra, A., Bacmeister, J., et al. (2019). Impacts of orography on large-scale atmospheric circulation. *Climate and Atmospheric Science*, *2*(10), 1–8. https://doi.org/10.1038/s41612-019-0065-9

Saunders, R. (2021). The use of satellite data in numerical weather prediction. *Weather*, *76*(3), 95–97. https://doi.org/10.1002/wea.3913

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., et al. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2194), 20200097. https://doi.org/10.1098/rsta.2020.0097

Shepherd, T., Polichtchouk, I., Hogan, R., & Simmons, A. (2018). *Report on stratosphere task force* (Technical Memoranda No. 824). 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*, *115*(D23), D23104. https://doi.org/10.1029/2010jd014457

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. https://doi.org/10.1175/jtech-d-18-0081.1

Trémolet, Y. (2006). Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, *132*(621), 2483–2504. https://doi.org/10.1256/qj.05.224

Vidard, P., Piacentini, A., & Dimet, F.-X. L. (2004). Variational data analysis with control of the forecast bias. *Tellus*, *56*(3), 177–188. https://doi.org/10.1111/j.1600-0870.2004.00057.x

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 Physics of Atmosphere and Ocean, English Translation*, *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. https://doi.org/10.3402/tellusa.v66.23294

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. https://doi.org/10.1029/2018ms001597

Wergen, W. (1992). The effect of model errors in variational assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, *44*(4), 297–313. https://doi.org/10.1034/j.1600-0870.1992.t01-3-00002.x

Zupanski, M. (1993). Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Monthly Weather Review*, *121*(8), 2396–2408. https://doi.org/10.1175/1520-0493(1993)121<2396:rfdvda>2.0.co;2